IDENTIFY PROBLEM

PROVIDE STRUCTURE

UNDERSTAND CONTEXT

DEVELOP STRATEGY

DEVELOP & EXECUTE TACTICS

IDENTIFY AND DEPLOY SOLUTION

SOLVING LARGE, COMPLEX, UNSTRUCTURED PROBLEMS

STATISTICAL ENGINEERING

HANDBOOK

iSEa

INTERNATIONAL STATISTICAL ENGINEERING ASSOCIATION

STATISTICAL ENGINEERING HANDBOOK

Lynne B. Hare
Editor-in-Chief

Chapter Editors:

1. Roger Hoerl
2. Lynne Hare and Roger Hoerl
3. Lynne Hare
4. Ronald Snee
5. Lynne Hare and Roger Hoerl
6. Ronald Does

Editorial Assistant: Lisa Petrine

Preface

This is not the statistics book you learned to detest when you were in school or college, the one you sold at semester's end or left to yellow and gather dust these many years.

This is not a statistics book at all. This is a book about sensible and highly effective methods of organization and of data driven decisions to capitalize on opportunities and to solve problems of all kinds from present, day-to-day issues to longer term problems of obtaining genuine meaning from large, unstructured data sets.

This is a book about statistical engineering as set apart from statistical theory or practice. The focus is on the doing. To that end, it begins with organizing to solve problems through broad organizational perspectives. Then and only lightly as needs demand, does it delve into statistical methods.

As such, statistical engineering fills a gap between theory and practice much the same as chemical engineering fills its corresponding gap with chemistry. It evolved out of needs perceived by statistically aware and highly experienced scientists and engineers who recognized and capitalized on the great advantages of statistical thinking; that all work consists of a series of interconnected processes; that all processes exhibit variability; and that keys to success are understanding and reducing variability.

Cutting through variability on an organizational scale requires an understanding of organizational structure including its various incentives, motivations and constraints, together with knowledge of internal politics and linkages with outside influencers. Navigating these waters is no easy task. But there is no other way. Strong, knowledgeable leadership is essential. One cannot be genius enough, but a leader can form effective teams whose combined knowledge, if capitalized upon, is more than up to the task.

Recognizing the great opportunity, Christine Anderson-Cook, Roger Hoerl, Peter Parker, Ronald Snee and Geoff Vining called a meeting for the initial planning of a society devoted to statistical engineering. It took place in Alexandria, VA, December 9th and 10th, 2017. Participants were William Brenneman, Stephanie DeHart, Laura Freeman, Will Guthrie, Lynne Hare, Roger Hoerl, Dean Neubauer, Peter Parker, Ronald Snee, Stefan Steiner and Geoff Vining.

Together, these people mapped the genesis of statistical engineering and formed the International Statistical Engineering Association (ISEA). Primary among its products is this Statistical Engineering Handbook. It is by no means the final word on the subject. Rather, it is intended to grow through the addition of subject material and case studies.

This handbook represents the diligence of those shown in the following pages. They wrote and re-wrote and took turns editing.

Lynne B. Hare
Plymouth, MA
2021

Handbook Authors

Stan Altan is Senior Director and Research Fellow in the Manufacturing and Applied Statistics department at Janssen Research & Development, LLC. Stan received his Ph.D. from Temple University. He is a fellow of the ASA and serves on the editorial board of the Statistics in Biopharmaceutical Research journal, an ASA publication. His professional interests are in experimental design, linear and nonlinear modeling and more recently statistical methods in support of continuous manufacturing of pharmaceuticals.

Stan contributed to Chapter 5 on Drawing Inferences.

Ronald J.M.M. Does received his Ph.D. in mathematical statistics from the University of Leiden, the Netherlands. He is professor of industrial statistics at the University of Amsterdam and an independent statistical engineer. Ronald is a Fellow of the American Society for Quality (ASQ), a Fellow of the American Statistical Association, an elected member of the International Statistical Institute, and an Academician in the International Academy for Quality. He has received ASQ's Shewhart and Lancaster Medals, and Hunter Award. Also, he was awarded by the European Network for Business and Industrial Statistics with the Box Medal. Ronald has co-authored ten books and more than 200 papers in the fields of statistics, psychometrics, process improvement, healthcare engineering, quality and management. Ronald's independent consulting includes a wide range of clients, including those in healthcare, finances, manufacturing and services. He also serves on the Editorial Boards of Quality Engineering and the Journal of Quality Technology for many years.

Ronald edited and contributed to Chapter 6 on Solution Identification and Deployment.

Lynne Hare, founder of Statistical Strategies, LLC, is a consulting statistician with 50+ years' experience, working at Unilever, NIST, and Nabisco/Kraft, from which he retired, having won their R&D Technical Leadership Award. He is a Fellow of the American Society for Quality and the American Statistical Association and winner of several awards including most recently the ASA Quality and Productivity Section's Gerry Hahn Award for sustained achievement in the field. He serves on the Editorial Review Board for both Quality Engineering and Quality Progress. Lynne holds an undergraduate degree in mathematics from Colorado College and M.S. and Ph.D. from Rutgers University.

Lynne serves as this handbook's editor-in-chief. In addition, he edited Chapter 3 on Data Collection and co-edited Chapter 5 on Drawing Inferences. His contributions also appear in sections of Chapters 2, 3 and 5.

Roger Hoerl received a Ph.D. in statistics from the University of Delaware. He is currently an assistant professor at Union College and Chair of ISEA (2020). The American Statistical Association has recognized him as a Fellow, with the Founders Award and with the Deming Lecture Award. He has also been named Fellow by the American Society for Quality and has received the Brumbaugh Award and the Shewhart Medal.

Roger edited and authored much of Chapter 1 and co-edited Chapter 5. He also contributed to Chapters 2 and 3 while serving as the ISEA Board member overseeing the development of this handbook.

Bart Lameijer is Assistant Professor of Operations Management at the Department of Operations Management of the University of Amsterdam Business School, the Netherlands. He is also co-director of the Institute for Business and Industrial Statistics at the University of Amsterdam Business School. His expertise is in Lean Management and Six Sigma implementation for Operational Excellence. His research interests center on both organization-wide and project-level data-driven process improvement methodology implementation.

Bart contributed to Chapter 6.

Steve Luko is a statistician and engineering fellow at Collins Aerospace, a division of Raytheon Technologies. His applications experience spans 40 years at Collins, Ingersoll Rand and Johnson & Johnson. Steve is a fellow of the American Society for Quality (ASQ) and a certified quality and reliability engineer. ASQ awarded him the Dorian Shainin Medal for applications in quality control which have been adopted by several companies. Active in leadership roles in the American Society for Testing Materials (ASTM), he received the Harold F. Dodge award for standards writing excellence. Steve has taught statistics and mathematics at several Connecticut universities, has published extensively in peer reviewed journals and has provided short courses in many varied statistical disciplines including design of experiments, reliability and statistical process control.

Steve contributed to Chapter 3.

Dean Neubauer (deceased) was an Engineering Fellow in the Process Engineering Directorate of the Manufacturing Technology and Engineering Division of Corning, Inc. where he had been employed since 1981 and where he holds sixteen patents. Dean has also taught in the MS program at the Center for Quality and Applied Statistics at R.I.T. He has been a mainstay on ASTM committees, active on management and editorial review boards such as Technometrics, Journal of Quality Technology and Quality Engineering, and co-author of *Process Quality Control* (4th ed., with Ellis Ott and Edward Schilling). He was a Fellow of the American Society for Quality, a Certified Quality Engineer, an American Statistical Association certified statistician and a Royal Statistical Society chartered statistician. His degrees are from Iowa State University and R.I.T.

Dean contributed to Chapter 3.

Susan O. Schall, founder and Lead Consultant, SOS Consulting, has spent her career solving productivity and quality problems across industries, saving over $250 million and transforming manufacturers to compete today and tomorrow. Prior to consulting, Susan held technical and leadership roles at Kodak, DuPont, GE Lighting and RR Donnelley. Susan has a BS in mathematics from SUNY, Collage at Fredonia, and BS, MS and PhD degrees in industrial engineering from Penn State University. She is also an ASQ Certified Quality Engineer, ASQ Certified Manager of Quality/Organizational Excellence, and a Six Sigma Master Black Belt. She has been recognized for her professional achievements and contributions to industrial engineering education by the Institute of Industrial and Systems Engineers (IISE) as Fellow in 2020 and with the Medallion Award in 2018. She is a Marquis Who's Who Top Executive and Woman if Influence.

Susan contributed to Chapter 2.

Ron Snee is Founder of Snee Associates, LLC, a firm dedicated to the successful implementation of process and organizational improvement initiatives, using Quality by Design, Continued Process Verification, Lean Six Sigma and other data-based improvement approaches that produce bottom line results. He has played a leadership role in 32 major corporate data-based improvement initiatives. Prior to entering the consulting field, he worked at DuPont for 24 years in a variety of technical and managerial assignments. This work has involved many data-based problem-solving ventures involving data collection, analysis and interpretation spanning different fields. These experiences accumulated over more than five decades have produced a deep understanding of the practice and theory of effective data collection. Ron is an Honorary Member and Fellow of the American Society for Quality (ASQ), Fellow of the American Statistical Association (ASA), Fellow of the American Association for the Advancement of Science and Academician in the International Academy for Quality. He has been awarded ASQ's Shewhart, Grant and Distinguished Service Medals, and ASA's Deming Lecture, Dixon Consulting Excellence, and Gerry Hahn Quality and Productivity Achievement awards. Ron has co-authored seven books and more than 330 papers in the fields of statistics, process improvement, quality and management. His work has been recognized by 30 major awards. He received his BA from Washington and Jefferson College and MS and PhD degrees in Applied and Mathematical Statistics from Rutgers University. He also serves as an Adjunct Professor in the Pharmaceutical Programs at Temple and Rutgers Universities.

Ron edited Chapter 4 and contributed to Chapters 1, 2 and 3.

Paul F. Velleman is Emeritus Professor of Statistical Sciences at Cornell University. He is a leading authority on Exploratory Data Analysis and co-author of seven statistics textbooks, the developer of the *Data desk* statistics package, and DASL, a free online library of data for teaching statistics. He is the author and designer of *ActivStats*, the first multimedia program for statistics education, for which he was awarded the EDUCOM Medal for innovative uses of computers in teaching statistics, and the ICTCM Award for Innovation in Using Technology in College Mathematics. He is president of the Data Description Inc., which develops *Data desk, DASL* and *ActivStats*. He is a Fellow of the American Statistical Association and the American Association for the Advancement of Science (AAAS). He was a founding member of the board of the non-profit State Theater of Ithaca and of the Community Foundation of Tompkins County, and currently sits on the board of Bay Chamber Concerts in Rockport, ME and the Camden Conference in Camden, ME.

Paul contributed to Chapter 4.

Geoff Vining received his Ph.D. in statistics from Virginia Tech, where he now teaches. Geoff was the primary initiator and founding Chair of ISEA. He has a long list of awards from the American Society for Quality, most notably Honorary Member, a select group of which there are only six living members. Geoff has an extensive publication and editing record, including writing six books and serving as editor of both Quality Engineering and the Journal of Quality Technology. Geoff has been quite active internationally, organizing conferences and connecting researchers and practitioners across North America, South America, Europe and Asia, in particular. He has extensive applications experience with large, complex, unstructured problems, especially with NASA and the Department of Defense, resulting in several awards.

Geoff contributed to Chapter 1.

Joseph G. Voelkel is Professor Emeritus in the College of Science's School of Mathematical Sciences at Rochester Institute of Technology (RIT), Rochester, New York. He is also a contract employee at the Rochester Data Science Consortium at the University of Rochester. His interests include Experimental Design, Quality Control and Improvement, Reliability, and non-standard and data-science problems in the physical sciences. Before RIT, he was a statistical consultant for Allied-Signal (now Honeywell), working in chemicals, plastics, fibers, water-treatment polymers and agricultural products, as well as inter-laboratory, epidemiological, and toxicology studies. He was a faculty member at the University of Wisconsin, specializing in cancer research. Joe's independent and RIT-based consulting includes a wide range of clients, including those in the optics, electronics, resin, plastics, automotive, laser, LED and bearing industries. He is a Fellow of ASQ, a Past-Chair of ASQ's Statistics Division, and currently serves on the Editorial Board of *Quality Engineering*.

Joe contributed to Chapter 4.

Lisa Petrine is a Program Manager in the College of Engineering at Penn State University. Prior to working at Penn State, Lisa was an Information Technology Analyst for Moravian College, Bethlehem, and Software Consulting Services, Nazareth, both in Pennsylvania. Lisa has a BS in Computer and Information Sciences (with a Business Administration minor) from Kutztown University, Kutztown, PA. As well as working as a Program Manager, Lisa also edits academic and technical/scientific writing.

Lisa worked with the authors to edit this handbook.

Summary

In Chapter 1, we define statistical engineering and explain the basic principles and frameworks underlying the discipline. We briefly review the history of statistical engineering, discuss the typical phases statistical engineering efforts go through, its core processes and how to utilize these core processes in addressing real problems.

Continuing in Chapter 2, we cover statistical engineering as a holistic approach typically involving a diverse set of tools and disciplines. These are integrated based on the context of the specific problem being addressed and overall strategy. While many of these technologies are quantitative, we integrate "soft skills" needed to solve complex problems sustainably. These methodologies, which we refer to as enabling technologies, typically cut across all the phases of Statistical Engineering.

In Chapter 3, we present principles and techniques for acquiring necessary and sufficient data for sensible, practical guidance with the ends being advancement of human well-being through the revelation of improvement opportunities and the identification of solutions. Discussions, presentations and examples focus on primary data issues including the theory of data collection, challenges inherent in the collection of data, the understanding and importance of data pedigree, the quantification of accuracy and precision under the heading of measurement systems analysis and finally, the essential planning of data collection including the statistical design of experiments (DOE), considering all the forgoing, for sound decision making.

Statistical engineering is a data-based methodology involving data analysis. In Chapter 4, three important aspects of data exploration are addressed: theory of data exploration, data cleaning and using Exploratory Data Analysis (EDA). A high-level discussion of EDA is presented to provide the foundation for data exploration. A "Global Positioning System" for an effective EDA journey is presented. Data are rarely "clean" and can have a variety of problems and limitations. Five types of data cleaning problems and methods for conducting data cleaning are discussed. The chapter concludes by showing how EDA can be used to understand data prior to doing formal statistical analyses. EDA helps the analyst to be alert to unexpected patterns, relationships and extraordinary cases. Some tools useful in using data analysis are described and illustrated.

In Chapter 5, we discuss a formal approach to drawing conclusions about an entire population or process of interest, based on a sample, or subset, of this population. Projecting from a sample of a population to the full population has risks associated with both bias and variation. Bias enters the picture when sampling units lack full and fair representation of the population. Sampling variation is the unavoidable difference between a sampled value and the true, but usually unknown corresponding value of the population. If sampling is carried out properly, avoiding pitfalls of bias and variation, we can derive accurate inferences about the entire population. In this chapter we also discuss the underlying theory of statistical inference, common reference probability distributions utilized in inference, and common methods of inference, such as confidence and prediction intervals, as well as hypothesis testing.

After the statistical engineer (SE) has learned about the problem it is time to search and select the most prevalent influences and to generate and select the most adequate and elegant solutions. In

Chapter 6, we discuss the process of solution identification and deployment as an iterative process. This process is characterized by phases of divergence and convergence. To successfully arrive at the best solutions collaborations are crucial. Consultation of experts in the problem area, hints and clues about possible influence factors and solutions from subject matter experts, and early understanding of the managerial appetite for anticipated investments are crucial for success through solution identification and deployment. Selection of the best solution requires understanding what is fundamentally causing the problem. Therefore, this phase begins with the identification of possible factors causing the problem and defines a process for identifying the most prevalent factors that can be controlled, compensated or eliminated. Finally, in this phase the SE starts to anticipate what is needed for solution deployment. Once possible solutions are revealed, initial considerations about what is needed for deployment starts. Thereby, the process of preparing the organization and key leaders for deployment is commenced. Possible hurdles can be identified and timely action to prevent or mitigate deployment obstacles can be taken.

Table of Contents

Chapter 1 Introduction to Statistical Engineering

Table of Contents

Preface

In this introductory chapter we define statistical engineering and explain the basic principles and frameworks underlying the discipline. We include a brief review of the history of statistical engineering; discuss the typical phases statistical engineering efforts go through, its core processes and how to utilize these core processes in addressing real problems. There is minimal discussion of the statistical tools typically applied in statistical engineering, as these are presented in subsequent chapters.

Section 1.1 - What is Statistical Engineering?

1.1.1 Objectives

The purpose of this section is to explain what statistical engineering is, i.e., how it is defined, how it works, why it is needed, as well as the basics of its underlying theory.

1.1.2 Outline

We begin with an elucidation of the definition of statistical engineering. Next, we explain why it is needed as a discipline, and then present the current state of the art in terms of its underlying theory.

1.1.3 Definition and Elaboration

The discipline of statistical engineering is *the study of the systematic integration of statistical concepts, methods and tools, often with other relevant disciplines, to solve important problems sustainably*.

Several words in this definition warrant explanation. First, statistical engineering is defined as a *discipline*, the *study* of something, not as a set of tools or techniques. Secondly, as an *engineering* discipline it does not focus on advancing the fundamental knowledge of the physical world, i.e., it is not a science. Rather, as with other engineering disciplines, it utilizes existing *concepts, methods and tools* in novel ways to achieve novel results. In this sense it is complementary to statistical science, just as chemical engineering is complementary to chemistry.

Concepts, *methods* and *tools* are each important and need to be *integrated*. That is, formal statistical *methods* (e.g., time series or regression analysis) and individual *tools* (e.g., residual plots) need to be integrated with *concepts (e.g.,* the advantages of randomization) and the need to understand the quality ("pedigree") of observational data prior to developing models (Hoerl and Snee 2018). When addressing straightforward issues, a single statistical tool may suffice. However, as noted by Hardin et al. (2015), when solving the challenging problems often faced by practitioners, obtaining a viable solution typically requires integration of multiple methods into an overall strategy and sequential approach.

Such integration should be done in a *systematic*, rather than ad hoc manner. Throughout the history of statistics, good statisticians have generally figured out how to integrate concepts, methods and tools to solve problems. One classic example would be Box and Wilson's (1951) integration of experimental design and regression into an overall sequential strategy for the empirical optimization of processes, which we know today as response surface methodology.

It would appear clear, however, that despite many historical examples of successful integration, there is little existing theory in the literature on how to best accomplish such integration in

general, that is, with a new problem. Due to a lack of theory, new integration problems are often attacked with a trial-and-error approach. However, the theory of statistical engineering, discussed below, provides guidance for a *systematic* approach, which is likely to be much more effective. In addition, such theory can be formally studied, taught and advanced over time.

By the word *theory*, we do not refer to mathematical statistics. Rather, we refer to development of an overall methodology, based on the scientific method, by which one might approach integration in a methodical (systematic) rather than ad hoc manner. Note that *theory* may be defined as: "A coherent group of general propositions used to explain a phenomenon" (Hoerl and Snee 2017). Note that neither this nor other common definitions of theory contain explicit requirements for mathematics, although mathematics is often important.

In addition, for many of the important problems facing practitioners, such integration must include *other disciplines* beyond statistics. For example, almost by definition, information technology (IT) is required to address "Big Data" problems (see the ASA statement on Data Science at http://www.amstat.org/misc/datasciencestatement.pdf). In fact, the authors of this handbook have found that IT is needed to some degree to solve most important real problems. Kendall and Fulenwider (2000) explain how critical IT is to successful Six Sigma projects. We feel that the same is true of statistical engineering. Challenging problems, such as developing personalized medicine protocols through genomics, for example, are virtually impossible to resolve without effective and innovative use of IT.

Other disciplines may be needed as well, including natural sciences, other engineering disciplines and social sciences, such as organizational effectiveness, psychology or social networking theory, depending on the specific problem being addressed. As one example, the improvement methodology Lean Six Sigma (Antony et al. 2017) is essentially the integration of diverse statistical methods, including control charts, experimental design and regression. It includes various quality concepts and methods, including Pareto charts, mistake proofing, and quality function deployment (QFD), in addition to the efficiency concepts and methods from Lean manufacturing. These efficiency concepts and methods could be considered under the umbrella of the discipline of industrial engineering.

As an engineering discipline, the ultimate goal of statistical engineering is to *solve important problems*. While this may seem obvious, an emphasis on solving important problems gives statistical engineering perhaps its most important attribute, being *tool-agnostic*. That is, statistical engineering is neither Bayesian nor frequentist, neither parametric nor non-parametric (or semi-parametric) and does not promote either classical or computer-aided designs, per se. Rather, as an engineering discipline, its "loyalty" is to solving the problem and generating results, not to a predetermined set of methods. Tools are important, but within a statistical engineering paradigm they are chosen based on the unique nature of the problem to provide the best possible solution, rather than predetermined based on personal preferences. Various philosophies and tool sets may be employed and integrated.

Further, statistical engineering seeks solutions that are *sustainable*. We argue that many solutions, including those published in professional journals, provide technical solutions. But all too frequently, these solutions are not sustainable over time. Of course, virtually no solution will

be permanent. Statistical engineering seeks solutions that are *sustainable* beyond the immediate time frame and hopefully last until the problem itself changes, or until new technology becomes available, enabling an even better solution.

In practice, purely technical solutions often overlook organizational, political or psychological constraints. To be sustainable, the solution must eventually be embedded into standard work procedures and best practices, typically via IT. An interesting example from the related discipline of data science is the classic Netflix competition, in which Netflix paid $1,000,000 to the team that developed the "best" model to predict customer ratings of movies.

As noted by Donoho (2017), however, the winning solution was never actually implemented by Netflix, because Netflix found that the time and expense involved in maintaining the 107 individual models utilized within the overall ensemble (see Fung 2013) was not worth the small improvement in accuracy. So, a team won the competition and the $1,000,000 award, but it did not actually solve Netflix's business problem. Clearly, the technical solution is only a part of solving important problems *sustainably*.

1.1.4 Why Statistical Engineering?

It is certainly logical to ask, "why a new discipline is needed?" Even allowing that one is, "why it should be statistical engineering?" As noted previously, good statisticians have integrated multiple statistical methods and tools from other disciplines, for a long time. In this sense, we could say that statistical engineering itself is old. However, as also noted above, such applications have typically been presented as isolated case studies utilizing ingenuity and creativity to provide novel solutions to complex problems. What has been missing is a concise presentation of an underlying theory as to how the researchers developed their solutions. A body of research is needed to fill in this gap, to develop an underlying theory as to how and why such problems should be addressed. In this sense, we say that statistical engineering is a new discipline, even though statistical engineering itself is old.

The main reason statistical engineering was needed in these case studies was to solve problems that were not straightforward "textbook" problems. Textbook problems are typically well structured, have a clear objective and a single, correct answer; generally, one that can be looked up in somewhere in the textbook. For example, a data set might be presented with paired data, such as "before and after" weights from a diet evaluation study. Clearly, with paired data a standard independent samples t-test would not be appropriate. Rather, a paired t-test is likely to provide the "correct" analysis. We can look this up in the textbook to verify that it is the appropriate analysis, making reasonable assumptions.

However, real problems faced by practitioners are not usually so well structured. The specific problem to be solved may not be clear. Appropriate data for solving the problem may not yet exist. For example, suppose an international corporation's reputation was damaged by the discovery that a supplier was, unknown to the corporation, using child or slave labor in a developing country. The corporation needs to address the issue immediately, so as not to support such human rights violations. Then it can perhaps begin a much longer process of rebuilding its reputation. But what exactly does "rebuilding its reputation" mean? How would this be measured

and verified? How should the company go about acquiring data to set a baseline on its reputation? The answers to these questions are not obvious, and there is certainly no "correct" answer to look up in a textbook.

Further, it is unlikely that one statistical method would suffice to solve this problem. Some type of survey or perhaps web scrapping of social media could be involved, followed by analysis of the data, perhaps with multiple tools. Additional data gathering and analysis steps might follow. In other words, there would be a need to first think through an overall strategy of "how" to attack the problem, then acquire data, then analyze the data using a mix of graphical and analytical tools. That is, there would be a need to link and integrate multiple tools in a sequential fashion, based on a strategy.

Very few statistical textbooks provide guidance on how to link and integrate multiple tools, especially through sequential cycles of data gathering and analysis. Rather, most textbooks provide details on individual methods, one method at a time: descriptive statistics, probability, confidence intervals, hypothesis testing, regression, etc. Further, a theoretical foundation is needed to provide guidance on how to accomplish this integration, including the underlying theory of statistical engineering, which we present shortly.

Several other authors have noted this gap in the current body of research on tool integration to solve complex problems. For example, Meng (2009) pointed to the same issue. Meng subsequently added a new course in the Harvard statistics department curriculum, Stat 399, which "…emphasizes deep, broad, and creative statistical thinking, instead of technical problems that correspond to a recognizable textbook chapter." Complex problems rarely correspond to a recognizable textbook chapter!

Shortly after the publication of Meng's paper, Susan Hockfield, then President of Massachusetts Institute of Technology (MIT) and a member of the General Electric (GE) Board of Directors, gave an interesting perspective on the relationship between science and engineering, which has obvious ramifications for statistical engineering and statistical science (Hockfield 2010). She noted that around the dawn of the 20th century, physicists discovered the basic building blocks of the universe (i.e., the periodic table), which could be considered a "parts list." However, it was engineers who figured out how this parts list could be put to best use, subsequently driving the electronics and computer revolutions. Similarly, Hockfield noted that biologists had recently discovered the basic building blocks of life (the human genome), another "parts list," and now engineers are finding creative ways to use this parts list, such as in personalized medicine.

A key point Hockfield made was that there has been for some time a consistent "separation of labor" between science and engineering across diverse disciplines, although it is important that they collaborate. To be more precise in terminology, common definitions of the word "science" are similar to: "the study and advancement of the fundamental knowledge of the physical or natural word" (https://www.merriam-webster.com/dictionary/science). Various definitions of engineering are also available (https://www.merriam-webster.com/dictionary/engineering), but accepted definitions generally emphasize "utilization of existing science and mathematics in novel ways to benefit humankind". An old saying in the engineering community is, "An engineer is someone who can accomplish for $1 what any fool can accomplish for $2." While science

emphasizes development of new fundamental knowledge, engineering finds creative ways to use this knowledge for the benefit of society.

We argue that this distinction between science and engineering applies to statistics quite well. Statisticians have been developing an excellent toolkit for over a century, which could also be considered a "parts list" using Hockfield's terminology. This is what the vast majority of statistics textbooks emphasize, as noted above. However, we argue that insufficient thought has gone into the engineering problem of how to best integrate multiple tools in creative ways to solve complex problems. At least, insufficient thought has gone into documenting the underlying theory of how to approach this engineering problem in general.

Acknowledging this problem, the American Statistical Association (ASA) published guidelines for the design of undergraduate statistics programs, noting (ASA 2014, p. 6):

> Undergraduates need practice using all steps of the scientific method to tackle real research questions. All too often, undergraduate statistics majors are handed a "canned" dataset and told to analyze it using the methods currently being studied. This approach may leave them unable to solve more complex problems out of context, especially those involving large, unstructured data…. Students need practice developing a unified approach to statistical analysis and integrating multiple methods in an iterative manner.

Unfortunately, the ASA report did not suggest a specific method to provide a "unified approach to statistical analysis and integrating multiple methods in an interactive manner". This is, in fact, the gap statistical engineering is intended to fill.

Michael Jordan, jointly appointed to the Department of Electrical Engineering and Computer Science, and Department of Statistics at the University of California, Berkeley, commented on the need for statistical engineering at a symposium celebrating the 50th anniversary of statistics at the University of Michigan (Jordan, 2019). During his presentation, he admonished the participants, "Let's embrace being engineers – and think about what 'statistical engineering' could look like, as a counterpart to 'statistical science'."

So, there appears to be a clear consensus that a deep theoretical foundation in individual methods, while certainly valuable and needed, is not sufficient. In addition to sound statistical science, the profession also needs a well-developed theory and practice of statistical engineering, to ensure that society benefits from the many advancements that have been made in statistical science.

At first glance, some may feel that what we are calling statistical engineering is nothing more than a rebranding of applied statistics. However, this would be analogous to saying that chemical engineering is nothing more than a rebranding of applied chemistry. As a simplistic example, consider parents who buy their children a chemistry set for a birthday or holiday. If the children mix vinegar and baking soda, they might create a toy "volcano", due to the subsequent chemical reaction. This is certainly applied chemistry! However, not many would consider this to be

chemical engineering. The children did not use the laws of chemistry to engineer a solution to a real problem.

Similarly, whenever someone applies a statistical method to real data, this constitutes applied statistics. In many applications, particularly with relative straightforward problems, one method found in a textbook will suffice. The problem has now been successfully solved through applied statistics. However, with more complex problems, a single method will rarely suffice. More likely, a novel solution will have to be engineered, using the "parts list" of statistical science tools, perhaps integrated with tools from other disciplines.

Another important distinction between statistical engineering and applied statistics is that statistical engineering has an underlying theory. While "the theory of applied statistics" would be an oxymoron, it is *applied* statistics, not *theoretical* statistics. As noted by Nair (2008), there is a clear and well-established delineation between theoretical statistics and applied statistics, although hopefully these are intertwined.

1.1.5 The Underlying Theory of Statistical Engineering

1.1.5.1 What is Theory?

As we present the theory of statistical engineering, we should acknowledge that it is in its early stages of development. Of course, the theories of all known disciplines are in essence, "works in progress", in that research in each continues. For example, mathematics is one of the oldest known disciplines and has been formally studied and researched for millennia. And yet, rigorous research in mathematics continues at universities and colleges around the world, with no evidence of slowing down. Having noted the ongoing development of the theories of all disciplines, the current state of statistical engineering theory is admittedly basic and relatively crude compared to more established disciplines, including traditional engineering disciplines. We anticipate that future research will add to the current body of knowledge, eventually producing rich literature documenting the theory of statistical engineering to a degree of rigor on par with other engineering disciplines.

The underlying theory of statistical engineering is quite different from the underlying theory of statistical science, which is based on mathematical statistics. Most of the theory of statistical science can be proven or derived using formal mathematics; calculus, real analysis, linear algebra, and so on. The theory of statistical engineering is not mathematical in nature, however. In other words, it is not based on a "theorem-proof" model. Rather, it is based more on empirical research, which demonstrates what does and does not tend to work to solve important problems sustainably, and why it does so. Of course, there may be proofs that certain tools work better than others under specific assumptions.

While some statisticians might not consider such theory to be a true theory, it is important to keep in mind that the fundamental theory of most disciplines cannot be proven mathematically. For example, no one to date has mathematically *proven* that the Keynesian theory of economics is "correct" or even "better" than its main alternative, New Classical Economics (https://www.econlib.org/library/Enc/KeynesianEconomics.html). Of course, no one has *proven*

that it is not correct either. Psychology, sociology, management science and geology are a brief list of disciplines that have extensive bodies of research and underlying theory, but which rarely publish "theory-proof" articles in their journals.

Madigan and Stuetzle, in their discussion of Lindsay et al. (2004, p. 409), essentially made this same point, "The issues we raise above have nothing to do with the old distinction between applied statistics and theoretical statistics. The traditional viewpoint equates statistical theory with mathematics and thence with intellectual depth and rigor, but this misrepresents the notion of theory. We agree with the viewpoint that David Cox expressed at the 2002 NSF Workshop on the Future of Statistics that 'theory is primarily conceptual,' rather than mathematical."

The word "theory" itself must be properly understood to understand the points above, as well as the theory of statistical engineering. As with engineering and science, many definitions of the word theory are possible (e.g., http://www.merriam-webster.com/dictionary/theory). However, reasonable and accepted definitions typically state something similar to, "a coherent group of general propositions used to explain a phenomenon." Obviously, there is no explicit requirement in such definitions for mathematics to be involved, although it often is. The underlying theory of physics, for example, involves considerable mathematics, but of course not all of the theory of physics is mathematical. If it were, physics would be considered a subfield of mathematics.

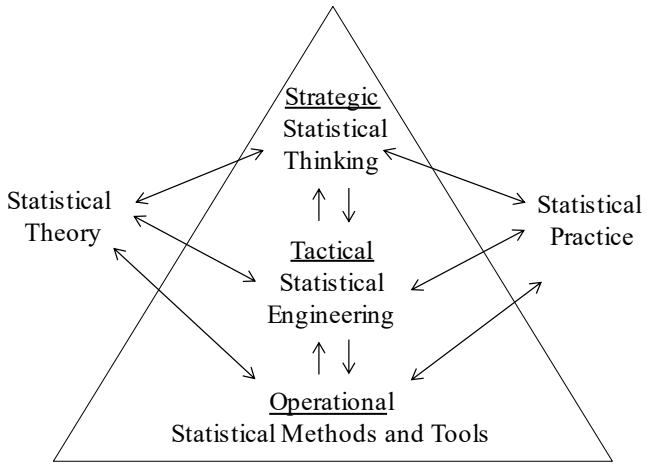
There is now, in fact, "a coherent group of general propositions used to explain" statistical engineering. These propositions are presented and explained below. There are two other aspects of the underlying theory that we feel are also important:

1. a conceptual model of the relationship between statistical engineering and the statistical methods
2. an overall model to guide application of statistical engineering to large, complex, unstructured problems

1.1.5.2 How Does Statistical Engineering Fit?

Figure 1.1 (Snee and Hoerl 2017) depicts the statistics discipline as a system, with strategic, tactical and operational levels, each of which has both a theoretical and an applied aspect. The strategic-tactical-operational model is one that has been used in the military, business, government and other organizations for a long time, perhaps millennia. The strategic level is where high-level decisions are made about the organization's fundamental purpose, what it views as success and how it will win in a competitive environment. This is where such things as vision, mission, values and so on are determined. Per Meng (2009), statistical thinking is at the strategic level for the statistics discipline; that is, how we think about statistics itself, and its relationship with other disciplines. This includes how to interpret the world from a stochastic versus deterministic viewpoint, how we think about data and its relationship to subject matter theory in problem solving, and so on.

Figure 1.1 The Statistics Discipline as a System



The operational aspect of this type of model is where the "rubber hits the road," that is, where the actual work of the organization is accomplished. In the military, it would be the "front lines", in manufacturing it would be the production floor, and in a hospital, where patients are seen and treated. In Figure 1.1, the methodologies of statistics, such as time series models, experimental design, statistical process control, and so on, would be at the operational level. In fact, when most people, both statisticians and non-statisticians, think about the statistics discipline, it is likely that they primarily think of this operational level – the tools themselves. Research on the tools over the decades has produced a rich and deep understanding of how and why these tools work, as well as invention of newer and more effective tools. When this theory is integrated with learning from actual applications, we refer to this combined body of knowledge as statistical science.

Note that Figure 1.1 illustrates a theoretical and applied component at all three levels. For example, at the operational level we both perform research on the theory of the individual methods and apply them to real problems. Similarly, we can debate the theory of statistical thinking: what should be the fundamental principles of the discipline? A stochastic view of the world would seem obvious as a core principle, but what about the proper relationship between statistics, data science, computer science, industrial engineering or operations research? There

could no doubt be serious debates as to how the statistics profession should view its boundaries and proper relationships with these other disciplines.

Of course, these concepts are hopefully applied in practice, such as the Food and Drug Administration (FDA) insisting the clinical trials be based on randomized experiments, rather than solely on observational data. Fortunately, the FDA understands the qualitative distinction between observational data and data from randomized experiments.

The tactical level of the organization exists to develop tactics to carry out the strategy. In the business world, senior executives set strategy – where to place the "big bets" in new product development, which businesses or markets to get out of, which to get in to, and so on. However, the employees on the "front lines" in manufacturing, sales, logistics, etc., are far removed, both physically and conceptually, from the executive office. Middle management therefore exists to take the strategic direction and figure out specific tactics within each function to ensure that the strategy succeeds. For success to occur, the "front lines" need to take actions that are supportive of the overall strategy. In some ways, this tactical level of middle management has the toughest job, which is one reason that "middle management" has a negative connotation in many circles.

In the statistics discipline, we have found a serious gap between the higher-level principles of statistical thinking and utilization of the individual tools. That is, distinguished statisticians may opine on the proper way to think about the discipline and how it can succeed in expository articles, but such opining is far removed from the tools research being done in academia, or from the routine applications of practitioners. In essence, there is no "middle management" in the statistics profession. In our view, the critical question of how researchers or practitioners should research and use statistical methods in such a way as to be consistent with the principles of statistical thinking has gone largely unanswered.

Wild and Pfannkuch (1999) identified this issue two decades ago and provided some suggestions as to how to address it. We propose that statistical engineering can further fill this gap and serve as the tactical element of the discipline, linking the individual methods with the fundamental principles of statistical thinking. That is, statistical engineering, as we discuss below, is based on fundamental statistical thinking principles. It applies these principles to guide the linking and integration of individual tools to solve a real problem, typically one that is large, complex and unstructured. Therefore, it is providing guidance on how to take the individual tools and utilize them in a manner consistent with the strategy. For example, statistical engineering provides a specific "unified approach to statistical analysis and integrating multiple methods in an iterative manner", one of the strategic principles mentioned in the ASA guidelines for undergraduate statistical education, discussed previously.

Again, while it is important to develop a theory of how to do this, it is equally important to apply this theory to real problems. Such application provides a feedback loop to the theory, noting what does and does not work in practice, when addressing real problems versus textbook problems.

1.1.5.3 A Coherent Group of General Propositions

The statistics profession has certainly learned and documented important principles over the decades concerning solution of large, complex and unstructured problems. However, we do not feel that they have been effectively integrated into a formal framework. If integrated, they are in some sense a "theory" or "a coherent group of general propositions used to explain a phenomenon."

Most experienced practitioners learn these principles and pitfalls "on the job," often through making their own mistakes. At this point, they might be considered principles of statistical practice or applied statistics. We argue that such principles can be studied, documented, debated and enhanced over time, as well as formally taught to students. Under these circumstances, they would be considered a theory. The logical expectation in most disciplines is that theory and practice should gradually converge over time; we believe that the same should be true of statistics.

The most critical propositions or principles of statistical engineering applied to large, complex, unstructured problems can be loosely grouped into the five major categories listed in Table 1.1 (Hoerl and Snee 2017). The first principle emphasizes the need for developing an understanding of the problem context. With straightforward problems, little time needs to be invested in studying the background or context. If someone asks you what time it is, you do not need to study the history of watchmaking to answer the question – just look at your watch or cell phone!

Table 1.1 Fundamental Principles of Statistical Engineering

1. Understanding of the problem **context**
2. Development of a problem-solving **strategy**
3. Consideration of the data **pedigree**
4. Integration of sound **subject matter theory** (domain knowledge)
5. Utilization of **sequential approaches**

Suppose a city wishes to address gang violence. One could come up with some "obvious" solutions, such as providing more police to patrol the streets, trying to infiltrate the gangs with informants or even modifying the criminal justice system. However, with large, complex, unstructured problems such as these, "obvious" solutions rarely work well. Rather, to have a serious impact on gang violence the city would likely need to develop a deep understanding of the gangs themselves; why people join them in the first place, how they recruit and operate, their specific criminal activities, how the gangs relate to one another, their internal codes of conduct, etc. An effective response is only likely to be identified after developing a deep understanding of these contextual issues. This same principle generally holds for large, complex, unstructured problems in business, engineering and healthcare.

The second principle highlights the fact that serious thought needs to go into development of a problem-solving strategy once the context is understood. With straightforward problems, the correct solution can often be found in a textbook; no overall strategy is needed. However, with complex problems, especially those that are unstructured, the plan of attack will often not be clear. In fact, there is rarely a single "correct" approach. Therefore, significant time and planning need to go into developing the overall approach to solve the problem. Hoerl et al. (2014)

discussed these points in greater detail within the context of Big Data problems. Unfortunately, the word "strategy" rarely appears in the indices of statistics textbooks, providing another illustration of the difference between statistical engineering and statistical science.

Understanding of the data pedigree, Table 1.1 point 3, is important in any data analysis. Too often analysts assume that the data are "perfect", representing a random sample from the population of interest. Almost without exception, they do not represent a random sample from the population of interest. Data always have some limitations, whether they involve biased or limited sampling, outliers, missing data, missing variables, the wrong timeframe of data collection or just outright errors, such as recording a "34" when the actual number was "43". Murphy's Law, which says that "anything that can go wrong, will go wrong," certainly applies to data collection.

The *pedigree* documents how the data were collected, what specifically they represent, how samples were obtain and measured, and what, if any, changes or deletions were made to the data over time (the "chain of custody"). Hoerl and Snee (2018) provide more detail on the concept and use of data pedigree, and an elaboration of this topic is also given in the Data Acquisition chapter of this handbook.

The next two points emphasize that statistical engineering views statistical and other tools from the perspective of the scientific method. Statistical methods are viewed as enablers of the scientific method, not substitutes for it. While this point may seem obvious, we note that few statistics textbooks formally discuss the scientific method, or how statistics fits within it. In particular, few discuss the critical importance of subject matter knowledge in acquiring data, analyzing it statistically, and interpreting the analyses; this is the emphasis of the fourth principle.

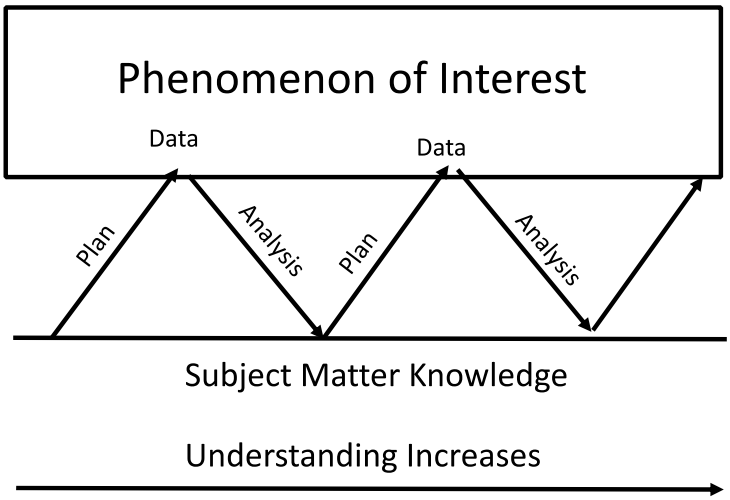
Subject matter (domain) knowledge is everything we know about the phenomenon under study, either from relevant theory, such as physics, epidemiology, or economics, or from previous data analyses. Such knowledge is needed from the very beginning of applications of statistical engineering, even in identifying the true problem, that is, the root cause, rather than just the symptoms. If scientists had all possible knowledge, they would not need statistics or statistical engineering. Statistics is only needed because scientific knowledge is not complete, and empirical approaches – based on data collection and analysis – are often needed to "fill in the gaps" in our scientific knowledge. Eventually, after the data analyses are confirmed, they augment our previous scientific knowledge, enhancing our understanding. This process continues through sequential cycles of the scientific method, eventually producing a mature discipline, such as physics or chemistry.

Sequential approaches are also core to the scientific method and are the emphasis of the fifth principle in Table 1.1. Most applications in statistics textbooks tend to be "one shot studies", where a data set is given, and the "correct" statistical method is applied, allowing the authors to move on to the next data set or next problem. The same is true of homework problems. For example, "what is the correct method to apply to this data?" Real problems, particularly large, complex, unstructured problems, are not so simple. There is no single "correct" method, and in

most cases multiple statistical methods and perhaps multiple disciplines are needed. In other words, a sequential approach is needed.

Each time practitioners determine the specific data needed, they do so based on their current understanding, that is, their current subject matter knowledge. They often have specific questions they need answered to "fill in the gaps". Once they obtain the data and begin to analyze it, typically with multiple tools, they may answer some questions, but others may arise unexpectedly. For example, why is every fourth data point high? Therefore, additional rounds of data gathering and analysis are typically needed. Fortunately, with each round, they become a little more knowledgeable and can ask better and more specific questions. Their understanding gradually increases through these sequential cycles of the scientific method, producing greater understanding. This is illustrated in Figure 1.2, based off a similar graph in Hoerl and Snee 2020, which is itself based off an earlier version from Box, Hunter, and Hunter (1978).

Figure 1.2 The Sequential Nature of Statistical Engineering



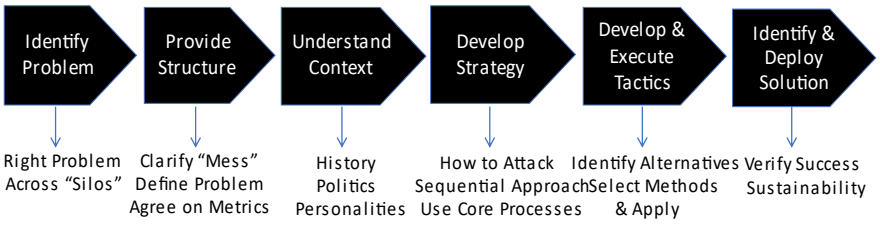
1.1.5.4 A Framework for Statistical Engineering Projects

As previously noted, there is no "correct" solution to large, complex, unstructured problems. Similarly, there is no "cookbook" that will lead practitioners step-by-step through the successful completion of all such projects. However, there is a framework to give some guidance as to how to think about approaching such problems. Figure 1.3, based on Hoerl and Snee (2017), shows the phases that statistical engineering projects typically go through. As an analogy, each child grows up to maturity along a different path; no two children, even "identical" twins, grow up exactly the same. However, the discipline of child development has documented the fact that virtually all children go through the same set of phases growing up, although uniquely. For example, "terrible twos", "fantastic fours", pre-teens, teenagers, etc., are layman's terms for these child development phases commonly used by parents.

So, it is important to keep in mind that Figure 1.3 provides a general framework, not a "cookbook". As previously noted in the discussion of fundamental principles, practitioners will generally need to develop a unique strategy for each problem, based its unique context. Therefore, while statistical engineering projects will generally go through each of these phases, they will do so in unique ways, just as children go through child development in unique ways.

It should also be noted that Figure 1.3 is similar in nature to other problem-solving frameworks, such as DiBenedetto et al. (2014), the Job Task Analysis (JTA) framework and "domains" from the Certified Analytics Professional (CAP) program (INFORMS 2018), and the Data Analytics Lifecycle (EMC Education Services 2015). While each of these frameworks has distinctive aspects, with Figure 1.3 focusing on large, complex, unstructured problems, there is enough overlap to provide confidence that each provides a reasonable approach.

Figure 1.3 The Phases of Statistical Engineering



The first phase in Figure 1.3 is to identify the problem. This might sound easy, and in some cases it is. However, as noted by D. K. J. Lin (2014, personal communication), "Finding a good problem is harder than finding a good solution." Also, large and complex problems typically cross organizational boundaries or "silos". Because it is usually easier to work "within" a silo than "across" silos, teams will often focus their problem-solving efforts in their silo, working on the symptoms of the larger problem. If multiple teams work on the same large problem, but each focuses on the symptoms within their silos, the net result is often teams working at cross purposes, each trying to push the problem from their silo to another silo. The real problem, crossing multiple silos, may not even be recognized, much less addressed.

A classic example of this phenomenon occurs when businesses attempt to effectively manage their overall order fulfillment system, from sales to production planning to warehousing and inventory to logistics, ultimately delivering the product to customers in a timely fashion. Obviously, this overall order fulfillment system is a large, complex system. In most businesses, it is broken up into individual silos, representing each functional area involved, such as a sales team, a production planning team, a warehousing and inventory team, logistics or product delivery, and customer management, which focuses on "keeping the customer happy". Periodically, there will be a business drive to reduce inventory costs and working capital, putting pressure on the warehousing and inventory team to reduce the inventory levels as low as

possible. At the same time, a team from logistics or customer management may be working on a project to provide more timely deliveries to customers, with no product outages.

It should be obvious that both teams are working on the same fundamental problem – order fulfillment. Both are only working within their own respective silos, focusing on the symptoms they see, high inventory costs and late or incomplete customer deliveries. The net result is that one team is figuring out how to lower inventories, while the other is working on how to increase them. No one has identified the real problem they should all be working on: the large, complex, unstructured problem of optimizing the order fulfillment system, whatever that might mean when properly structured. The overall system could no doubt be improved, but this would require cooperation and everyone having the same understanding of what the real problem is and what success would look like. They would need to identify the *right* problem.

Once the right problem has been identified, it usually needs to be properly structured. As noted by X. Tort (2018, personal communication), what we typically see initially, unfortunately, is a "mess". It is virtually impossible to solve a mess. Rather, we first need to convert the mess into a formal problem. Once we have a formal problem, we can move forward to solve it. The process of converting a mess into a problem is what we call *providing structure*.

In our order fulfillment example, we may initially see a mess in which we have too much finished product inventory (perhaps some is expiring before we can deliver it to customers), too much work-in-progress inventory, upset customers who do not know where their product is or why it was late or incomplete, manufacturing disruptions, dysfunctional teams that do not like each other and will not work across silos, recurring quality issues resulting in more work-in-progress and late shipments, and perhaps pressure from senior management demanding that the situation be "fixed" ASAP, but not providing any methodology to fix it.

As a next step, the organization would define the right problem properly, considering the overall order fulfillment system. Since it is typically impossible to minimize inventory while at the same time minimizing late customer deliveries, what exactly would success look like? How would it be measured? It should be clear that there is no obvious problem statement or a single, quantitative objective to be maximized. Considerable work may be required to convert this mess into a formal problem that can be attacked, and to obtain organization alignment across silos and with senior leadership.

The next phase is to understand the context of the problem, which is one of the fundamental principles of statistical engineering and was discussed in the previous section. As noted there, large and complex problems have resisted solution for a reason; "obvious" solutions do not generally work. Only once the problem's root causes (the history of previous efforts – including why they failed and the technical, political and social background of the problem) are properly understood, can the team develop a viable approach to solution. This usually requires a lot of hard work but is absolutely necessary for big problems.

Once the right problem has been properly identified, the mess has been converted to a formal problem statement, and the context of the problem is properly understood, the team is in position to develop a strategy to address it. As noted above, a strategy is needed because multiple

methods and perhaps multiple disciplines will be required, and all of these need to be integrated into an overall approach or game plan. In sports, one aspect of a head coach's responsibilities is to prepare a "game plan" for each opponent. The individual game plans may be quite different from each other, depending on the strengths and weaknesses of the opponents that the team faces. This game plan is a strategy that the coaching staff believes will maximize the chances of success. However, if the players are not all on the same page, and some are not following the game plan, failure is likely.

Similarly, a key role of project leadership is to develop a game plan, or strategy, to solve the problem, and then ensure that everyone on the team, even people from different silos, are all on the same page. This is easier said than done, because people from different silos and with different skills sets may have their own ideas about how the project should proceed. They may not agree with the strategy and start to proceed on their own "closet projects". Such splintering of the team rarely works well, just as it does not in sports. The statistical engineering strategy will typically involve application of a series of statistical and other methods, linked and integrated in a logical manner. Note that the strategy for solving a particular problem is obviously at a much lower level than the overall strategy for a business, university or other organization. Both are examples of strategy, however.

Once a strategy has been developed and everyone is on the same page, the team needs to develop and employ tactics to carry out the project. A strategy, while critically important, is just a plan. To win on the sports field, the team needs to block, tackle, pass, catch, etc., in order to implement the strategy. Tactics are more detailed elements of the overall strategy that provide specific direction at the operational level. For example, suppose our strategy for order fulfillment includes a decision that for now we will prioritize customer fulfillment (minimizing late deliveries) over inventory reduction. We still need specific methods for fulfilling orders; a high-level plan is not sufficient. In the tactics phase, we figure out specific methods to fulfill orders more consistently, and then deploy these in operations to see how well they worked. The tactics will generally involve selection of individual statistical and other methods within each of the core processes discussed above.

Once the strategy and tactics are in place, the team can "take the field" and begin implementing them, i.e., solving the problem. For statistical engineering problems this will result in several statistical and non-statistical tools utilized in a sequential strategy. The results of the first analysis may change the ensuing tactics, just as when sports teams find themselves way behind at halftime they may "ditch the game plan" and start over, or perhaps make less dramatic halftime adjustments.

In the course of applying these methods in a systematic fashion, the team should begin to learn and identify specific actions they could take to address the problem. In most cases, these actions will need to be piloted to verify that they work and do not cause unforeseen issues. Gradually, a final solution is identified and deployed. If it does not work as well as anticipated, the team may need to reloop back to the strategy or tactics phases. Once a satisfactory solution is obtained, the team still needs to consider sustainability. Therefore, a "control plan" is typically needed to embed the solution into standard work processes, as well as to identify how the system should be monitored over time, and what steps employees should take when backsliding is detected.

Even in a best-case scenario, there will be opportunity for further learning and improvement. Therefore, a new improvement initiative or project may make sense, to follow up on the first team's results. The cycle of improvement from the scientific method continues.

1.1.5.5 The Core Processes of Statistical Engineering

The methods needed within the statistical engineering strategy are often selected from five major categories, or "core processes", which represent the major "whats" of statistical science. That is, the core processes are not individual methods or tools, such as regression analysis or control charts, which could be considered "hows". They are called "processes" because they represent the major high-level activities performed in applications of statistics. Virtually all individual statistical methods fit conceptually into one of these processes. Of course, other non-statistical tools and competencies will be needed in the other phases of statistical engineering projects, as we explain shortly. In the typical order in which they are applied, the core processes are:

- Data Collection – proactively obtaining the highest quality data possible for the problem at hand and documenting the data pedigree
- Data Exploration – understanding the data, observing patterns and trends, and beginning to develop or refine hypotheses, based on graphical and numerical methods
- Model Building – developing different types of formal models depending on the data and problem being addressed
- Drawing Inferences (Learning) – considering what broader conclusions can be drawn about the phenomenon of interest beyond this particular data set
- Solution Identification and Deployment – determining the best course of action to take based on what has been learned from the previous processes, deploying it and ensuring sustainability

Note that each of these high-level processes includes a verb – they represent some action, rather than a specific tool. There are many tools to be considered for use within each process. The mix of tools will typically vary for each problem. There is also a set of overarching competencies that is generally needed to achieve success. These competencies are needed not only in the strategy and tactics phases, but rather across all phases of statistical engineering applications. These overarching competencies include project management, teamwork, communication and other competencies are discussed in the chapter on overarching competencies.

1.1.6 Summary of Key Points

Key points that we would like to emphasize from this introductory section include:

- Statistical engineering is not a "buzzword". It has been carefully defined to represent the engineering of solutions to statistically oriented problems.
- Large, complex, unstructured problems are particularly amenable to a statistical engineering approach.
- Statistical engineering emphasizes *integration*, i.e., integration of methods and integration of disciplines.

- There is an underlying theory to statistical engineering that is admittedly a work in progress.
- Part of this theory is a set of generic phases through which most applications of statistical engineering progress. This framework provides general guidance to those applying statistical engineering.
- The overall strategy and tactics utilized in applications will typically involve a series of methods selected from statistical *core processes*, linked with other methods. The specific methods selected will depend on the unique aspects of the problem at hand.

Section 1.1 References

American Statistical Association Undergraduate Guidelines Workgroup (2014) "2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science," Alexandria, VA: American Statistical Association. Accessed at http://www.amstat.org/education/curriculumguidelines.cfm

Antony, J., Hoerl, R.W., and Snee, R.D. "Lean Six Sigma: Yesterday, Today, and Tomorrow", The International Journal of Quality & Reliability Management, 34,7, 1073-1093, 2017.

Box, G. E. P. and Wilson, K. B. "On the Experimental Attainment of Optimal Conditions," Journal of the Royal Statistical Society, Series B, 13, 1–45, 1951.

Box, G.E.P. Hunter, W.G., and Hunter, J.S. Statistics for Experimenters, John Wiley & Sons, Hoboken, NJ, 1978.

Di Benedetto, A., Hoerl, R.W., and Snee, R.D. "Solving Jigsaw Puzzles: Addressing Large, Complex, Unstructured Problems", Quality Progress, June, 50-53, 2014.

Donoho, D. "50 Years of Data Science", Journal of Computational and Graphical Statistics, 26:4, 745-766, 2017.

EMC Educational Services. Data Science and Big Data Analytics, John Wiley & Sons, Hoboken, NJ, 2015.

Fung, K. "The Pending Marriage of Big Data and Statistics", Significance, 22-25, 2013.

Hardin, J., Hoerl, R., Horton, N.J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, D., Temple Land, D., and Ward, M.D. "Data Science in Statistics Curricula: Preparing Students to 'Think With Data'", The American Statistician, 69, 4, 343-353, 2015.

Hockfield, S. J. "Technical Challenges of the 21st Century," Niskayuna, NY: Presentation at GE Global Research, 2010.

Hoerl, R.W. and Snee, R.D., Statistical Thinking: Improving Business Performance, 3rd ed., John Wiley & Sons, Hoboken, NJ, 2020.

Hoerl, R.W. and Snee, R.D. "Statistical Engineering: An Idea Whose Time Has Come?", The American Statistician, 71, 3, 209-219, 2017.

Hoerl , R.W. and Snee, R.D. "Show Me the Pedigree!", accepted for publication in Quality Progress, 2018.

Hoerl, R. W. Snee, R. D., and De Veaux, R. D. "Applying Statistical Thinking to 'Big Data' Problems," Wiley Interdisciplinary Reviews: Computational Statistics, July/August, 221–232, 2014.

INFORMS. Certified Analytics Professional Program and Examination: Candidate Handbook, 6th ed., Catonsville, MD, 2018.

Jordan, M.I. "Decisions and Contexts: On Gradient-Based Methods for Finding Game-Theoretic Equilibria," 2019.

Kendall, J. and Fulenwider, D.O. "Six Sigma, E-Commerce Pose New Challenges," Quality Progress, July, 31-37, 2000.

Lindsay, B. G., Kettenring, J., and Siegmund, D.O. "A Report on the Future of Statistics" (with discussion), Statistical Science, 19, 387–413 , 2004.

Meng, X. "Desired and Feared-What Do We Do Now and Over the Next 50 Years?," The American Statistician, 63, 202–210, 2009.

Nair, V. "Industrial Statistics: The Gap Between Research and Practice,"
Youden Memorial Address, ASQ Statistics Division Newsletter, 27, 5-7, 2008.

Wild, C. and Pfannkuch, M. "Statistical Thinking in Empirical Enquiry," International Statistical Review, 67, 223–248, 1999.

Section 1.2 - History and Background of Statistical Engineering

1.2.1 Objectives

The purpose of this section is to explain the origin and development of statistical engineering.

1.2.2 Outline

This section provides a brief history of statistical engineering with an emphasis on how it began, the key milestones along the way and the leaders who developed and formed the discipline. Statistical engineering developed because there was a need for it, a gap to fill. In particular, some felt that the traditional industrial statistics paradigm was no longer working as well as it should. Further research and application were needed to make statistical engineering a viable and sustainable discipline and continue to this day. This led to the development of statistical engineering.

1.2.3 Initial Use of the Term Statistical Engineering

Churchill Eisenhart (1950), the founder of the Statistical Engineering Division of the National Bureau of Standards (now National Institute for Standards and Technology), first published the term "statistical engineering." However, Eisenhart's focus was more on what we would today call "engineering statistics" – the application of statistics to engineering and physical science problems.

Even earlier, Dorian Shainin utilized the term statistical engineering in the 1940s as a way of using statistical and other methods to solve industrial problems (Shainin 2012). Again, the focus was on the application of statistical tools to the solution of engineering problems. Steiner et al. (2008) provide a detailed discussion of the Shainin methodology from a statistical perspective. Steiner and MacKay (2005) provided a statistical engineering framework for reducing variation in manufacturing processes and discuss numerous examples of its use, principally in the auto industry.

As noted in Section 1.1, ISEA defines *statistical engineering* as: "*the study of the systematic integration of statistical concepts, methods, and tools, often with other relevant disciplines, to solve important problems sustainably*."
This current definition is an expanded view of statistical engineering focusing on performance improvement in all types of organizations, including manufacturing, financial, service, non-profits and healthcare, to name a few. Particular emphasis is on the solution of problems that have a major impact on an organization, i.e., problems that are typically large, complex and unstructured.

1.2.4 Development of the Current View Through Publications and Conferences

A key stimulus development of the current view of statistical engineering was the publication of an article entitled "The Future of Industrial Statistics: A Panel Discussion" (Steinberg 2008).

This article detailed the myriad of issues facing the industrial statistics profession at that time. However, in the view of some readers, not enough solutions to these problems were provided.

After considerable thought and debate, Ron Snee (DuPont, retired) and Roger Hoerl (GE R&D) concluded that a more promising route to success for the profession, in industry and beyond, was to identify and solve problems that have a major impact. Rather than being narrow "textbook" problems with "clean" and precise correct answers, these problems were typically messy, i.e., large, complex and unstructured. They further concluded that neither the statistics nor engineering professions had developed a coherent theory as to how such problems should be addressed.

Building on the work of Steiner and MacKay (2005, 2008), among others, Snee and Hoerl extended the concept of statistical engineering towards the current view. See for example Snee and Hoerl (2009, 2011a,b), and Hoerl and Snee (2009, 2010a,b). They provide a definition upon which the current ISEA definition is based and illustrated how statistical engineering could fill the gap between statistical thinking (Snee 1990, Hoerl and Snee 2020) and statistical tools. They showed how to deploy the tools in such a way as to be consistent with the principles of statistical thinking.

The first public presentation of the current view of statistical engineering to the broad statistics community was at the 2010 Joint Statistical Meetings in Vancouver, Canada. The session was in honor of statistician Gerry Hahn's 80th birthday. In May 2011, NASA held a "Statistical Engineering Symposium" which was attended by more than 150 representatives of the government, academic and private sectors. Engineers, statisticians, scientists, managers and project leaders shared lessons learned, techniques and strategies for improving awareness of the value of statistical engineering. The symposium resulted from the leadership of Peter Parker, a statistician at NASA.

Michael Gilmore, a Presidential appointee as the Director, Operational Test & Evaluation within the Office of the Secretary of Defense gave an opening keynote. Dr. Gilmore served as the senior advisor to the Secretary of Defense on operational and live fire test and evaluation of Department of Defense weapon systems. Several other symposia related to statistical engineering have resulted from this initial event, involving NASA and the Department of Defense, under the leadership of Parker and Laura Freeman of the Institute for Defense Analysis.

In July 2011, Phil Scinto from Lubrizol published "Statistical Engineering Examples in the Engine Oil Additive Industry" in *Quality Engineering* (Scinto 2011). In April 2012, a special issue of *Quality Engineering* appeared due to the forward thinking and efforts of Editor Connie Borror and Special Issue editors Christine Anderson-Cook and Lu Lu. This 350+ page issue is a *tour de force* on the nascent theory and practice of statistical engineering and was made available online as a free download. Included were papers on:

- Foundations of statistical engineering
- Roles for statisticians
- Principles and examples
- Leadership for statistical engineering

- Statistical engineering education guidance
- Nine case studies

In March 2013, a session on statistical engineering was on the program of the first Stu Hunter Research Conference held at Heemskerk, the Netherlands. Stefan Steiner gave a presentation on statistical engineering (Steiner 2014), while Roger Hoerl was a discussant (Hoerl 2014). In 2014, Hoerl, Snee, and Richard De Veaux presented two short courses on the role of statistical engineering in enhancing "Big Data" projects, one at the American Statistical Association (ASA) Applied Statistics Conference in February and the other at the ASA Joint Statistical Meetings in August.

While most of the publications to date on statistical engineering have appeared in quality-oriented journals, more recently Hoerl and Snee (2017) published "Statistical Engineering: An Idea Whose Time Has Come" in the *American Statistician,* a publication of the American Statistical Association. This helped introduce the principles of statistical engineering to a broader statistical audience.

1.2.5 Role of the American Society for Quality (ASQ)

In April 2010, Christine Anderson-Cook, then Chair-Elect of the ASQ Statistics Division, became aware of statistical engineering from the article "Closing the Gap" (Hoerl and Snee 2010b) and decided that this should be an initiative that the Statistics Division promote. Upon stepping into the role of Chair in July 2010, she made statistical engineering a primary initiative for her year as Chair. In January 2011, the Statistics Division published a collection of statistical engineering related papers (http://asq.org/divisions-forums/statistics/quality-information/statistical-engineering).

Since 1999, the ASQ journal *Quality Progress* has published an approximately monthly column called "Statistics Roundtable" (now Statistics Spotlight). Many of the seminal articles on statistical engineering noted above appeared in this column. In 2016, the ASQ Statistics Division published a collection of articles from this column entitled: "Statistical Roundtables: Insights and Best Practices" (Anderson-Cook and Lu 2016). This publication consisted of a collection of "...articles that have stood the test of time and remain relevant, informative, and educational for a broad audience." The first chapter in this publication was entitled "Statistical Engineering," and included 13 Statistics Roundtable articles on statistical engineering. The authors of the articles in the statistical engineering chapter are, in alphabetical order: Christine Anderson-Cook, Alexa DiBenedetto, Lynne Hare, Roger Hoerl, Stu Hunter, Bob Mason, Ron Snee and John Young.

It was the Statistics Division, through Anderson-Cook's leadership, that proposed the special issue of *Quality Engineering* focused on statistical engineering, discussed above, as well as the publication on Statistics Roundtable best practice articles. *Quality Engineering* is also an ASQ publication.

1.2.6 Formation of the International Statistical Engineering Association (ISEA)

At the Fall Technical Conference held in October 2017 in Philadelphia, a handful of individuals interested in statistical engineering held an informal planning meeting. They decided to hold a more formal meeting in December of that same year and to invite a broader group of interested parties. This broader group met in Arlington Virginia in December and developed a plan to accelerate the establishment of statistical engineering as a unique discipline. It was at this meeting that the definition of statistical engineering given above was originally developed. This meeting was singular in that eleven people attended a two-day, unfunded meeting that was not connected in any way with an existing professional society or conference. These were: Will Brenneman, Stephanie DeHart, Laura Freeman, Will Guthrie, Lynne Hare, Roger Hoerl, Dean Neubauer, Pete Parker, Ron Snee, Stefan Steiner and Geoff Vining. Vining organized and led both the Philadelphia and Arlington meetings.

Part of the plan developed at the Arlington meeting was to establish a new professional society focused on statistical engineering. In July 2018, this plan became a reality when the International Statistical Engineering Association was legally incorporated. Vining became the founding Chair of the society. The purpose of ISEA is to advance the theory and practice of statistical engineering, including its inclusion into academic curricula, and to enhance the professional qualifications and standing among its members. These high-level objectives of ISEA are intended to encompass the following:

- To promote unity, effectiveness of effort and ethical professional conduct among those who devote themselves to the theory and practice of statistical engineering.
- To provide for the creation of conferences, conventions and other meetings of its members for the exchange of ideas and experiences in the development, application, and use of statistical engineering principles.
- To create and disseminate a body of knowledge for statistical engineering.
- To facilitate the proper inclusion of statistical engineering in statistical and other professional publications, including textbooks and academic curricula.

More information on the ISEA can be found on its website www.isea-change.org.

1.2.7 Summary: The Work Continues

ISEA continues to grow after starting with 14 founding members, to having more than 250 members in early 2019. The association's first annual Statistical Engineering Summit was held in October 2018, in West Palm Beach, drawing roughly 65 attendees. The 2nd Summit will be hosted by the National Institute for Standards and Technology (NIST) in Gaithersburg, MD, in October 2019. ISEA agreed to assume responsibility and oversight of the Stu Hunter Research Conference, renamed the Stu Hunter Research Conference on Statistics and Statistical Engineering. The plan is to make this the premier research conference on statistical engineering going forward. The University of Amsterdam (the Netherlands), University of Waterloo (Canada), Virginia Tech, software corporation Stat-Ease and Procter and Gamble have immediately seen the value of statistical engineering and have begun supporting ISEA as Corporate Members of the society.

Section 1.2 References

Anderson-Cook, C. and Lu Lu "Special Issue on Statistical Engineering," *Quality Engineering,* Vol. 24, No. 2 (2012).

Anderson-Cook, C and Lu Lu. *Statistical Roundtables: Insights and Best Practices,* Quality Press, Milwaukee, WI, 2016.

Eisenhart, C. "Statistical Engineering," *National Bureau of Standards Technical News Bulletin,* Vol. 34, No. 3, March 1950, 29-40.

Hoerl, R. W. "Discussion of Statistical Engineering and Variation Reduction," *Quality Engineering,* Vol. 26, No. 1 (2014): 61-64.

Hoerl, R. W. and R. D. Snee, "Post Financial Meltdown: What Do Services Industries Need From Us Now?" *Applied Stochastic Models in Business and Industry*, December 2009, pp. 509-521.

Hoerl, R. W. and R. D. Snee, "Moving the Statistics Profession Forward to the Next Level," *The American Statistician*, February 2010, (2010a): pp. 10-14.

Hoerl, R. W. and R. D. Snee, "Closing the Gap: Statistical Engineering Can Bridge Statistical Thinking with Methods and Tools," *Quality Progress*, May 2010, (2010b): pp. 52-53.

Hoerl, R. W. and R. D. Snee, "Tried and True—Organizations Put Statistical Engineering to the Test and See Real Results," *Quality Progress*, June 2010, (2010c): pp. 58-60.

Hoerl, R. W. and R. D. Snee, "Statistical Thinking and Methods in Quality Improvement: A Look to the Future," *Quality Engineering*, 22, 3, (2010d): pp. 119-139.

Hoerl, R. W. and R. D. Snee, *Statistical Thinking – Improving Business Performance*, 3rd Edition, John Wiley and Sons, Hoboken, NJ, 2020.

Hoerl, R. W. and R. D. Snee, "Statistical Engineering – An Idea Whose Time Has Come," *The American Statistician,* Vol. 71, No. 3, (2017): 209-219.

Scinto, P. "Statistical Engineering Examples in the Engine Oil Additive Industry," *Quality Engineering*, Vol 23, No. 2, (2011): 125-133.

Shainin, R. D. "Statistical Engineering: Six Decades of Improved Process and Systems Performance," *Quality Engineering,* Vol. 24, No. 2, (2012): 171-183.

Snee, R. D. "Statistical Thinking and Its Contribution to Total Quality," *The American Statistician*, 44, (1990): 116-121.

Snee, R. D. and R, W. Hoerl, "Turning to Service Sectors," *Industrial Engineer,* October, (2009): 37-40.

Snee, R. D. and R, W. Hoerl, "Further Explanation; Clarifying Points About Statistical Engineering," *Quality Progress*, December, (2010): pp. 68-72.

Snee, R. D. and R, W. Hoerl, "Engineering an Advantage," *Six Sigma Forum Magazine,* Guest Editorial, February (2011a): 6-7.

Snee, R. D. and R, W. Hoerl, "Proper Blending: Finding the Right Mix of Statistical Engineering and Traditional Applied Statistics," *Quality Progress,* June, (2011b).

Steinberg. D. M. “The Future of Industrial Statistics: A Panel Discussion,” *Technometrics,* Vol. 50, No. 2, (2008): 103-127.

Steiner, S. H. and R. J. MacKay, *Statistical Engineering: An Algorithm for Reducing Variation in Manufacturing Processes,* ASQ Press, Milwaukee, WI, 2005.

Steiner, S. H, MacKay, R. J. and J. S. Ramberg “An Overview of the Shainin System™ for Quality Improvement,” with Discussion, *Quality Engineering,* Vol. 1, No. 1, (2008): 6-45.

Steiner, S. H. “Statistical Engineering and Variation Reduction,” *Quality Engineering,* Vol. 26, No. 1, (2014): 44-60.

Section 1.3 – Achieving Success in Each Phase of Statistical Engineering

1.3.1 Objectives

The purpose of this section is to provide guidance on how to successfully conduct each of the typical phases of statistical engineering applications, i.e., to provide keys to success for each phase.

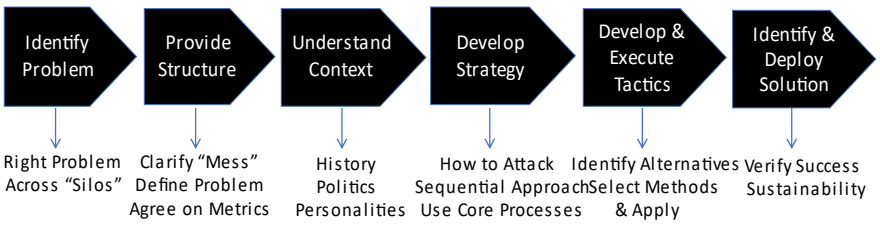
1.3.2 Outline

We first present guidance for conducting each phase of statistical engineering applications, including keys to success, based on experience and the existing literature. We discuss each of the six major phases of statistical engineering in sequence. Next, we discuss how the key principles of statistical engineering tend to apply within these phases.

1.3.3 Guidance and Keys to Success by Phase

As discussed in section 1.1.1, there are no "correct" solutions to large, complex, unstructured issues. Nor is there a "cookbook" that will lead practitioners step by step through successful completion of all such projects. However, Figure 1.3, reproduced below, illustrates the phases that statistical engineering applications typically go through to capitalize on opportunities, particularly with large, complex, unstructured challenges. While statistical engineering applications will generally go through each of these phases, they will do so in unique ways, just as children go through child development in unique ways. For examples of each of these phases, we will repeatedly reference the financial default prediction case study of Hoerl and Snee (2017).

Figure 1.3 The Phases of Statistical Engineering



1.3.3.1 Identify Problem

The first phase of typical statistical engineering efforts is to identify the "real" problem or opportunity. Note that statistical engineering can and should be applied proactively to avoid problems and drive innovation and continuous improvement. We use the word "problem" in a generic sense to indicate any opportunity, issue or problem.

We refer to the "real" problem because teams often work on symptoms rather than underlying causes of problems. This rarely works well. As a simple illustration, suppose an organization realizes that it is constantly recruiting, not because of growth, but because of high turnover. Turnover is costly, not only in terms of the recruiting costs required to replace employees who left, but also because the work of the organization is constantly in transition from one group of team members to another. So, the organization might identify "employee turnover" as the problem to be solved. This might lead to "solutions" to turnover, including the offering of higher salaries, better benefits and more employee "perks," such as free coffee, lunches, etc.

While these efforts may be of some benefit, we argue that high turnover is typically just a symptom of a deeper problem. The key question to ask is why employees are leaving? Digging deeper might lead the organization to realize that it has a leadership problem; employees do not feel that the leaders of the organization have a clear vision and strategy to succeed long term. Perhaps women or ethnic minorities do not feel welcome and valued in the organization. There could be many underlying causes, but typically high turnover, low morale, absenteeism, etc., are just symptoms of the fundamental problem.

Further, within large organizations teams often work on the portion of the problem that is seen within their "silo," rather than the entire problem. In Section 1.1 we discussed the classic example of organizations managing their overall order fulfillment system, from taking orders to ultimately delivering the product to customers. When teams work on this problem within their own "silos," such as sales, manufacturing and distribution, they generally end up working against each other in a "tug of war" or "tractor pull," with little or no improvement to the entire system. Clearly, the overall order fulfillment system is a large, complex system. Therefore, a *key to success* in this phase is to apply a systems perspective and look for the entire problem, which typically goes beyond the issues seen within one department.

A system is "a set of processes that work together to accomplish an objective in its entirety" (Hoerl and Snee 2020). "In its entirety" means that the system includes all related and necessary processes. Problems often arise when the overall objectives are not understood or shared universally. For example, salespeople may have little understanding of all the processes that go into producing and delivering product to fulfill an order. The same lack of system vision might be true of people in manufacturing, distribution, customer service and other areas. There may not be a single employee in the organization who understands the entire system from end to end. This is why correct identification of the fundamental and entire problem is a key to success; it is much easier said than done.

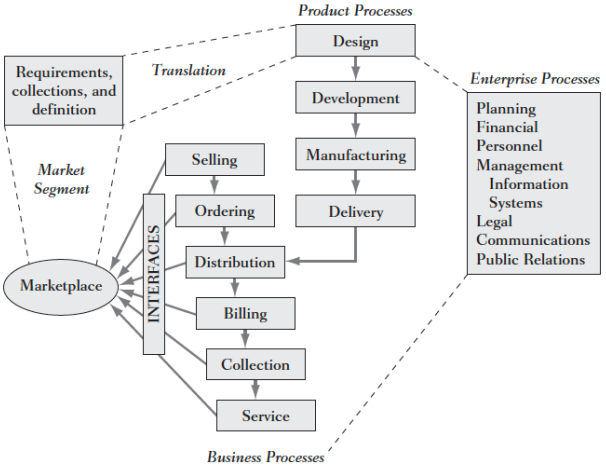
An initial step towards identifying the right problem is to document the process where the problem was first identified, in "distribution" for example. Next, it is usually helpful to consider to what wider system this process belongs. System identification is critical. Then, by working with a larger, more cross-functional team, it should be possible to document the overall system, at least at a high level, and clarify the entire problem. While not needed for small problems, such a systems viewpoint is critical in large, complex, unstructured problems. In the example noted above, the overall system would be the order fulfillment system, and the fundamental problem may be that it is inefficient, possibly involving both missed deliveries and excess inventory. Admittedly, we have not yet rigorously defined these terms; however, they will need to be clarified in the next phase, in which we provide structure to the problem.

Figure 1.4 shows the major systems at IBM Europe (from Hoerl and Snee 2020). Note that IBM views their overall business to be made up of three major systems:

- Business processes
- Product processes
- Enterprise processes

IBM's business processes system is roughly equivalent to what we previously referred to as the order fulfillment system, although it also includes billing and after-sales service. Kauffman (1980) provides a more detailed discussion of systems thinking.

Figure 1.4 IBM Europe Core Processes



Returning to the default prediction case study of Hoerl and Snee (2017), the initial issue was a shocking loss of $110 million by GE Capital on bonds of WorldCom, which went into bankruptcy. This obviously created a "mess", with lots of political pressure to find a "quick fix". However, rather than simply fire the financial analysts involved and declare the issue over, GE Capital dug deeper and realized that there was a much broader problem related to how GE Capital bought, managed and sold securities. After some discussion and clarification, it became clear that the real problem was the need to predict companies likely to go into default, ahead of the market. The WorldCom fiasco could have occurred with numerous other securities.

To summarize, the challenges in this phase are to look beneath the immediate symptoms to identify the "real", or fundamental problem, and to consider this problem in its entirety, which will typically span organizational units or silos. *The key to success in this phase is the ability to view issues from a systems perspective.*

1.3.3.2 Provide Structure

Once the right problem has been identified, it must be properly structured. We quoted Tort above, who noted that what we typically see initially is often a "mess". It is virtually impossible to solve a "mess". The process of converting a mess into a problem is what we call *providing structure*.

This is not nearly as easy as it might seem, because there is no "formula" or "algorithm" that enables people to convert a mess into a structured problem. Rather, this requires the application of *critical thinking*. Critical thinking can be defined as: "the objective analysis and evaluation of an issue in order to form a judgment." Note that in this definition, the word "objective" is key. That is, critical thinking does not start with an assumed answer and work backwards, but rather objectively reviews all data on the issue to develop an informed opinion. This opinion leads to a judgment, which in this case is a more precise statement of the problem.

While most people would claim that their own opinions are objective, psychology tells us that this is rarely the case. As an obvious example, if one asks individuals of diverse political viewpoints why so many governments around the world struggle to run balanced budgets, the answers are likely to be quite different. Some respondents are likely to suggest that the governments are not collecting enough revenue and need higher taxation, while others will suggest that the governments spend too much and need to reduce their budgets. Both viewpoints could be considered true mathematically, but people's political leanings are likely to impact their judgment on this issue.

The same phenomenon occurs in business and industry, science, education, sports and many other aspects of society. In many cases people will develop relatively fixed positions on technical or organizational issues and are often not willing or able to view them from a different perspective. They have "put a stake in the ground" and are not willing to move from it. Working through the mess of issues, opinions, political pressure, and perhaps true crises, requires clear and objective thinking! Otherwise, the statement of the formal problem may embed a pre-determined solution.

For example, a problem statement such as: "Document evidence that excess surfactant in the reactor is leading to the recent low yields in the process" assumes that excess surfactant in the reactor leads to low yields. Whenever a presumed solution is embedded in the problem statement, there is clear evidence of a lack of critical thinking. A much better problem statement would be: "Identify and correct the root causes of recent low process yields."

Critical thinking also enables teams to think through problems *logically*, from symptom to root cause. Emotional thinking generally leads people to jump to conclusions, as in the problem

statement noted above, while critical thinking enables them to reason objectively based on the available data and theory. Therefore, to properly structure the "mess" into a formal problem, critical thinking is necessary. Further, use of formal frameworks such as the SIPOC model can be very helpful. As shown below, the acronym SIPOC stands for suppliers; inputs; process; outputs; customers.

Suppliers -> Inputs -> Process -> Outputs -> Customers

Any process, physical or otherwise, tends to have each of these elements. For example, consider processing loans in a bank. The customers are individuals or businesses requesting loans. The loan decision itself, and the terms associated with it, would be the outputs. The process would consist of the individual steps that a financial institution goes through when approving loan applications, i.e., the information that would go on a flow chart or value stream map. For example, there might be online applications that are initially evaluated by computer models. A loan officer, who looks at specific criteria to make a decision on the application, may then review those applications prioritized by the computer models.

Depending on the size of the loan requested, an underwriter or higher-level authority might have to review and approve the loan, perhaps modifying the terms in the process. Inputs that might be required in this case include credit scores, obtained from third parties, which would be suppliers, and possibly input from employers or landlords to confirm the information on the loan application. Sometimes the customers or suppliers may be other groups within the same organization, such as manufacturing being a supplier to distribution. The key point is that the SIPOC model is not limited to industrial process, but also applies to healthcare, finance, academia, and all other processes.

The SIPOC model helps apply critical thinking by providing a framework in which to make sense of the "mess" and structure it as a problem. Going back to our order fulfillment example from Section 1.1, we may initially be trying to wade through customer complaints or lost customers due to late shipments, high levels of finished product rejected for quality reasons, high process downtime and in-process inventory, sub-standard raw materials, and unreliable suppliers who are frequently late with shipments. How can we attack this mess? By looking at this situation from a SIPOC perspective, we can see that the fundamental issue is that the process is not producing the outputs that we intend it to. This problem is leading to customer complaints about late shipments. Further, some of the causes of late shipments are:

- The internal process, which has high downtime and in-process inventory
- Poor raw materials coming from suppliers
- Unreliable suppliers, in terms of receiving raw materials on time

The SIPOC model will help us realize that we need to focus on producing the product output, both in terms of quality and throughput that is intended. If that does not satisfy customers, then we have misunderstood their requirements and need to meet with them to address this issue. To improve our outputs, we need to both look inside our own operations, to address the downtime and in-process inventories and begin working with suppliers to improve both their timeliness and quality. So, the SIPOC model helps separate cause from effect and focus our improvement

efforts. We now have a rough problem, not just a mess. Once the immediate process is clarified, we may still need to consider the overall system of which this process is part. That is, we may need to continue to apply systems thinking.

To turn this rough problem into a formal problem that a team could attack, we need specific objectives with defined metrics, clarification of any constraints (including financial or time constraints) and possibly answers to a few other questions. These are elements that are typically included in the project charter, created in the Define phase of a Lean Six Sigma project (Snee and Hoerl 2018). In the case of statistical engineering, however, the problem will likely be larger and more complex than a Lean Six Sigma project. Therefore, problem definition (structure) is a phase in statistical engineering. More effort needs to go into understanding the context of the problem, after which an overall strategy for attacking the problem can be developed. Note that for Lean Six Sigma projects, DMAIC is essentially the standardized strategy used to attack problems. So, what Lean Six Sigma does in one phase for small to medium sized problems, statistical engineering does over three phases for large, complex, unstructured problems.

In defining metrics, the team might decide to focus on reducing process downtime. However, there are numerous ways that "downtime" can be measured. For example, some organizations include scheduled maintenance as part of downtime, others do not. How should downtime be quantified if part of the process is down at times, but other parts are still running? Clearly stating that "downtime" is a key metric is not helpful until the term has been clearly defined. "Operational" definitions are those that mean the same thing to different people at different times (Deming 1982). The key items, but generally not the only items, that need to be documented to adequately structure the problem include:

- Overall objectives, including priorities if there are multiple objectives
- Scope of problem (what is "in scope" versus "out of scope")
- Operationally defined metrics
- General timeline anticipated for addressing the problem
- Any important constraints (financial, legal, organizational, etc.)

Note that at this point the team is documenting the problem, not yet planning a solution. Therefore, the timeline and constraints are only needed to ensure that everyone is on the "same page". Once the context of the problem is better understood and an overall strategy to address the problem developed, which are the main outputs of the next two phases, these aspects of the problem will need to be reevaluated.

Applying this concept to the default prediction case study discussed earlier, the team needed to first define some key terms, such as "default", which does not have a generally accepted definition in the financial literature. Next, specific objectives with defined metrics were developed and agreed upon, to clarify the joint objectives of identifying defaults ahead of time (with at least 3 months' notice) and also avoiding "false alarms", i.e., to quantify Type I and Type II errors. The scope was narrowed to predicting defaults for US, public, non-financial institutions. A practical constraint was the need to transfer to GE Capital something that would be easy for analysts not trained in statistics to understand and utilize. Further, it was agreed that

this would be a tool to aid the traders, not a replacement for them. That is, it would not be an automated trading system. The team now had a structured problem that it could attack.

To summarize, keys to success in the "Provide Structure" phase include:

- *Critical thinking*
- *Systems thinking*
- *Use of formal frameworks, such as SIPOC, to aid in structuring the problem*
- *Operational definitions and metrics*

1.3.3.3 Understand Context

The next phase focuses on understanding the context of the problem, which is one of the fundamental principles of statistical engineering. As noted in Section 1.1, large and complex problems have resisted solution for a reason; "obvious" solutions do not generally work. Only after the problem's history, previous improvement efforts – including why they failed, and the technical, political and social background of the problem are properly understood, can the team develop a viable approach to a solution. This usually requires a lot of hard work, which may not be needed for more straightforward problems but is absolutely necessary for large, complex, unstructured problems.

Consider the problem of global poverty. There is virtually no one on earth who is "in favor" of poverty. Trillions (dollars, Euro's, etc.) have been spent to alleviate poverty around the globe. Unfortunately, however, poverty still exits and continues at unacceptably high levels. This is not to say that nothing has been accomplished in the war on poverty, just that true success has not yet been achieved. Those who have dedicated their careers to alleviating global poverty, such as Banerjee and Duflo (2011), make it clear that this is a large, complex, unstructured problem. For example, someone living on $10,000 a year in the US would be considered in poverty, while someone in Bangladesh on the equivalent of $10,000 a year would not. So even defining poverty in a meaningful way is complicated.

As numerous authors have pointed out, including Banerjee and Duflo, many well-intentioned efforts to reduce poverty have failed. Others have fortunately succeeded, at least to some degree. The efforts that have succeeded have tended to be based on a deeper understanding of the context of poverty. That is, these social scientists developed a better understanding of why people in a particular area are poor, what they can and cannot control in their lives, what role the government plays in their lives, and local culture.

For example, consider the relationship between population growth and poverty. Numerous social scientists have suggested that a first step to addressing poverty is controlling population growth via family planning (Banerjee and Duflo 2011). This has led several countries, including India in the 1970s and China more recently, to implement somewhat draconian policies aimed at population control. However, as Banerjee and Duflo note (p. 106): "The problem is that it is impossible to develop a reasonable population policy without understanding why some people have so many children. Are they unable to control their own fertility (due to lack of access to contraception, for example), or is it a choice? And what are the reasons for these choices?"

Clearly, if one wishes to attempt to address poverty through population control, understanding the *context* of population growth is an absolute prerequisite.

The key point here is that successfully addressing large, complex, unstructured problems in business, science, academia, industry, etc., also requires a solid understanding of the problem context. For example, suppose a corporate R&D center wishes to drive more research that provides true technical breakthroughs, rather than incremental improvement. This is clearly a large, complex, unstructured problem! How should we even define "breakthrough" research? There is no "cookie cutter" approach. Further, the approach that might work at one organization, say Google, might not work at a different organization, such as Boeing. These are very different organizations in very different businesses with very different cultures.

Regardless of the specific organization, leaders of the effort would need to spend time understanding the R&D context to have a high probability of success. For example, what is the current state of research? Why is that the case? What cultural or political issues have resulted in this current state? In many research organizations, for example, there is little or no tolerance for failure. Failing on a research project might sabotage one's career. When this is the case, it is to be expected that project leaders and researchers would feel more comfortable focusing on incremental improvement, where ultimate success is much easier to control. Further, leadership may have unknowingly, or perhaps to save money, hired researchers who do not have the technical expertise to perform more fundamental research.

A great deal of "digging" will likely be required, both among the researchers and the leaders of the organization, to understand the root causes of the current lack of breakthrough research. Points of view external to the organization, perhaps from peers or competitors, might be necessary to get an unbiased view. Ongoing use of *critical thinking* is necessary to obtain this deeper understanding and avoid jumping to "obvious" conclusions that will not lead to success. Because this problem is likely to cross organizational boundaries, use of *systems thinking* would be important to see how recruiting processes, funding processes, reward and recognition process, and research processes all connect to form a system.

Further, the ability to ask *neutral and open-ended questions* is critical. There is an art to asking questions. Questions that can be answered "yes" or "no", and questions that point towards a certain answer, such as those that begin with: "Don't you agree that…", are not effective. Rather, the questions need to be neutral, not pointing towards any particular answer and also open-ended.

Examples of such questions are: "Based on your experience in this organization, what would you say are the highest research priorities?"; "What degree of risk do you feel is appropriate to take on a long-term research project?"; or "What attributes are most highly valued in candidates for employment?" Answers to these questions are more likely to lead to a deep understanding of the context to the breakthrough research problem. Of course, they take time and require an open mind, one that does not have a pre-determined "solution".

An old saying goes: "Whenever you have more than one person involved, there will be politics." All organizations have political issues to some degree. Of course, some present greater obstacles

than others. Woodrow Wilson was President of Princeton University from 1902-1910, and then became President of the United States in 1913. It is rumored that when asked why he left Princeton to run for US President, he replied, "I couldn't take the politics anymore." Whether Wilson stated that or not, proper understanding of organizational context requires a willingness to dig into the politics. Unfortunately, politics are often "unstated", and it is often easier to pretend that they do not exist. Therefore, to accurately identify the underlying political context requires careful questioning. The anticipated timeline, skills needed and constraints identified in the "Provide Structure" phase should be updated once the context is clarified.

To understand the context in the default prediction problem discussed previously, the team spent significant time reviewing the literature of default prediction, because it is a classic problem in finance. In fact, researchers in this field have won several Nobel prizes in economics, including Black and Scholes (1973). Further, there were several commercial default prediction systems available on the market. The team formally evaluated these to understand how they worked and how well they could predict. They took a "field trip" to Wall Street to speak with several financial institutions about these issues, in addition to interviewing financial analysts and executives at GE Capital. In these interviews, they discovered some political pressure to use a commercial system, at least as part of the solution, since GE Capital had already paid a significant amount of money to have access to one.

To summarize, keys to success in the "Understand Context" phase include:

- *Critical thinking*
- *Systems thinking*
- *Asking neutral and open-ended questions*
- *Recognition of the cultural and political environment*

1.3.3.4 Develop Strategy

Once the fundamental problem has been properly identified, the "mess" has been converted to a formal problem statement - with defined metrics, and the context of the problem is properly understood, the team is in position to develop a *strategy* to address it. A strategy can be defined as: "a plan of action or policy designed to achieve a major or overall aim". As noted in Section 1.1, such a "plan of action" is not typically needed for straightforward problems. Rather, diagnosis of the "correct" method or approach, often provided in textbooks, will suffice. However, large, complex, unstructured problems can rarely be solved with a single method. Rather, multiple methods and perhaps multiple disciplines will be required, and these need to be integrated into an overall "plan of action". Further, this approach will likely need to be sequential in nature, as it involves multiple methods.

But how should one go about developing an overall strategy to attack such problems? While there is no simple "three step method" to developing sound strategies, there are some principles available that have been proven effective. A strategy is basically a "road map" that determines the most logical path to get from where one is, to where one wants to be – the "plan of action" to achieve an "overall aim". In sports, where one wants to be is clear – on the winning side of the score. However, to determine how to get there, the team needs to honestly document where it is

(its strengths and weaknesses, as well as the obstacles to be overcome) the opponent. So, developing strategy is basically determining how to get from point A to point B, with recognizing that there are obstacles in the way.

It should be clear that *systems thinking* (understanding how all the pieces fit together) and *critical thinking* (being able to think objectively and rationally about a problem) are again critical. While *creativity* is always advantageous, it is particularly important when developing strategy, because almost by definition there is no "cookie cutter" answer to developing strategy. A creative solution needs to be engineered. Also, effective strategies for big problems typically involve a technical component and a non-technical component. These will need to be integrated into one overall strategy.

The non-technical component is needed because of the importance of culture, politics and organizational issues in virtually all environments. Dealing with the non-technical aspects often involves such things are forming cross-function teams, which can also help technically, communicating broadly, meeting with stakeholders individually to build trust, and other human factors. These are aspects of *project management* and *teambuilding* in general. We address these skill sets in Chapter 2, which discusses overarching skills needed for statistical engineering.

If we ignore the non-technical complexities of the problem, there is little chance that a technical solution, even a very good one, will work. In fact, for some problems, the non-technical aspect is the most challenging, such as in highly politicized environments. Finding a technical solution might be easier than getting consensus to implement it. When this is the case, more sophisticated approaches to *organizational effectiveness* will likely be required, as explained in Weisbord (2012). This 2012 publication is a 25-year anniversary 3rd edition, building on the original 1987 classic on organizational effectiveness. While it contains too much detail to cover here, one of Weisbord's main points is that the history of improving organizational performance has gone through four major revolutions.

The first was Frederick Taylor's "scientific management" around the turn of the 20th century, which suggested that industrial work could be scientifically studied and improved. Taylor's work was so influential that it led to the development of the discipline of industrial engineering. Weisbord refers to this phase as "experts solving problems". Workers and others learned from the experts, the engineers, and did as they were told. While making drastic improvements, a limitation of Taylorism is that it wastes the intellects of most workers. Therefore, the next revolution was a move to "everyone solving problems", which became a popular approach in the decades after World War II. Worker involvement became critical and led to such initiatives as "quality circles" in Japan and self-managed work teams in the US. However, the focus was on solving fairly straightforward problems, not improving the overall system. Lean Six Sigma could be considered an example of this more inclusive approach to solving problems.

The next revolution brought in systems thinking and involved cadres of "efficiency experts" who came into organizations, studied the entire work system, and made radical changes, often resulting in downsizing. Weisbord refers to this phase as "experts improve whole systems", which was a major movement in the 1960s and 1970s. As with Taylorism, the individual worker was typically left out and perhaps became a victim of the final system. The last revolution noted

by Weisbord was a move to "involving everyone in improving whole systems", i.e., getting the entire workforce involved not only in solving smaller problems, but also in improving the overall system. Weisbord (2012) provides several examples of organizations that involved employees in completely redesigning entire work systems in the 1980s and more recently.

For the technical aspects of the problem, the team needs to think through the methods that are likely to be needed and in what logical order they should be applied. The technical methods could involve statistical as well as non-statistical approaches. Figuring out how to do this for a complex issue may seem like an insurmountable problem, so again *creativity* is needed. Fortunately, statistical engineering has benchmarked another engineering discipline, chemical engineering, to find a useful approach. As explained in Section 1.4 of this chapter, which focuses on use of the Core Processes of statistical engineering, during the development of the discipline of chemical engineering, engineers recognized that chemical plants tended to be made up of a sequence of common operations. These are referred to as "unit ops" in the chemical engineering literature.

A unit operation involves a physical change or chemical transformation. For example, mass transfer, reaction and heat transfer are clear examples. Depending on the source used, there are roughly seven unit operations in chemical engineering. There are many ways that each operation, such as reaction, could be performed. "Reaction" is really a category, not a well-defined step. Polymerization would be one possible method of reaction, and fermentation is another. The advantage of documenting the unit operations is that is significantly simplifies the design of a chemical plant. Fundamentally, engineers need to answer two macro questions:

1. What sequence of unit operations would take the available inputs and convert them to the desired outputs?
2. For each unit operation, e.g., reaction, what specific type of reaction would be best in this case?

This same pair of questions is the key to designing everything from a micro-brewery to a glass factory making covers for smart phones to a recycling center for paper or plastics. All these processes follow the same general model, utilizing the two macro questions noted above, although the specifics will be quite different. See Section 1.4 for more detail on the unit operations from chemical engineering.

By utilizing this same unit operations (or "unit ops") approach from chemical engineering, statistical engineering can be considered to have five "unit ops", as well as a set of overarching competencies that apply throughout statistical engineering efforts. We refer to these as the five "core processes" of statistical engineering, rather than unit operations, because they do not involve physical or chemical transformation. As explained in Section 1.1, the core processes of statistical engineering are:

- Data Acquisition – proactively obtaining the highest quality data possible for the problem at hand
- Data Exploration – better understanding the data, observing patterns and trends, and beginning to develop or refine hypotheses, based on graphical and numerical methods

- Model Building – developing different types of formal models, depending on the data and problem being addressed
- Drawing Inferences (Learning) – considering what broader conclusions can be reached about the phenomenon of interest beyond this particular data set, based on the analyses performed
- Solution Identification and Deployment – determining the best course of action to take based on what has been learned from the previous processes, deploying it and ensuring sustainability. Additional data or looping back to previous core processes may be required here.

The remainder of this handbook is organized such that each chapter focuses on one these core processes, as well as one chapter on the set of overarching competencies.

While there are a very large number of statistical methods and tools available, from designed experiments to linear and non-linear models to confidence intervals, etc., we argue that each of them performs one of these core processes. For example, designed experiments are a means of obtaining high pedigree data. Most graphical methods help explore the data. Model building is used to develop models. Statistical inference tools, such as confidence intervals, help analysts understand what conclusions can be reasonably reached about an entire population based on analysis of a sample. Eventually someone has to utilize the analysis to deploy a solution in order to solve a problem. In terms of statistical methods in this last core process, control charts can help ensure that a solution is sustainable over time, by quickly identifying any deterioration in process performance.

This framework can help develop a viable strategy and accompanying tactics, by reducing much of the work to answering the *two macro questions*:

1. What sequence of core processes would most likely solve the problem at hand, given the context?
2. What specific statistical or other methods would be most effective for each core process?

There is an important advantage in developing strategies in statistical engineering in contrast to chemical engineering; in almost all cases the ordering of the core processes will follow the order listed above. This is not the case in chemical engineering.

While the core processes are listed in their logical sequence, complex problems may require that some processes are used more than once, or that subsequent processes need to be modified based on earlier processes. This is also the case in chemical engineering. So, the core processes should not be viewed in a linear, step-by-step manner when dealing with large, complex, unstructured problems. For example, suppose that a team intends to fit a linear regression model in "Model Building", but in "Data Exploration" it notices time-dependent data, and so it may decide to fit a time series model instead. Conversely, it may decide that it needs to go back and collect more or different data.

Therefore, the team working on the problem should give some thought to the specific methods to be employed for each core process, but the final decisions will be made in the next phase,

"Develop and Execute Tactics". The focus in "Develop Strategy" is on question 1 – the core processes, while the focus in "Develop and Execute Tactics" is on question 2. We comment further on the utilization of core processes in Section 1.4.

Note that while the two macro questions related to core processes are critically important, as is considering both the technical and non-technical aspects, developing a successfully strategy will likely involve other considerations as well. For example, what other methods, beyond statistics, might be appropriate? Which specific skill sets or individuals are needed on the team? Is there an existing framework, such as DMAIC from Lean Six Sigma, which might work for the overall strategy? If there are multiple objectives, how should these be prioritized? The anticipated timeline, skills needed and constraints identified in "Provide Structure", and potentially revised in "Understand Context", should be finalized at this point. Therefore, simply listing the core processes noted above does not constitute a strategy.

Applying this to the default prediction case discussed previously, that team developed an overall strategy at this point, involving both the organizational and technical aspects of the problem. Major elements of the strategy, in roughly sequential order, were:

- Forming a cross-functional team, involving GE Capital, GE Global Research in the US, and GE Global Research Bangalore (India).
- Incorporating diverse skill sets on the team, including statistics, quantitative finance, economics, machine learning and computer science.
- Managing expectations, to make sure senior executives did not have expectations for a "quick fix" or perfect solution.
- Recognizing the need to obtain high pedigree data as an initial step. Since GE Capital was built on acquisition, it did not have an existing data set appropriate for this project. Further, high pedigree data sets in finance are generally proprietary and expensive; hence the team recognized that it would take time to find the right data at the right price. (Data Acquisition)
- Figuring out how to augment this data over time without having to continue paying for the data, such as thorough web-scraping techniques. (Data Acquisition)
- Considering the possibility of utilizing probability of default from a commercial default predictor as a starting point, both for technical and political reasons, and then augmenting it with a "slope" or "momentum" metric. This would speed up the project, be more likely to be accepted, and could still provide a competitive advantage. (Model Building)
- Finding a filtering methodology to help with the slope metric, given the volatility of financial data over time. (Data Exploration and Model Building)
- Comparing traditional statistical methods with simulation and machine learning techniques, to determine which approaches might do the best job of predicting default from the probability of default and slope metric. (Model Building and Drawing Inferences)
- Figuring out how to present the final model and its results to executives. (Solution Identification and Deployment)

- Developing a control plan to test over time if the model were losing predictive capability due to changes in the economy. If the model were deteriorating, this would signal the need to reevaluate it. (Solution Identification and Deployment)

In summary, there is no "cookbook" approach to developing a winning strategy. However, the following are keys to success in this phase:

- *Critical thinking*
- *Systems thinking*
- *Creativity*
- *Organizational effectiveness understanding and tools, including project management and team building*
- *Consideration of the two macro questions related to core processes*

1.3.3.5 Develop and Execute Tactics

Once the team has developed an overall strategy, it needs to develop and employ tactics to implement the strategy. A strategy, while critically important, is just a high-level plan. Tactics are more detailed elements of the overall strategy that provide specific direction at the operational level. In our order fulfillment example discussed in Section 1.1, we noted that if the strategy included a decision to prioritize customer fulfillment over inventory reduction, we would still need specific methods for fulfillment. In the tactics phase, we would figure out specific methods to fulfill orders more consistently, and then deploy these in operations to see how well they worked. Tactics will generally involve selection of individual statistical and other methods within each of the core processes.

Table 1.2 shows some of the statistical methods that are frequently applied to achieve the goals of each of the core processes. In other words, these methods are among those that should be considered when answering macro question 2, as to which specific methods should be selected. Table 1.2 should not be viewed as an exhaustive list, but rather as a set of commonly used tools, especially since non-statistical methods may also be required. That is, *integration of non-statistical tools*, often related to information technology, is also critical.

A key point noted in Section 1.1 is that statistical engineering is "tool agnostic", meaning that it selects tools based on the specific need, rather than pre-selecting "favorite" tools. Therefore, *critical thinking* must be applied to select the most appropriate tools for each core process. This decision will depend heavily on the specific context of the problem, the overall objectives, and the pedigree of the data that have been collected.

Table 1.2 Common Methods Utilized in Each Core Process

Data Collection

- Data collection protocols
- Observational data collection
- Automated data collection
- Design of experiments
- Data Pedigree documentation
- Querying databases
- Database integration

Data Exploration

- Data cleaning and manipulation
- Exploratory data analysis (EDA)
- Visualization methods for high-dimensional data
- Statistical process control (Phase I – diagnosis of stability)

Model Building

- Linear models (regression and ANOVA)
- Generalized linear models
- Bayesian models
- Fixed and random effects models
- Predictive analytics models (machine learning)
- Time series models
- Reliability models
- Applied mathematical models (quadratic programming, etc.)
- Model verification and validation (residual analysis, out of sample prediction, etc.)

Drawing Inferences (Learning)

- Point estimation methods
- Interval estimation methods (e.g., confidence intervals)
- Hypothesis testing

Solution Identification and Deployment

- Human factors engineering
- Piloting of solutions
- Control plans
- Statistical process control (Phase II – monitoring of stable processes)

The tactics will then need to be implemented, both on the technical and non-technical side, in the appropriate order, based on the strategy. It is important to avoid the pitfall of "paralysis by analysis", and therefore a *bias towards action* helps. The results from initial core processes may change the ensuing tactics, just as when sports teams find themselves way behind at halftime they may "ditch the game plan" and start over, or perhaps make less dramatic halftime adjustments. Additional analyses, or perhaps additional data may be required. The team may need to loop back to earlier core processes, or even earlier phases of statistical engineering. As noted several times, addressing large, complex, unstructured problems is rarely a linear process. Therefore, *flexibility* in executing the tactics is important.

Applying this to the default prediction case, the team developed more detailed tactics to carry out the overall strategy. For example, on the organizational side, they divided up the main tasks between the groups in US research center, the Indian research center and GE Capital. They set up weekly conference calls to compare notes and determine next steps. On the technical side, they chose "KMV" (now "Moody's KMV") as the commercial system to obtain initial probabilities of default. A proprietary smoothing algorithm was selected, and then the slope metric was calculated on the basis of the smoothed default probabilities. The computer scientists on the team found websites they could scape to obtain new data, and they eventually created a direct feed from Wall Street.

A model based on classification and regression trees (CART), which would be considered a machine learning approach, outperformed traditional statistical approaches, such as time series models and Monte-Carlo simulation. The CART model was modified based on results from a Markov Chain analysis, which revealed that the system had memory, i.e., if a company had been in default in the past, it was more likely to go back into default, all other things being equal. The team invited a new member to join at this point, a specialist in censored data analysis, to consider how to monitor prediction accuracy in real time, given that there was a three-month window in which companies could default.

In summary, key success factors in "Develop and Execute Tactics" are:

- *Critical thinking (being tool agnostic)*
- *Integration of non-statistical tools*
- *Having a bias towards action*
- *Flexibility*

1.3.3.6 Identify and Deploy Final Solution

In the course of applying these methods in a systematic fashion, the team should begin to learn and identify specific actions they could take to address the problem to achieve the overall goals identified in the "Provide Structure" phase. In most cases, these actions will need to be *piloted* to verify that they work, and do not cause unforeseen issues. For example, before "mothballing" a computer system and replacing it with a totally new one, it is usually a good idea to run them in parallel for a while, in case there is a problem with the new system.

If the planned actions do not work as well as anticipated, the team may need to loop back to the strategy or tactics phases. Once a satisfactory solution is identified, it still needs to be implemented. A brilliant solution is worthless if it is not actually implemented. There may be *cultural or political issues that need to be considered* when deploying the solution, such as involving people in the process. Good solutions are sometimes rejected for non-technical reasons.

Consider the famous Netflix competition, in which Netflix gave a $1,000,000 prize to a team that developed the most accurate model to predict customer ratings of movies. Unfortunately, because the context of the problem was not well understood by the participants, the winning model was never implemented by Netflix (Fung 2013). From a purely technical viewpoint the winning model was a success, but from a business viewpoint it was a failure, because it was never implemented.

Even after an identified solution has been successfully implemented, the team still needs to worry about sustainability. As the old saying goes: "It's easy to quit smoking; I've done it dozens of times." Therefore, a "control plan" is typically needed to embed the solution into standard work processes, as well as to identify how the system should be monitored over time, and what steps employees should take when backsliding is detected. Also, even in a best-case scenario, there will be opportunity for further learning and improvement. Therefore, a new improvement initiative or project may make sense, to follow up on the first team's results. This continues the cycle of improvement based on the scientific method.

The default prediction team presented their model as a two-dimensional map, labeled green (buy or hold), yellow (consider selling down), and red (sell). Probability of default and slope were the two dimensions of the map. Further, while much of the team's work was done in the programming languages R and SAS, they created Visual Basic code that could be run from a Microsoft Access database, to allow GE Capital analysts to utilize it, without learning statistical software. These two aspects of the solution, a map labeled red, yellow, and green, as well as use of Microsoft Access, enabled easy transition to the client group at GE Capital.

The default prediction model was validated internally by the team and then externally by GE Capital. The external validation did not use any data that was used by the team to develop the model. Rather, GE Capital ran the model on its own portfolio, simulating what the financial results would have been over the past 8 months if all trades had been based solely on the model. The results were positive in the hundreds of millions of dollars versus the actual results over that time period, providing strong validation for the model. The validation results were so positive

that GE Capital *embedded* the model into its deal approval system. That is, going forward, traders would have to list the color code from the model on any proposal for a major purchase or sale of securities. *Embedding* statistical methods into organizational processes and systems is critically important to ensuring sustainability.

Lastly, the team also provided a *monitoring* tool to the GE Capital client group that evaluated predictive accuracy over time, using censored data techniques borrowed from reliability. The outputs from the censored models were plotted on a modified control chart for easy interpretation. To protect the technology, GE Capital filed US and international patent applications, both of which were approved. Details on this case study can be found in Hoerl and Snee (2017).

In summary, the key success factors in the "Identify and Deploy Final Solution" are:

- *Piloting potential solutions*
- *Consideration of cultural and political issues, as well as technical*
- *Embedding the solution into standard work processes*
- *Monitoring the solution over time to ensure sustainability*

1.3.4 Application of the Fundamental Principles to the Phases

In Section 1.1 we discussed both the typical phases and the fundamental principles of statistical engineering efforts. Now that we have explained the phases in more detail and discussed the keys to success in each, a logical question one might ask is how the fundamental principles relate to the phases. That is, in what sense are they "fundamental", and do they apply equally to each phase?

Recall from Section 1.1 that the fundamental principles of statistical engineering are:

1. Understanding of the problem **context**
2. Development of a problem-solving **strategy**
3. Consideration of the data **pedigree**
4. Integration of sound **subject matter theory** (domain knowledge)
5. Utilization of **sequential approaches**

Principles 1 and 2 are so critical that they constitute their own phases, as we discussed. Consideration of the data pedigree is critical whenever data is collected, plotted or analyzed. This will tend to occur in the "Develop and Execute Tactics" phase, as the strategy including data collection and analysis is deployed. However, if utilization of existing data is being considered, then the pedigree of the data should be documented during the "Understand Context" phase, as the team evaluates the background of the problem. This evaluation will help the team develop an appropriate strategy, in terms of deciding whether these data are sufficient, or if newer or more accurate and precise data are needed.

Subject matter theory, from principle 4, is critically important throughout all phases. Without some level of subject matter theory, it is not possible to even identify the fundamental problem. For example, could someone who has never even seen a rugby match identify the real problem leading to poor performance by a rugby team? Similarly, to properly structure the problem we are likely to utilize the SIPOC model, as discussed in Section 1.3.3.2. However, if the team has no subject matter knowledge, how can it document the process? How could it identify the key outputs? SIPOC is considered a "knowledge-based tool", in the sense that it is generally developed based on subject matter knowledge, rather than data.

Understanding the context of the problem requires subject matter knowledge, to be able to ask logical, open-ended questions. And understanding the context better develops further subject matter knowledge. This phase is essentially about deepening one's understanding of the subject, both technically and non-technically. The next phase, "Develop Strategy", then utilizes this enhanced subject matter knowledge to put together a "game plan" to address the problem. This would be a pointless exercise, of course, without subject matter knowledge. Going back to the rugby example, trying to develop a game plan to improve a rugby team's performance without understanding the game of rugby would certainly be pointless!

As noted above, most of the statistical and non-statistical tools will be applied in the "Develop and Execute Tactics" phase. It is well known that integration of subject matter knowledge is critical to properly apply, and learn from, statistical methods (Box et al. 2005). Lastly, in deploying the final solution, the team needs to understand both the technical and non-technical aspects of the problem, if the solution is to be sustainable. Embedding solutions into work processes, a key to success in this phase, obviously cannot be accomplished without understanding the work process.

Clearly then, subject matter knowledge is critically important to each phase of statistical engineering, which explains why it is one of the fundamental principles. In fact, one could argue that this is a key differentiator between statistical science and statistical engineering, or between statistical engineering and data science, also known as algorithmic modeling or machine learning. Statistical science is the science of the statistical methods, typically documented via mathematics, not the application discipline, whether it is biology, chemistry, business, or something else.

Further, as noted by Breiman (2001) and Shmueli (2010), machine learning approaches to data analysis focus on prediction accuracy, rather than on trying to understand the "physics" of how the input variables (x's) actually relate to the output variables (y's). That is, they do not utilize subject matter knowledge in constructing the model form, nor do they attempt to develop further understanding of the phenomenon under study. Machine learning models may fit well within an overall statistical engineering approach, as in the default prediction case, in which a CART model was used. In such applications, the machine learning model addresses one aspect of the overall strategy, but not the entire strategy.

The use of sequential approaches, principle 5, is implied using a six-phase approach. Solving straightforward problems that have answers one can look up in textbooks do not require

statistical engineering. However, a statistical engineering application will not generally resolve a large, complex, unstructured problem *completely*. In the default prediction case, once the original project was closed, several follow-up projects were investigated, such as developing default prediction models for financial, privately held, or non-US corporations.

In summary, the fundamental principles are integrated well with the phases of statistical engineering. However, they are integrated in different ways. Two of the principles are phases themselves ("Understand Context" and "Develop Strategy"), while subject matter knowledge is utilized across all phases.

1.3.5 Summary of Key Points

Key points that we would like to emphasize from this section include:

- Large, complex, unstructured problems do not have "correct" answers that can be looked up in textbooks. However, there is a theory to attacking them involving a six-phase sequential process.
- Significant effort needs to go into finding the underlying problem, as opposed to symptoms that might appear in individual "silo's".
- A "mess", an unstructured problem, needs to be properly structured before a strategy to solve it can be developed.
- Large and complex problems are not easily "fixed" and require a carefully thought-out plan of attack; i.e., a strategy.
- Statistical tools are most helpful once the team is executing specific tactics as part of the strategy.
- Ensuring sustainability of the final solution should not be an afterthought but based on careful planning.
- Keys to success have been documented for each of these phases, *critical (objective) thinking* and *systems thinking* being particularly important.
- The fundamental principles of statistical thinking integrate well with the six major phases of applications.

Section 1.3 References

Banerjee, A.V. and Duflo, E. Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty, PublicAffairs, NY, 2011.

Black, F. and Scholes, M. "The Pricing of Options and Corporate Liabilities," The Journal of Political Economy, 81, (1973): 637–654.

Box, G.E.P., Hunter, J.S., and Hunter, W.G. Statistics for Experimenters, 2nd ed., John Wiley & Sons, Hoboken, NJ, 2005.

Breiman, L. "Statistical Modeling: The Two Cultures", Statistical Science, 16, 3, (2001): 199-215.

Deming, W.E. Out of the Crisis, MIT Press, Cambridge, MA, 1982.

Fung, K. "The Pending Marriage of Big Data and Statistics", Significance, August, 20-25, (2013).

Hoerl, R.W. and Snee, R.D. Statistical Thinking: Improving Business Performance, 3rd ed., John Wiley & Sons, Hoboken, NJ, 2020.

Hoerl, R.W. and Snee, R.D. "Statistical Engineering: An Idea Whose Time Has Come?", The American Statistician, 71, 3, (2017): 209-219.

Kauffman, D.L. Systems One: An Introduction to Systems Thinking, Future Systems, Inc., St. Paul, 1980.

Shmueli, G. "To Explain or to Predict", Statistical Science, 25, 3, (2010): 289-310.

Snee, R.D. and Hoerl, R.W. Leading Holistic Improvement with Lean Six Sigma 2.0, 2nd ed., Pearson Education, London, 2018.

Weisbord, M.R. Productive Workplaces: Dignity, Meaning, and Community in the 21st Century, 3rd ed., Jossey-Bass, San Francisco, 2012.

Section 1.4 – Utilization of the Core Processes

1.4.1 Objectives

The purpose of this section is to explain the concept of core processes, where they came from and how they apply to statistical engineering projects.

1.4.2 Outline

This section begins with a discussion of the origins of the discipline of chemical engineering, which has many parallels with statistical engineering. In fact, the International Statistical Engineering Association (ISEA) benchmarked chemical engineering when considering the development of a new discipline and professional society. We begin with a discussion of the history of chemical engineering, including when and why it was developed. Next, we discuss the key principle of unit operations in chemical engineering. The following section explains the statistical engineering analogues to unit operations, which we refer to as core processes and are the foundation of statistical engineering.

1.4.3 The Origins of Chemical Engineering

The origins of the commercial chemical industry are in Germany, but the origins of chemical engineering as a discipline are in the United States (Auyang 2003). The history of this evolution is insightful for understanding statistical engineering, and in particular its relationships with statistics, data science, operations research, and subject matter expertise for solving complex problems.

In the mid-19th century, the chemical industry in Germany focused on specialty organic chemicals, especially dyestuffs. Production volumes were small, and the processes were batch. Each process was designed specifically for that particular company's production needs for that particular chemical, i.e., the processes were tailored and "niche." Chemists worked directly with mechanical engineers to design, build and operate that process. The resulting processes produced "one-off" solutions that were not easily translatable to other chemicals. The high profit margins on these specialty chemicals more than allowed for this inefficient approach.

In the late 19th century, the nascent chemical industry in the United States focused on commodity rather than specialty chemicals, which were much larger scale and had relatively small profit margins. Efficiencies in the design, construction, and the operation of the chemical process were essential for maintaining a reasonable profit. Learning from previous projects became critical for organizational success.

Auyang (2003) summarized the differences between the German and American chemical industries in Table 1.3.

Table 1.3 German and American Chemical Industries During the 19th Century

Germany	USA
Economy of scope	Economy of scale
Fine chemicals: dyestuffs, drugs	Heavy chemicals: soda, petroleum
137,000 tons in thousands of dyes	2,250,000 tons of sulfuric acid
Advanced science, small volume	Capital intensive, high volume
Product innovation – chemistry	Production process – engineering
Chemist and mechanical engineer	Chemical engineer
Industrial R&D – proprietary	University R&D – open science

These differences highlight why the production and engineering focus in the United States led to the creation of a new discipline; see the references given below for a more detailed explanation. In short, developing efficient chemical processes required integrating a thorough understanding of chemistry with critical aspects of mechanical engineering. Chemical engineers were neither chemists nor mechanical engineers.

Rather, they were new hybrids. As shall be seen, statistical engineering utilizes statistical methods, but also requires the ability to link and integrate multiple methods to solve a complex problem, which is fundamental to engineering.

George Davis gave a series of twelve lectures in 1887 at the Manchester Technical School, which is recognized as the catalyst for the new chemical engineering discipline. One of the people attending those lectures was Lewis Norton, who later in 1888 initiated the first four-year bachelor's degree program in chemical engineering at the Massachusetts Institute of Technology (MIT). This was designated as Course X (Course 10) based on Norton's notes on industrial practice in Germany and Davis' lectures. The curriculum was a fusion of industrial chemistry, which until that time primarily consisted of "cookbook" approaches to industrial chemical processes, with mechanical engineering. The emphasis was on the engineering. Several universities followed MIT's lead, and developed similar curricula, including the University of Pennsylvania (1894), Tulane University (1894), the University of Michigan (1898) and Tufts University (1898).

The American Institute of Chemical Engineering (AIChE), the leading professional organization for chemical engineering, was founded in 1908. Its founding involved considerable controversy and politics with the older and much larger American Chemical Society (ACS). Some wondered why a new society was needed. In their view, ACS essentially "covered" chemical engineering, eliminating the need for a new society. (Ironically, many statisticians took the same view of ISEA when it was founded in 2018.) Paralleling the organizational genesis of ISEA, AIChE grew out of an initial meeting of a small group of motivated individuals (twelve chemists and engineers) who came together to discuss founding a new discipline. They formed the "Committee of Six" to explore "the possibility of forming a chemical engineering organization." The group met for six months and decided that an organizational meeting was in order. This meeting occurred in January 1908, with the Committee of Six joined by fifteen other chemists and chemical engineers. At that meeting, the group decided to form the new organization whose

first official meeting was June 22, 1908. This meeting was attended by 40 people. William Walker was named the first President of AIChE.

The new AIChE made several strategic decisions to minimize conflict with the older and larger ACS. First, it established very restrictive criteria to join, at least ten years of experience (five with a bachelor's degree), which effectively excluded academic chemists. Second, it focused on how it could complement rather than compete with the ACS. For example, the experience requirement did allow "production chemists," a group not well-represented in the ACS, who tended to be academic chemists, to join AIChE. Third, the new society generally took a very conservative approach to programs and projects to avoid conflict with the ACS. That is, it sought to avoid services provided by ACS, but to focus on novel services not currently available. This approach laid a foundation for cooperation and collaboration between the new chemical engineers and the traditional chemists.

Early in its history, AIChE decided to use academic curricula to define the new discipline. Fundamental to these efforts was a 1915 letter from Arthur Little, MIT alumnus and later the 1919 President of AIChE, to the President of MIT, espousing the concept of "unit operations" to distinguish the new discipline of chemical engineering from other disciplines. This letter produced fruit, in that there is little disagreement today that unit operations is a concept unique to chemical engineering, separate from chemistry.

Chemical engineering's focus on academic curricula to define the discipline soon led to the need for standardization. Inconsistent use of nomenclature and significant variation in course content led AIChE to be one of the first engineering societies to insist on accreditation of academic curricula. In 1925 it issued its first list of accredited academic curricula in chemical engineering. In 1932, it was a founding member of the Accreditation Board for Engineering and Technology (ABET). AIChE insisted on enforced standards of practice for the profession from its relatively early days. These efforts helped to standardize the concept of unit operations, and their definitions.

These highlights of the history of chemical engineering are based on the following references:

- Auyang (2003)
- The History of Chemical Engineering website (www.pafko.com/history/index.html)
- "The First Century of Chemical Engineering" (www.sciencehistory.org/distillations/magazine/the-first-century-of-chemcial-enginering)
- "History – MIT Chemical Engineering" (http://cheme.mit.edu/about/history/).

1.4.4 The Unit Operations of Chemical Engineering

Today, unit operations are the basic building blocks for commercial chemical processes. There are multiple depictions of unit operations in the chemical engineering literature, but most are consistent with the following:

1. Separation
    a. Solid-Liquid
    b. Liquid-Liquid
    c. Solid-Gas (Vapor)
    d. Liquid-Gas
2. Chemical Reaction
3. Solid Size Reduction
4. Fluid Flow
5. Heat Transfer

A critical point is that a unit operation such as distillation (liquid-liquid separations based in differences in boiling points),\ is fundamentally the same for ethanol-water as for petroleum processing. While the details (the "hows") are different, the fundamental purpose (the "what") is the same. Unit operations are the basis for thinking about designing the steps within a chemical process and serve as a useful way to construct the curricula for teaching the subjects. A classic course within the chemical engineering curriculum is the "Unit Operations Lab" where undergraduate students interact with the equipment performing specific unit operations, for example, a distillation column.

While these specific terms and operations may not be familiar to those lacking a background in chemical engineering, there are clear analogies to other disciplines. For example, credit scoring, which is quantitative evaluation of the creditworthiness of an individual or corporation, could be considered a "unit operation" in finance. Diagnosis is a "unit operation" in healthcare. Evaluation, through homework, tests, or quizzes, is a "unit operation" in education.

Unit operations emphasize another critical aspect of chemical engineering: systems thinking. A chemical process is a system of unit operations with the output from one unit (stage) as the input to the next. For example, a typical chemical process consists of the following stages:

- Receipt of raw materials
- Initial processing of the raw materials
    - Size reduction for solids
    - Purification (separation processes)
- Chemically reacting the raw materials
- Purification of final product
- Packaging of final product
- Proper preparation and disposal of waste products

Design of a new chemical plant, an incredibly complicated task, can be greatly simplified by viewing the entire system as a sequence of five potential unit operations, some of which may occur multiple times. Once the system is designed as a sequence of unit operations, then detailed

design involves determining the specifics for each unit operation. That is, if reaction is needed at a certain point in the process, what is the specific type of reaction, using what raw materials, using which catalysts, at what temperature, and so on? Clearly, systems thinking is absolutely fundamental to chemical engineering, as it is to statistical engineering.

The Fifth Edition of the Perry's Chemical Engineers' Handbook (Perry and Chilton, 1973) has been extremely influential to ISEA in shaping how it approached this Statistical Engineering Handbook. Perry's text includes chapters on the unit operations enumerated above. However, it also includes chapters on methodologies fundamental to chemical engineering that transcend the specific unit operations. Such overarching topics include:

- Basic mathematics and statistics, including optimization
- Physical and chemical data
- Process control
- Materials of construction
- Engineering economics

This statistical engineering handbook likewise incorporates some methodologies and skills that transcend the core processes of statistical engineering and tend to be necessary in each of them. Chemical engineering provides a useful reference point for defining the statistical engineering core processes and for providing a context for statistical engineering theory.

1.4.5 The Core Processes of Statistical Engineering

Statistical engineering utilizes the phased approach discussed in Section 1.4.4 to drive improvement and solve challenging problems. In essence, it "engineers" a solution via an overall systems approach, integrating multiple core processes. ISEA utilizes the term "core processes" for unit operations to differentiate itself from chemical engineering, but the concepts are clearly similar. These core processes are the "whats" of statistical science. That is, the core processes are not individual methods or tools, such as regression analysis or control charts, which would be considered "hows." Rather, these core processes represent the major high-level activities performed in applications of statistics. Virtually all individual statistical methods fit conceptually into one of these processes. Other non-statistical tools and competencies are needed in statistical engineering projects. While the specific set of such tools depends on the problem, there is a set of overarching competencies that is often needed throughout a statistical engineering project.

In the typical order in which they are applied, the core processes are:

- Data Collection – proactively obtaining the highest quality data possible for the problem at hand and documenting the data pedigree
- Data Exploration – understanding the data, "wrangling" data to obtain a usable format and structure, observing patterns and trends, and beginning to develop or refine hypotheses, based on graphical and numerical methods
- Model Building – developing different types of formal models, depending on the data and problem being addressed

- Drawing Inferences (Learning) – considering what broader conclusions can be drawn about the phenomenon of interest beyond this specific data set
- Solution Identification and Deployment – determining the best course of action to take based on what has been learned from the previous processes, deploying it and ensuring sustainability

Note that each of these high-level processes involves a verb form – they represent some action, rather than a specific tool. The solution approach will generally utilize a sequence of core processes, for example data collection, data exploration, model building, etc., to create efficient and effective solutions. Re-looping through the sequence multiple times might be required. Initially, the approach will be identified as part of strategy development, and then considered again in the tactics and solution identification and deployment phases.

Data collection, as noted above, can be done in a number of ways, from online acquisition (e.g., web scraping), designed experiments, observational data, historical data, surveys, etc. While each of these approaches has different advantages and disadvantages, they are all methods of obtaining data in the first place, based on our understanding of the problem and our view of what data and analyses might be needed to address it.

Once data are obtained, it is good to think about its pedigree and structure, "wrangle" it into a usable format, and visualize it prior to fitting formal models. The logical sequence is: think, plot, calculate, repeat. This is because graphics force us to see what we were not expecting. There may be surprises in the data that modify our modeling approach. Time series graphs, scatterplots, boxplots or more sophisticated interactive computer plots would all be examples of ways to visualize the data to gain deeper understanding of it and the processes that produced it.

After obtaining and visualizing the data, we are likely ready to fit formal models. These may be traditional statistical models, such as regression, machine learning models (i.e., neural networks, support vector machines), or even models from other disciplines. Once we have the model, we need to draw practical conclusions, i.e., inferences, about the data and the process that produced it. This may or may not be straightforward. For example, many of the machine learning models, multi-layer neural networks ("deep learning") in particular, are notoriously hard to interpret. In essence, we are trying to understand what actionable conclusions we can reach that are beyond the limited scope of this one data set.

The last core process typically applied is to determine and deploy the final solution. Unfortunately, many projects develop models or solutions that are never implemented. No matter how good the model is, it will not have impact unless utilized. This often involves "embedding" the model into the work process. For example, in processes related to consumer credit, it might involve embedding a credit scoring model into the loan approval process, requiring that every application be evaluated via the model. For the improvement to be sustainable, some type of "control plan" must be put into place to ensure that the gains are maintained over time. In the case of the credit scoring model, it would be wise to reevaluate this model for accuracy over time, to ensure that its performance does not deteriorate as economic conditions change. In all such cases, it is essential that the model be updated periodically to assure that it is currently accurate.

In addition to the core processes noted above, certain methodologies and skills, including team and project dynamics, cut across all the core processes of the engineered solution, much like mathematics and engineering economics cut across all chemical unit operations. These overarching competencies include:

- Organizational anthropology (effectiveness)
- Change management
- Organizational collaboration
  - Teamwork and group dynamics
  - Interdisciplinary collaboration
  - Communication skills
- Project management

The core processes are discussed in detail in the remaining chapters.

1.4.6 Summary of Key Points

Key points that we would like to emphasize from this section include:

- The development of statistical engineering has, to some degree, been modeled after the development of chemical engineering.
- Chemical engineering developed as a unique discipline because of a need to be more efficient in determining how individual chemical processes were linked and integrated to produce a product.
- The concept of unit operations was key to this development, providing an overall strategy to design and improvement of chemical plants.
- Statistical engineering has a similar concept, namely core processes, which constitute the major "whats" from statistical science.
- A key challenge in statistical engineering is to determine how to select individual methods and tools (the "hows") from these core processes, and then link them together in an appropriate way to solve the problem at hand.

Section 1.4 References

Auyang, S.Y "Why Chemical Engineering Emerged in America Instead of Germany", http://www.creatingtechnology.org/eng/chemE.htm (2003).

Perry, R.H. and Chilton, C.H. Chemical Engineer's Handbook, 5th ed., McGraw-Hill, NY, 1973.

Section 1.5 – Chapter Summary

Statistical engineering, the engineering of solutions to large, complex, unstructured problems of a statistical nature, has been around as long as statistical science. However, unlike statistical science, the underlying theory and principles of statistical engineering have not been carefully documented, researched or published in literature. One purpose of this handbook is to address this oversight.

Hopefully, this chapter has made clear the fact that large, complex, unstructured problems cannot be fully addressed with individual tools, no matter how powerful. Rather, an overall strategy, or plan of attack is needed, one which will typically link and integrate multiple tools, and perhaps multiple disciplines. While each such problem is unique, and will require a unique strategy, there are common principles, common phases and common core processes (methodologies) that can be drawn upon. That is, just as chemical engineers can look to the unit operations of chemical engineering for the building blocks upon which to design a chemical process, so too statisticians and other analysts can look to the core processes and principles of statistical engineering to find their building blocks.

Further, the typical phases of statistical engineering, while not providing "7 easy steps to solving complex problems," do provide an overall framework, or way to think about attacking problems that at first glance may seem insurmountable, i.e., that are a "mess". Utilization of this framework will, of course, require substantial use of statistical and other tools. The remaining chapters in this handbook focus on some of the most commonly applied statistical tools, organized by the core processes. The final chapter, Chapter 7, discusses some of the overarching competencies required for statistical engineering, such as leadership, teamwork and project management. The statistical tools are of little benefit in attacking complex problems unless integrated with these competencies.

Chapter 2 - Enabling Technologies

Table of Contents

Preface

As we saw in Chapter 1, Statistical Engineering is a holistic approach, which means that applications typically involve a diverse set of tools and disciplines. These are integrated based on the context of the specific problem being addressed and the overall strategy developed by the team. While many of these technologies are quantitative by nature, the holistic approach also requires integration of "soft skills" that are needed to solve complex problems sustainably. These methodologies, which we refer to as enabling technologies, typically cut across all the phases of Statistical Engineering, hence we cover them now.

It is said that the best leaders have a sense of vision and an ability to communicate it. We recognize great leadership; we know it when we see it, but that statement alone does not suffice to convey leadership characteristics, environment, or technologies for initial and sustained success, especially in Statistical Engineering. Great leadership is often born out of necessity and often out of foreseen opportunity.

While our authors cannot promise to make readers into great leaders, they are able to identify and summarize traits, characteristics and key drivers. First among them is vision.

Think, "I have a dream." It is not, "Hey folks, I have this all mapped out." The vision is inspirational, or it is nothing. It is a statement of belief. It entices followers to adopt it as their own. And it is always present.

This chapter's opening section elaborates on leadership by detailing vision and the necessary and supporting organizational alignment. It includes steps for completing successful Statistical Engineering projects, even while facing resistance to change as detailed by Kotter (1996). Further, support is provided by descriptions of the need for social change and some means of bringing them about. This is followed by an annotated listing of guidance for leaders, compactly written.

The key to leadership and to subsequent organizational success is excellent communication. Many whose task is to glean meaning from large, unstructured data sets are aware of the need to work in teams to attain their goals. They can be likened to large colonies of ants. One might imagine that the ants could get their work done faster if they did not stop to rub antennae so often. Of course, their behavior represents time out for communication, and that communication facilitates, rather than delays accomplishment.

The second section emphasizes the need for clarity of communication and shows steps for its assurance, so that all team elements have uniform understanding. It goes on to discuss the advantages of story-telling and proper ways to carry it out. People learn more from stories than they do from explanations of facts. That is indeed why "I have a dream" works. It relates the vision in story form.

This section also provides advice about presentations to both large and small groups, written communications and their associated graphics, and quite useful information about the presentation of statistical findings.

Teams cannot thrive without meetings of some kind, face-to-face or virtual. By the same token, meetings in great abundance can be a team's kiss of death. So, when and when not, to have meetings is important. Meeting composition, team size and an understanding of human dynamics are keys to success. These topics are covered in some detail. So too is meeting structure for maximum effectiveness. The proper structure helps to assure a level playing field for participation by all. The result is greater meeting effectiveness leading to improved decision making and productivity.

The section goes on to present powerful tools for teams including:

- Brainstorming
- Affinity mapping
- Interrelationships digraphs
- Multi-voting
- Cause and effect diagrams

and a few others.

Section 2.1 - Leadership

In this section we discuss what leadership is, the work of leaders, leading statistical engineering projects, and we provide general guidance for leaders. Leadership is essential to the successful completion of SE projects because of the nature of the projects. SE projects are about solving large, complex unstructured problems that result in organizations and people working in a new and more effective manner. Problem solving is never easy. Leadership is needed to help people and organizations make the needed changes.

2.1.1 What Is Leadership?

There are many definitions of leadership. One definition is, "leadership is the capacity to translate intention into reality and sustain it." Another definition is, "enabling people and organizations to move from one way of working to another way of working." Along the way behavior and mindsets change along with changes in what is valued and rewarded.
The Work of Leaders can simply be stated as:

1. Craft a Vision - Imagining a future state that the group will make real
2. Build Alignment - Unite people toward a common goal
3. Champion Execution – Help individuals and teams accomplish goals

While these responsibilities on the surface appear to be sequential, in practice it is an on-going process with false starts, recycles and hopefully success in the end.

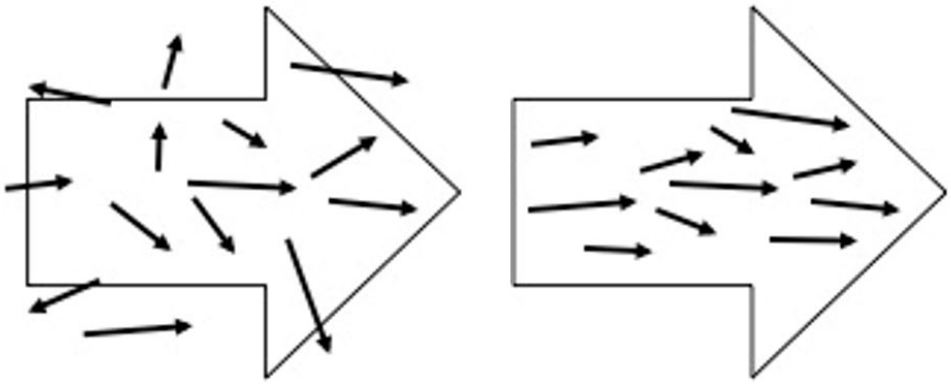
But how do leaders actually do this work? In the book, *The Work of Leaders* by Straw et al. (2013), the authors identify three drivers each for vision, alignment and execution that make it achievable. We summarize the drivers in this section (2.1.1). In the following section we dig deeper into the work of leaders in SE projects and discuss specific jobs, roles and tools involved in leadership of SE projects. In the last section we show how leaders can use the drivers of vision, alignment and execution to develop leadership initiatives and check on the progress of such initiatives.

2.1.1.1 A truly great vision elevates our work. It sparks our imagination. It touches our human need to do something bigger than ourselves. The need for vision has been recognized for a long time. The Bible tells us that, "Where there is no vision, the people perish" (Proverbs 29:18).

It is the job of the leader to facilitate the construction of the vision with input from the organization. Each person in the organization must see themselves in the vision and how their contributions can result in the realization of the vision. A vision must stretch the organization but still believed to be attainable. Over time as the organization progresses the vision can be revised to provide additional and new growth for the organization. A vision for statistical engineering in an organization might be "Statistical Engineering provides a competitive advantage for the organization."

2.1.1.2 Building Alignment is critical in moving from an imagined future state (vision) to reality. It is a dynamic, ongoing process that requires constant realigning as conditions and needs change. True alignment will meet *both* the rational and emotional needs of employees, customers and partners. Building alignment can be seen as "Aligning the Arrows" as seen in Figure 2.1.

Figure 2.1 Aligning the Arrows – Before and After



On the left we see the parts of the organization going in different directions. On the right the parts of the organization are aligned going in the same direction toward the vision. Fred Smith, Founder and CEO of Federal Express, tells us that: "Alignment is the essence of management." We believe that this admonishment holds for leadership as well.

2.1.1.3 Execution is making the vision a reality. Execution is how organizations take good ideas and turn them into results. Larry Bossidy, retired CEO of Honeywell, tells us that, "When I see companies that don't execute, the chances are that they don't measure, they don't reward, and don't promote people who know how to get things done" (Bossidy and Charan 2002). Harvard Business School research has identified components necessary for people to do good work. Two components deal with a sense of achievement: passion for a task and a working environment that stimulates creativity. Leaders are responsible for making sure people have what they need to do their work effectively, including creating the work environment. Leaders that champion execution, *defend* the time needed by the team to accomplish the work, *advocate* for the team, praising and providing feedback, *lobby* for resources and support from other work areas, and *cheer* on the team to maintain momentum.

2.1.2 Leadership for Statistical Engineering Projects

In the previous section we learned that the work of leaders involves three things: Creating a vision, building alignment and championing execution. In this section we dig deeper into the work of leaders in SE projects and discuss specific jobs, roles and tools involved in leadership of SE projects. Additional discussions of leadership of SE projects and improvement initiatives can be found in Snee and Hoerl (2012, 2018).
The leadership needs in conducting SE projects include:

- Identifying the right problem to be addressed and getting it framed properly
- Creating a strategy for the project
- Developing a sense of urgency regarding the problem and developing a guiding collation to communicate the importance of the project to the organization
- Obtaining resources to complete the project including people, funds, equipment, etc.
- Providing the needed education and training
- Communicating the project importance and progress toward solution
- Driving towards successful completion, and ensuring sustainability of the solution

Leadership is about change. So first one needs to understand the stages of change that are very nicely defined by Kotter's Eight Stages of Change summarized in Table 2.1.1 (Kotter 1996, 2008).

Table 2.1 Kotter's Eight Stages of Successful Change

Stage	Purpose	Primary objective
1	Establish a Sense of Urgency	Examine competitive realities and "Mission Critical" opportunities
2	Create a Guiding Coalition	Form a group with the power and influence to lead the change
3	Develop a Vision and Strategy	Create a vision to help direct the effort and the strategies to achieve the vision
4	Communicate the Change Vision	Use multimedia to help direct the effort including role modelling to demonstrate expectations
5	Empower the Organization for Broad-Based Action	Remove obstacles, change systems and structures, encourage risk taking
6	Generate Short Wins	Plan for short-term wins Recognize and reward involved persons
7	Consolidate Gains and Produce More Change	Initiate new projects and actions to change ineffective and inefficient processes and activities
8	Anchor the new Approaches in the Culture	Recognize and reward new ways of working Develop means to sustain the gains

We see in Table 2.1 that the work of leaders and the leadership for SE projects described are covered by the stages of Kotter's model.

2.1.2.1 Shortening the Path to Acceptance - Respond to Legitimate Concerns.

There is always resistance to any new direction including new ideas, initiatives, improvement projects, ways of working, etc. People will have legitimate concerns regarding SE projects that have to be addressed. Ignoring legitimate concerns may enable short advances but the concerns will continue to rise until they are addressed. A typical concern is: why is this project needed? Why is it needed now? This concern has to be addressed clearly and concisely. The answer has to be repeated throughout the life of the project. The value has to be clear to the organization.

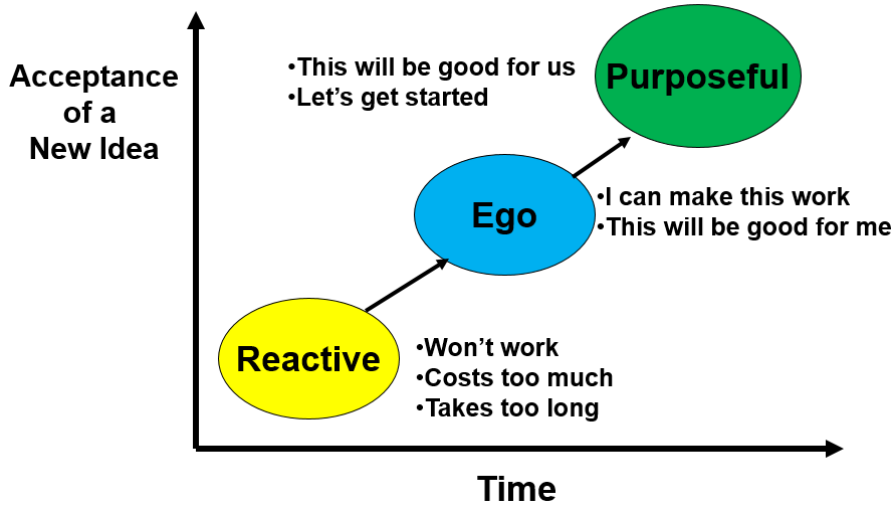
People also need to understand how this initiative is responding to competitive trends. Generally speaking, these include the facts that competition is tougher, products are more complex; we are in a global market, etc. Failing to respond to competitive trends will greatly affect the health of the company as well as its employees, stockholders and suppliers.

Other barriers to adoption include: technical – "It won't work in this case because…," financial – "We can't afford it," psychological – "It's too painful to change," and general resistance – "We tried this before and it didn't work," or "There is no need, we are already doing this." Again if these types of barriers are not addressed, progress will be slowed.

The first big step in dealing with resistance is recognizing that it will be present to some degree in everyone when a new idea is presented. Figure 2.2 shows the three modes of behavior people go through when presented with a new idea: "Reactive," "Ego" and "Purposeful." This model has been used for several years in the DuPont Company for dealing with culture change initiatives. Typically, everyone starts in the "Reactive" mode, thinking this won't work here. We can't afford it. It will take too long; we need a quick answer.

After thinking about the new idea for a while some people move to the "Ego" state. They begin thinking that *I can make this work for me.* It will make me more effective. Some people will then move on to the "Purposeful" state concluding that *this will be great for the organization. We will all benefit. Let's get after it.*

Figure 2.2 Modes of Behavior when Presented with a New Idea



People move through these modes of behavior at different speeds. Some move quickly. Some never get out of the reactive state. Some get stuck in the ego state and never get to the purposeful state. The job of the leader is to recognize the three states exist; we all generally start off in the reactive state and progress from one state to another at different rates.

Many believe that the most important thing one can do initially is to focus on Stage 6 of Kotter's model and generate short term wins. As they say, "Nothing succeeds like success." This piece of strategy is implemented using a project plan that includes a few projects that are important,

doable and can be completed quickly. This demonstrates to the organization that there are important improvements that need to be made and that we can successfully accomplish these improvements in a timely fashion.

2.1.2.2 Some Tools for Leaders of SE Projects.

This chapter sub-section discusses a variety of tools that are useful for leaders of SE projects. Specifically, they are the non-technical skills and tools discussed in Section 2.1., Leadership essentials (Section 7.2), Teaming tools (Section 7.3), Communication (Section 7.4 and Change Management (Section 7.5). These tools and approaches are very effective in helping an organization through the process of change.

It is also important to recognize that a critical role of the leader is to ensure continued communication with the organization and all involved in the project. These communications continue through the life of the project, including the development and implementation of strategy and plans, as well as progress made and important results. This can be effectively accomplished by recognizing that there are two important but separate activities at work: "rationalization" and "socialization" of new ideas and initiatives (Table 2.2, Scholtes 1998). Rationalization includes recognizing the need, developing the plan for the response, and implementing and monitoring the change. Rationalization is primarily a cognitive or mental activity. Socialization involves interactions with people to help them understand the need, getting appropriate personnel involved in planning the changes needed, and in the communication and feedback on the effectiveness of the change. Socialization, as the term implies, is related more to human interactions and direct involvement of people. Both rationalization and socialization are needed to truly internalize change. Typically, leaders are much better at rational activity than socialization. Attention to both aspects of change is needed to have a successful SE project.

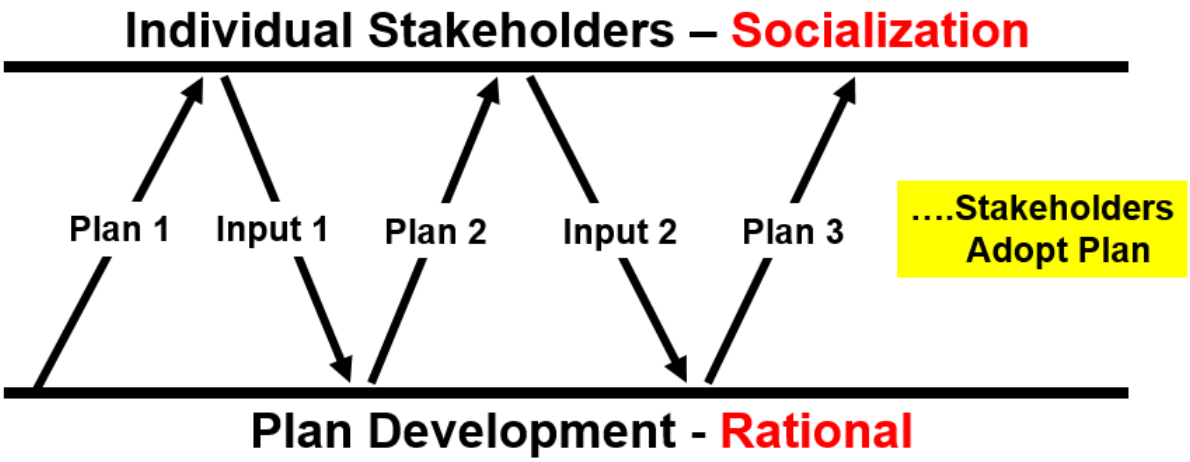
Table 2.2 Socialization of Change

Step in the Change Process	Approach
1. Understand the Need	Rational
Discuss the Need in Groups	Socialization
2. Plan the Response	Rational
Participation in Planning Response	Socialization
3. Implement and Monitor the Change	Rational
Communication and Feedback on the Effectiveness of the Change	Socialization

Figure 2.3 shows how the rationalization and socialization can interface and work together. A plan is created and shared with a group of one or more stakeholders. Input is received from the group, and the plan is revised as needed and appropriate. The revised plan (Plan 2) is shared with another group, input received and plan revised again. This process of "shopping the plan around" continues until sufficient stakeholders have been involved and integrated; that is, until we have

"critical mass" for the plan. The plan is now ready to be presented in a group meeting for discussion and adoption.

Figure 2.3 Building Stakeholder Value and Support – Shopping the Plan Around



This process best begins through input from the stakeholders who are most likely to be supportive. That is, we follow the path of least resistance. When reviewing the plan with the stakeholders you are under no obligation to include ALL the input received, only that which enhances the plan. However, it is best to be flexible rather than dogmatic. Eventually the plan, in an advanced state, is reviewed with those suspected to be less supportive. You will hear what their concerns are and should be ready to address these concerns when the stakeholders as a group review the plan. When the stakeholders see the nearly-final plan they will be able to see their concerns included in the revised plan. The stakeholders now "own" the plan.

2.1.3 Guidance for Leaders

In this section we provide guidance for leaders – tips and traps – by discussing the critical drivers of vision, alignment and execution. This information will be useful in deciding what leadership actions to take and what to check to assess the progress and performance of leaders.

2.1.3.1 Drivers of Vision

A truly great vision elevates our work. It sparks our imagination. It touches our human need to do something bigger than ourselves. The drivers of Vision are: Exploration, Boldness and Testing Assumptions.

"The human is the only animal that thinks about the future."

Exploration: Remaining Open. Many of us have a need for closure – to check things off the to-do list, to remove ambiguity, to create a clear path forward. Unfortunately, if that need for closure is high, you will tend to run with the first good idea you have and accept a vision that is not a good fit. Being open does not mean indecision. Rather, remaining open is about not

making a decision **too early**, resisting the temptation to run with the first idea, giving ourselves permission to take time to let the brain wander into unchartered territory.
Exploration: Prioritizing the Big Picture. It is always easier to put together a puzzle using the picture on the front of the box as a guide – literally the "big picture." The six questions of strategic clarity help leaders define the big picture:

1. Why do we exist?
2. How do we behave?
3. What do we do?
4. How will we succeed?
5. What is most important, right now?
6. Who must do what?

Boldness: Adventurous *Be bold, but not reckless*. Bold leaders stretch the boundaries beyond current practice and/or knowledge. Our nature too often is to "play it safe." Leaders who want to be more adventurous need to ask themselves: *What's the worst thing that could happen? What's the best?* Once you have identified your worst fear, you can confront it. Knowledge of the best thing can help you instill confidence in yourself and your team.

Boldness: Speaking Out. The desire to not look like a fool is a strong instinct. As leaders, we need to rein in our self-preservation instincts and go out on a limb, speaking out to voice ideas that seem unconventional and/or impractical. Most bold ideas are born into a fragile existence. If the idea is powerful, analysis and ingenuity will turn it into a practical, winning idea. This takes courage. Build up to it by pitching your ideas to others informally to get a sense of how others will react and to polish your delivery. Do not apologize or back down too quickly when you get negative feedback. Instead use the feedback to refine the idea and your pitch. Also, challenge others skepticism, do not let them off easy playing 'devil's advocate.' Some people will need time to reflect on the idea and come to see it from their own perspective. That is always part of change. Expect it.

Testing Assumptions: Seek Counsel. People are predictably prone to overconfidence when it comes to checking their intuition. When we believe we have come up with the greatest idea ever, our instincts are to protect it from criticism and rejection. Due diligence is an opportunity to enhance, tweak and understand the vision at a deeper level. Seek counsel by inviting people whose skills, knowledge and experience you respect into your vision – test it out with them and let it unfold. This is *not* asking for approval, but input. This is best done individually to prevent 'groupthink." Do not limit your counsel to folks you work with either – you may get fresh perspectives from a supplier, a customer, a neighbor, or your spouse.

Testing Assumptions: Explore Implications. Sometimes the vision seems so clear, so compelling we are more likely to believe arguments that support it, even when those arguments are unsound. One way to avoid this is to conduct a "pre-mortem" on your vision. We are all familiar with post-mortems or "after project reviews." Don't wait until *after* the vision is achieved to review it. Ask your team to imagine the vision has failed and to identify all the reasons for the failure before starting to work the plan. This will not only give you confidence in the vision and insights to improve the vision, but will sensitize your team to early warning signs of failure and give them time to respond in a way that will enhance the probability of success.

2.1.3.2 Drivers of Alignment.

Gaining alignment is critical in moving from an imagined future state (vision) to reality. It is a dynamic, ongoing process that requires constant realigning as condition and needs change. True alignment will meet *both* the rational and emotional needs of employees, customers and partners. This means that you must reach *both* the head *and* the heart through Clarity, Dialogue and Inspiration.

Clarity: Explain the Rationale. Clear communication is crisp. It is communication that provides enough information, but not too much; it is well-structured and efficient. It is simple but addresses real-world complexities. But crisp is hard work. Leaders often overlook communicating what is obvious to them but a mystery to everyone else. This means leaders need to share enough specifics to anticipate questions without overwhelming the hearer in details. A simple reason for a change should help people follow your logic and reach the same conclusion. Providing rationale is particularly important in times of uncertainty or large change. Speculation and gossip will occur if leaders do not step forward to offer clarity on the situation; people will fill in the gaps in communication, often with information that is far from the truth. A way to address this is through transparency – people at all levels have access to essential information when they want/need it. When crafting communications, leaders should look at the situation from the listener's point of view and then monitor people's reaction for comprehension.

Clarity: Structure Messages. Being crisp and clear takes time to structure the message. Meandering, unfocused communication leaves people confused and questioning leadership. Start by identifying the "headline." This should be no more than 8 words. Next, nail down the talking points. Ask: *"If people walk away with nothing else, what two or three points do I want them to remember?"* Finally, once you have structured your message, refer back to it often and consistently. Repetition and familiarity will shape people's attitudes and feelings.

Dialogue: Exchanging Perspectives. The factor with the highest correlation to job satisfaction is "a chance to have my opinion heard and considered." The word dialog means "through meaning," suggesting, "a free-flow of meaning through a group, allowing the group to discover insights not attainable individually." Dialog is an opportunity to give people voice which opens the door to shared ownership and accountability. To exchange perspectives, leaders need first to give people a safe place to open up, a place where they do not feel rushed or threatened. Second, leaders need to practice "reflective listening." Reflective listening is the leader summarizing what someone said using own words and checking for understanding.

Dialog: Being Receptive. Being receptive is not about the message or process of dialog (crisp and reflective listening); it is about the vibe the leader is sending out during dialog. People sense, both consciously and unconsciously, whether you are receptive and approachable. Your tone of voice and body language verify your receptiveness. People can quickly sense skepticism or disapproval, so leaders should try to suppress these emotions and to hold back on challenging the response from others. Remember dialog is a time for openness, not debate.

Inspiration: Be Expressive. Inspiration helps leaders obtain buy-in. It breathes life into the vision, galvanizing people. It is about bringing positive energy to the group and its goals. Being expressive is connecting to the audience. To do this, a leader must first be clear in their own mind why they are passionate about the vision. Next, leaders need to be specific. Like structuring the message for alignment, the leader should choose three key points. These points should speak to people's hearts, not just their heads. Employees perceive the environment through the eyes of their leaders. The moods, opinions, and actions of leaders rub off on their employees. If the leader is cynical or pessimistic about the vision, it will be reflected by the group. Likewise, if the leader appears excited and committed to the vision, the group will be too.

Inspiration: Be Encouraging. Leadership is about relationship. If the relationship goes sour between leader and follower, followers/employees will gradually lose their commitment to work. Being encouraging means giving support, confidence and hope to someone. It makes people feel good about themselves, their team, and their work. To encourage, give people a common aspiration, something the whole group can latch onto and be inspired by. Traditional strategies are to identify a "common enemy," a "noble cause," or a 'rallying cry." This requires knowing your audience – what motivates you may not motivate others.

2.1.3.3 Drivers of Execution.

Execution is making the vision a reality. Execution is how organizations take good ideas and turn them into results. Harvard Business School research has identified components necessary for people to do good work. Two components deal with a sense of achievement: passion for a task, and a working environment that stimulates creativity. Leaders are responsible for making sure people have what they need to do their work effectively, including creating the work environment. Leaders that champion execution, *defend* the time needed by the team to accomplish the work, *advocate* for the team, praising and providing feedback, *lobby* for resources and support from other work areas, and *cheer* on the team to maintain momentum. The drivers of execution are: Momentum, Structure, and Feedback.

Execution: Momentum. Momentum is the ability to accentuate the positive, making success more certain and challenges few and manageable. Consistent with Newton's First Law of Inertia, *a body at rest tends to stay at rest, and a body in motion tends to stay in motion*, it is harder to create momentum than to sustain it. Fifty percent of change efforts fail at the first stage of "Create a Sense of Urgency" (Kotter 1996, 2008). Complacency is hard to overcome. Momentum starts with the mindset of the leader. It is the mentality that the work we do contributes to the success of the team. But leaders cannot do this alone; they need to create a culture of momentum.

"The speed of the leader determines the pace of the pack."
- *Ralph Waldo Emerson*

Execution: Be Driven. Being driven is pushing yourself and others forward, believing things could always be better, never wasting an opportunity. There is an unspoken assumption that we do not wait around. A fast-paced organization does not have to be stressful. Leaders should "model the way," never asking team members to work harder than they are working or to

maintain a level of sustained activity that they are not committed to do. Leaders set high standards and commit their team to deadlines associated with external events. Why an external event? External events are harder to break or change as someone outside the organization is depending on the team.

Execution: Initiate Action. Leaders take responsibility for change when they see it rather than look the other way; leaders have initiative. Initiating takes energy, especially when it is about initiating around difficulties that arise during execution. Time is the biggest obstacle; leaders are already busy. To initiate action leaders must continually challenge priorities. What is the most important thing we ought to be doing to make a difference *right now*? Leaders must help team members take on new initiatives in their daily work, perhaps by writing the initiatives into annual objectives.

Structure: Providing a Plan. One cannot be a good leader without being a good manager, where being a good manager means you: plan, organize, direct, coordinate, and control work. A plan is a framework to bring together people, strategy, and operations. A plan ensures everyone is on the same page and provides a common foundation for the team to rely on. The leader's role in creating the plan will depend on the leader's role, experience of the team, and type of work. A front-line leader in a small organization may be very immersed in the details while the CEO of a large organization may only champion the leaders that report to him to invest in planning. The best way to obtain a team's buy-in to a plan is to engage them in the planning process. Planning is inherently an iterative process that takes time. Leaders need to provide the time to create a realistic plan.

"By failing to plan, you are preparing to fail"
- *Benjamin Franklin*

Structure: Analyze in Depth. A good plan that can be successfully executed needs to have rigor, depth of planning. Analyzing in-depth is about appreciating the true purpose of execution and understanding all the moving parts. This requires critical thinking to anticipate the cause-and-effect relationships that play out in execution of the plan. Leaders involve the team in this analysis and create an environment in which there is consistent and timely communication across functions and shared understanding of how the pieces are connected (process thinking). Leaders also challenge the team to think critically about what may happen so undesired variation (statistical thinking) can be eliminated or minimized. Leaders must be deliberate about providing plenty of time for such analysis.

Feedback: Addressing Problems and Offering Praise. Perfect feedback requires complete transparency – all the cards face up on the table. This is often not possible in the real world. To provide feedback, a leader must be involved, getting hands dirty to understand what is really going on in the trenches. People do not always speak up about problems due to organizational politics. A leader must speak up (be bold). Addressing problems can be tough as it means disrupting harmony. No one likes confrontation or risking hurting someone's feelings. If candor is done recklessly, it can kill transparency.

Leaders must make themselves vulnerable by acknowledging their own mistakes. Also, leaders should facilitate regular, semi-formal dialogues about what is not working. Remember to focus on the problem, not the people – the goal is to find a solution to the problem, not assign blame. People also need to know what is working right and to feel valued. Do not assume people know you appreciate the good work they do. Prioritize celebrations or milestones and build recognition into all your plans. Make sure the recognition is personal and fits the accomplishment.

Leading through Vision, Alignment and Execution is a simple model, but not simplistic. It is hard work that requires focus and intention. Start by understanding your tendencies and current performance using the drivers. Additional discussion and guidance on leadership can be found in Hunter (2008), Scholtes (1998) and Taylor (2014).

2.1.4 Putting It all Together

We conclude this section with the example of a plant manager making a major change in a manufacturing facility in which there were several candidates for change. This is certainly a large, complex unstructured problem. The plant manager started the discussion by talking about the competitive landscape and the need of the plant to respond. The planning process for the change began. Several small improvement teams were chartered to create some important building blocks for the needed changes. Training in the new manufacturing methodology was developed and completed. The plant manager held several "all hands" meetings to discuss direction and progress. As expected, some employees expressed support for the new direction, others were skeptical.

At the end of a year the plant manager and his staff decided it was time to put the new approach in place. He held another all-hands meeting and announced that, "It is time to get on the bus." In other words, the plant manager and his staff had taken the time to discuss and obtain input on the plan. Now, it was time to support it, even if some employees still had reservations. An important aspect of leadership is knowing when it is time to further discuss a potential decision, and when it is time to make the decision and move forward.

In this case, the decision had been made, and it was time to move forward to the new way of working. Much input was received, pro and con. One employee responded that, "He wanted to drive the bus." The major switch to the new process was completed and was successful. Production cycle times were reduced by 37% in the first month.

Senior executives were so impressed with the results that they asked the plant manager to discuss his experiences and results with the management staffs of other plants. He later received a major promotion to a more responsible leadership position.

Section 2.1 References

Bossidy, L. and R. Charan. *"Execution- The Discipline of Getting Things Done"*, Crown Business, Random House, Inc., New York, NY, (2002).

Hunter, James C. *The Servant: A Simple Story About the True Essence of Leadership*. Crown Business, Random House, Inc., New York, NY, (2008).

Kotter, J. P. *Leading Change,* Harvard Business School Press, Boston, MA, 1996.

Kotter, J. P. *A Sense of Urgency,* Harvard Business School Press, Boston, MA, (2008).

Scholtes, P. R. *The Leader's Handbook*, McGraw-Hill, New York, NY, 1998.

Snee, R. D. and R. W. Hoerl, "Leadership – Essential for Developing the Discipline of Statistical Engineering", *Quality Engineering*, Vol. 24, No. 2, April-June 2012, 162-170, (2012).

Snee, R. D. and R.W. Hoerl. *Leading Holistic Improvement with Lean Six Sigma 2.0,* FT Prentice Hall, New York, NY, 2018.

Straw, Julie, Scullard, Mark,, Kukkonen, Susie, and Barry Davis. *The Work of Leaders*. San Francisco, Ca: John Wiley & Sons, Hoboken, NJ, 2013.

Taylor, Kris. *The Leader's Guide to Turbulent Times: A Practical, Easy-to-Use Guide to Leading in Today's Times*. Evergreen Leadership, 2014.

Section 2.2 - Communication

2.2.1 Objectives

The purpose of this section is to discuss the communication skills required to successfully conduct statistical engineering projects. Conducting such projects is challenging, due to the nature of large, complex, unstructured problems. Therefore, considerable cross-functional and cross-disciplinary teamwork is needed, especially when it comes to actually deploying the solution, and ensuring sustainability.

2.2.2 Outline

We begin with a brief discussion of communication at the strategic, or leadership level. After setting this broader context, we discuss specific communication skills needed at the operational and tactical levels, typically associated with statistical engineering projects.

2.2.3 Strategic Communication

Communication is not so much an intellectual process as an emotional one. The whole point of leadership is to mobilize the workforce around what is most important. Therefore, leaders must appeal to the head and the heart when communicating. It is important that key messages come from different sources and through various channels using a variety of tools. Today's technologies (email, teleconferencing and social media) can be useful. But effective communication has little to do with technology. The world is full of organizations in which employees feel uninformed despite access to newsletters, intranet sites, Facebook groups, and town halls; these methods often lack interaction with the leaders, each other, and the message.

The most effective way to communicate is for members of the leadership team to come out of meetings with a clear message and promptly share with their direct reports, and then have those direct reports do the same for their direct reports. This is called "cascading communication." When employees hear all leaders saying the same thing after a major meeting or decision, they start to believe that alignment and clarity exist. This will create momentum for action.

The process for accomplishing cascading communication starts at the end of a leadership meeting or decision when the leaders agree on what they are going to bring back to their organizations. This requires the leaders to review their decisions, decide which are ready to share, which are not, and commit to the message and timing (within 24 hours is a good standard).

Face-to-face communication is best as it gives employees a chance to ask questions for clarification, and to hear the tone and see the body language in which the message is delivered. It is also best to communicate with the entire group so everyone hears the same message at the same time and can benefit from each other's questions.

Clear communication is crisp and structured. It is communication that provides enough information, but not too much; it is efficient. It is simple but addresses real-world complexities. Meandering, unfocused communication leaves people confused and questioning leadership. This means leaders need to share enough specifics to anticipate questions without overwhelming the hearer with details. A simple reason for a change should help people follow the logic and reach the same conclusion. Providing rationale is particularly important in times of uncertainty or large change. Speculation and gossip will occur if leaders do not step forward to offer clarity on the situation; people will fill in the gaps in communication, often with information that is far from the truth. When crafting communications, leaders should look at the situation from the listener's point of view and then monitor people's reaction for comprehension.

Crafting a crisp structured message starts by identifying the "headline." This should be no more than eight words. Next, nail down the talking points. Ask *"If people walk away with nothing else, what two or three points do I want them to remember?"* Finally, once you have structured your message, refer back to it often and consistently. Repetition and familiarity will shape people's attitudes and feelings.

2.2.3.1 Everyone Needs to Know the Score

Leaders need to communicate in ways that stick. The old adage, "A picture is worth a thousand words," is really true. Simple tables, graphics and drawings can be effective ways of painting a picture of the current situation or possible future. They can also help to weave information (facts/data) into a story.

Scorecards and other visual management techniques help leaders and help team members manage and achieve performance results. They are timely, easy to understand, and often provide a graphical depiction of the performance of key performance indicators (KPIs) – like in sports, they let members of the team know if the organization is "winning" (achieving its targets for success). They help employees think and act like owners.

In its simplest terms, a balanced scorecard is a set of measures (scores) that translate the organization's strategies and goals into a comprehensive set of measures that provides a strategic framework for communicating clarity and driving alignment across the organization. There is usually a hierarchy of scorecards. The scorecards are linked vertically and horizontally to each other. Vertical linkages connect the individual work team scorecard to organizational strategy and top-level goals; it helps work teams focus on strategic priorities and the organization's vision. Horizontal linkages connect customer's needs to process measures across work teams.

Historically, organizations have measured and communicated performance financially; this approach focuses on improving cost, quality and time of existing processes. The financial reporting process, however, is anchored in an accounting model that does not include the intangible and intellectual assets of an organization. These assets and capabilities are critical for success in today's competitive environment.

The objectives and measures of a balanced scorecard are more than a collection of financial and non-financial performance measures. The measures represent a balance between external

measures related to shareholders and customers and internal measures of critical processes, innovation, and learning and growth; they are balanced between outcome measures, or results of past efforts, and the process and infrastructure measures that enable future performance.

World-class organizations use the balanced scorecard as a strategic management system consisting of four steps:

1) Clarify and translate vision and strategy
2) Link strategic objectives and measures
3) Plan, set targets and align strategic initiatives
4) Feedback and learning

2.2.3.2 Using Stories for Strategic Communication

In most organizations, few employees have the analytical skills to critically look at the data they have at their fingertips, let alone use it for good decision-making. Few organizations have addressed how to effectively share knowledge/data/information among employees; workers often have insufficient knowledge to make key decisions and take effective action for improved productivity. Enter stories. Facts inform but stories resonate. Strategy, culture and systems do not change behavior in the same way stories do.

Stories are the most ancient forms of communication. Prehistoric people conveyed stories with drawings on stone walls; Egyptians told stories with hieroglyphics. Jesus used parables to change people's thinking and beliefs, redirecting lives. The use of stories in today's businesses, non-profit organizations and government is only just being recognized as a way to engage people in the organization's mission.

Stores of our lives form the basis of all we are and do. Lest leaders think this is all fluff, stories and storytelling have bottom-line impact. Stories can convey tribal knowledge, demonstrating the value of specific initiatives and of the organization to its customers and community. Stories which envision the future build trust, enable mutual respect and have the power to reframe perspectives and create alignment. All of these benefits positively engage workers which in turn increases productivity and profitability. When organizations provide customers with something to talk about, they will talk about it, positive or negative. Positive stories will attract more business; this is the ripple effect stories have on organization's profitability. Organizations that use stories as part of their sales process create two powerful advantages: they better understand their customer needs by listening to their stories, and they build trust, an essential ingredient to a long-term customer relationship. One hospital that captures and shares patient stories sends the message that patients are valued. Sharing patient stories also provides ordinary people a way to give back to the hospital staff that helped them heal while providing hope to others.

Key to acting out any story is to be authentic – remain true to yourself; bring yourself to the platform. You are the vehicle through which information is being transmitted. If you are not comfortable with yourself, transmission of the message will not be strong. Inexperienced storytellers spend a disproportionate time on content, not enough on delivery.

You do not have to be a good actor to tell a good story. Rather, to tell a good story, first choose the right story. This will typically be one that:

1) includes vivid details
2) includes a lesson learned
3) can be used in a business context
4) will call people to a higher standard
5) you enjoy telling.

The story does not have to be personal. You can use current events, inspirational historical individuals, TV shows, movies. Good stories often involve a 'turning point' – a time when someone made a change in their life – geography, relationship, job. career, responsibility, perspective, accomplishment, or tragedy/injury.

Once you have chosen the right story, craft the story. Crafting a good story takes time and multiple drafts. Stevenson (2008) suggests nine steps to structuring a good story:

1) Set the stage/scene – create context; frame the story to help audience know where you are starting
2) Introduce the characters - use physical and emotional descriptions
3) Begin the journey – leave the safety and comfort of the initial scene
4) Encounter obstacle – a person, decision, physical or emotional problem; this is the most dramatic part of the story. Help your audience experience it.
5) Overcome the obstacle- plant the seed for the lesson to be learned
6) Resolve the story –; the let audience know how everything turned out; tie up loose ends, leaving no unanswered questions
7) Make the point - share the lesson learned
8) Ask the question – engage the audience in their experience with something similar
9) Restate the point – summarize the story and call your audience to action

The following story was used to explain why a statistical mindset is so critical. In this case, it was to see and understand the "story" of variation that was creating waste and an environmental issue. Further, it shows how engaging the workforce to improve and sustain improvements over time can produce a cycle of positive reinforcement. Extending this mindset into all decisions of the organization, including new product design, can have further benefits that enable an organization to compete today and tomorrow.

The lead operator at a large food manufacturer was assigned a project to reduce disposal costs of a packaging line that was experiencing excessive waste. She assembled a cross-functional team to explore the situation. From the beginning, the team met with resistance from plant engineering and leadership that wanted to use waste disposal as a test for implementing new robotic handling technology. Through simple observation, the team learned that product overflowed jars or bounced off jars onto the floor during filling was collected in drums and sent to a local landfill daily. Therefore, there was an additional environmental issue, above and beyond the cost issue. Landfill costs were based on weight. Since collecting the waste was done at the end of the shift and operators were eager to just go home, they did not take the time to weigh each drum. Instead, they estimated the weight based on fill height, ignoring variation in products, drums, and fill geometry.

The landfill did not weigh individual drums either and relied on the weight stated on the paperwork sent with each delivery to bill the manufacturer. Landfill costs were the highest costs at the plant outside of materials and labor. As a result, plant management was eager to reduce these costs. The team mapped the process flow and conducted a measurement system analysis of weight at the loading dock scale. This led to the discovery that most drums were not being weighed and that the estimated weight was significantly higher than actual weight; the plant was being billed too much by the landfill. Once this was shared and rectified with the landfill, automation of waste handling was no longer financially viable and engineering staff abandoned their design plans for robotic handling of waste.

This alone could have been the end of the project as it met the original improvement target, but seeing the impact of variation on plant costs, the project team passionately asked to continue the project to identify root causes of the product on the floor. Reductions in waste would not only improve the bottom line, but also reduce the environmental impact of the plant. Through root cause analysis, followed up with simple statistical and graphical analysis to quantify and understand the variation, the team was able to identify and remove several root causes, reducing waste by 50%, resulting not only in a dramatic financial savings, but also a huge reduction in material going to landfill. Statistical thinking enabled the team to understand the story behind the waste and make lasting improvements and encouraged them to keep looking for other opportunities.

2.2.4 Presentations

Statistical engineering projects are of no value unless the process changes identified by the project are implemented. This typically requires that the study results be presented to management and others to build support for the proposed changes and obtain the needed resources. These presentations can take many forms, such as informal one-on-one discussions, formal presentations to various groups, and written reports. Key requirements for any presentation or report are that it be clear, concise, and accurate. The following information will help you prepare for such interactions.

2.2.4.1 Presentations to Individuals or Small Groups

The simplest and most frequent communication is a presentation to a single person or to a small group. One should not take such interactions lightly. As John Wooden, renowned UCLA basketball coach, pointed out, "Failing to plan is planning to fail." Careful preparation can make the difference between getting and losing the support you need. First, identify the purpose for the meeting and your expected outcomes (i.e., what you would like to happen as a result of the meeting). Next, construct an agenda for your meeting that will produce your desired outcomes.

A typical agenda might include:

- Introductions
- Meeting purpose and desired outcomes
- Project description
- Study design and data collection and analysis

- Results, interpretation, and conclusions
- Accomplishments to date, or since last review
- Progress toward goals as reflected in the key project metrics
- Recommendations, needed resources, and help
- Key learnings and issues
- Next steps and meeting conclusion

An agenda for a shorter presentation to management might include:

- Project description
- Key results
- Accomplishments to date, or since the last review
- Progress toward goals as reflected by the key project metrics
- Issues, needs, and next steps

The agenda tells your audience what problem you worked on, the work you did, and your recommendations based on your work. You will find that the positive outcome of the meeting will make the thorough preparation well worthwhile. Such a presentation may use a projector for computer-generated visuals and may involve a handout and use of a whiteboard to make your points. Handouts and whiteboards often work well for small groups.

2.2.4.2 Presentations to Large Groups

The preparation for presentations and project reviews for large groups is similar to that for small groups. The key difference is that you will be more dependent on projected visual aids to make your points. The discussion following your presentation may also be more formal because of the large audience size. The content of the presentation can be similar to that of the small group. It is particularly important that the visuals be easy to read and understand when the audience is large. Some guidelines for creation of visuals are:

- Use one slide for each 2 minutes of presentation.
- Use no more than 30 words per slide.
- Use no more than 8 lines per slide.
- Do not use acronyms or abbreviations.
- Use 20-point or larger font size.
- Use high contrast between lettering and background—dark lettering on a light background or vice versa.
- Include a "take-away box" at the bottom of the slide to reinforce key points.

These guidelines will help keep the information content per slide reasonable. However, in some instances you may choose to violate these guidelines. You can get away with more information per slide in smaller groups from a readability standpoint but understandability may still be an issue. Remember that when these guidelines are violated you may waste valuable meeting time explaining your slides instead of your project.

2.2.4.3 Presentation Pitfalls

Some common presentation mistakes to avoid include:

- Talking to the projection screen (always face the audience when speaking)
- Saying "uh" between sentences (a very common nervous habit)
- Speaking too fast or using slang when addressing an international audience
- Reading a speech (boring and insulting to the audience—speak in your own words, even if you must memorize what you want to say)
- Speaking in a monotone (try to vary the inflection of your voice)
- Unreadable visuals (discussed above).

2.2.5 Written Reports

You may often need to follow up the presentation with a written report. This often serves to document the work and conclusions, and contribute to a "corporate memory." However, reports take time to prepare. As a compromise in many instances, an electronic copy of the presentation slides, if they are complete, will be sufficient for this purpose. When a written report is needed, its contents should include the following items (in order):

- Cover page if appropriate
- Key conclusions in an executive summary
- Project background
- Study design
- Data collection and analysis
- Results and interpretation
- Discussion, conclusions, recommendations

The contents and style of the report should always match the needs and culture of the intended audience. The executive summary is a concise statement of the key conclusions, recommendations, and take-aways from your project. Keep it general and short—do not present details in this section.

2.2.5.1 Summary or Abstract

Sometimes, the only written report required is a one-paragraph summary or abstract of the project. This still provides important documentation. Such a paragraph typically contains three parts: problem/issue description, work done, and results/impact/implications. Of course, it should also document the individuals involved in the project, in case someone reading the abstract is interested in more details. Depending on the required length, it is usually appropriate to include a two- or three-sentence description for each of the three areas—that is, a total of six to nine sentences. It is particularly important in the business world to include the results, impact, and implications. A mere description of the work done is not sufficient and may irritate some of your audience.

2.2.5.2 Graphics

A graphic should be included to illustrate each of the key points of your report. Graphics should be clear and understandable. Graphics can contain too much information (i.e., a "busy" graph) in the same way that a slide can contain too many words. Graphics may have been used during your research for data exploration, analysis, and communication. In presentations and reports, graphics will help communicate your results. However, the graphics used in the exploration and analysis phases of the project are not always appropriate for communication of results. Plan to revise or replace charts that are too obscure or complicated. In addition to the display of summary results, graphics can be used to display process flows, graphs of models, and procedures.

2.2.5.3 Presenting Statistical Results

A graph is the best way to present statistical results, not only in presentations, but in written reports; as well. Include statistical results in the text of the report or in tables when you need to support decisions, conclusions, and recommendations. Readers of your report will want to know what data you used as a basis for your conclusions, and its pedigree. When possible, supporting statistics should be accompanied by some measure of uncertainty such as confidence limits.

Tables are another effective way of reporting data and statistical results. Tables should be clear, concise, and as simple as possible. Keep in mind that the objective is to help the reader understand what analyses you did and how you reached your conclusions. Clearly label the table title and the names of the rows and columns. Use table footnotes where necessary to help the reader understand and use the tables. As much as possible, each table should stand on its own and not require reference to the text to understand the table contents. The table should be constructed so that it is easy to make comparisons of interest.

2.2.5.4 Written Report Pitfalls

Some common mistakes to avoid in written reports include:

- Burying the key conclusions at the end (see advice above concerning the executive summary)
- Explaining how you did what you did before explaining why you did it and what you actually did (discussed above)
- Using technical language beyond the understanding of the intended audience. (KISS: Keep it simple, stupid! The object is to communicate, not to impress.)
- Getting bogged down in details, such as a complex financial analysis (state the conclusions in the body and include the details as an appendix)
- Making the report a one-sided "position paper" (objectively state the results and provide data to back up key recommendations)

2.2.5.5 Summary

Many technical projects that have had high potential and impressive results have languished and failed to produce tangible benefit. One reason for this is poor communication of the projects to others, especially decision makers. Only recommendations that others clearly understood are likely to be implemented. Therefore, communication of the work should never be considered "grunt work," necessary but not important. Communication of the project results, via both presentations and written reports, is a critical aspect of solving the problem, and should be treated as such.

Section 2.2 References

Chang, Richard and Mark W. Morgan. *Performance Scorecards: Measuring the Right Things in the Real World*. New York: Jossey-Bass, 2000.

Kaplan, Robert S. and David P. Norton. *The Balanced Scorecard*. Boston: Harvard Business School Press, 1996.

Kaplan, Robert S. and David P. Norton. *The Strategy Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment*. Boston: Harvard Business School Press, 2001.

Hayes, Robert H. and Steven C. Wheelwright, "Link Manufacturing Process and Product Life Cycles," Harvard Business Review, January-February, 1979.

Hayes, Robert H. and Gary Pisano, "Beyond World-Class: The New Manufacturing Strategy," *Harvard Business Review*, January-February, 1994.

Hayes, Robert, Gary Pisano, David Upton, and Steven Wheelwright. *Operations, Strategy, and Technology: Pursuing the Competitive Edge*. Hoboken, NJ: Wiley & Sons, Inc., 2005.

Hoerl, R.W., and Snee, R.D., *Statistical Thinking: Improving Business Performance*, 3rd ed., John Wiley and Sons, Hoboken, NJ, 2020.

Heskett, James L., Thomas O. Jones, Gary W. Loveman, W. Earl Sasser, Jr., and Leonard A. Schlesinger. "Putting the Service-Profit Chain to Work." *Harvard Business Review*, March-April 1994.

Lafley, A. G. and Roger Martin. Playing to Win: How Strategy Really Works. Boston, MA: *Harvard Business Review*, 2013

Labovitz, George and Victor Rosansky. *The Power of Alignment: How Great Companies Stay Centered and Accomplish Extraordinary Things*. New York: John Wiley & Sons, Inc., 1997.

Lencioni, Patrick. The Advantage: *Why Organizational Health Trumps Everything Else in Business.* San Francisco, CA: Jossey Bass, 2012.

Rummler, Geary A. and Alan P. Brache. *Improving Performance: How to Manage the White Space on the Organization Chart, 2nd Edition*. San Francisco: Jossey-Bass Publishers, 1995.

Shook, John. *Managing to Learn: Using A3 Management process to Solve Problems, Gain Agreement, Mentor and Lead*. Cambridge, MA: The Lean Enterprise institute, 2008.

Silverman, Lori. *Wake Me Up When the Data Is Over*. San Francisco, CA: Jossey-Bass, 2006.

Simons, Annette. *The Story Factor*. Cambridge, MA: Basic Books, 2001.

Stack, Jack. *The Great Game of Business*. New York: Currency Books, 1992.

Stevenson, Doug. *Story Theatre Method: Strategic Storytelling in Business*. Colorado Springs, CO: Cornelius Press, 2008.

Straw, Julie, Mark Scullard, Susie Kukkonen, and Barry Davis. *The Work of Leaders: How Vision, Alignment and Execution Will Change the Way You Lead*. San Francisco: John Wiley & Sons, 2013.

Taylor, Kris. *The Leader's Guide to Turbulent Times: A Practical, Easy-to-Use Guide to Leading in Today's Times*. Evergreen Leadership, 2014.

Treacy, Michael and Fred Wiersema, "Customer Intimacy and Other Value Disciplines." *Harvard Business Review*, January-February 1983.

Section 2.3 - Effective Teamwork

2.3.1 Objectives

This section provides basic guidelines and advice regarding teams, their formation, their organization, the leadership and other team roles required and criteria and tips for team success.

2.3.2 Outline

Beginning with advice concerning team composition, leadership and maintenance; this section moves on to cover critical elements of performance, size, key disciplines, structure, and roles of the various participants. From there, it continues to discuss decision making, conflict resolution, behavior and responsibilities.

Next, selected tools for idea generation, coordination and consolidation for priority setting are described and exemplified.

A list of references is provided so readers may pursue selected topics further.

2.3.3 Selecting, Leading and Maintaining Teams

Teams have existed since humans began living in social groupings. Most people assume they know how teams work – after all, they have had first-hand experiences all their life – family, baseball teams, scouts and project teams at work. Despite this and the growing recognition of what teams offer in the workplace, the collective impact of teams on the performance of an organization is woefully underexploited.

"It is the long history of humankind (and animal kind, too) that those who learned to collaborate and improvise most effectively have prevailed."

\- Charles Darwin

2.3.3.1 What is a team and when to use one rather than individuals

Katzenbach and Smith (1993) define a team as:

"a small number of people who are committed to a common purpose, performance goals and approach for which they hold themselves accountable."

Teams outperform individuals when*:*

- The task is complex and/or cross functional
- Creativity is needed
- The path forward is unclear
- Efficient use of resources is needed
- Fast learning is necessary
- High commitment is desirable for implementation and achievement of results.

2.3.3.2 Lessons from studies of team performance

Katzenbach and Smith studied teams to identify four key lessons for maximum performance:

1. No team arises without a challenge meaningful to those involved. Teamwork is not the same thing as a team. A common set of demanding performance goals considered important by the group will lead, most of the time, to both performance and a team. Performance is the primary objective; a team remains the means, not the end.

2. Leaders can foster performance best by building a strong performance ethic rather than by establishing a team environment alone. Simply organizing around teams and calling groups 'teams' will not generate the same results as a true team.

3. Biases toward individualism exist, but do not need to get in the way of team performance. Most of us grew up with a strong sense of individual responsibility. Parents, teachers, coaches and other leaders have shaped our views and focus on individual accomplishment; rugged individualism is highly valued in US society. Building shared value and commitment are key to ensuring individualism does not get in the way.

4. Discipline within the team and across the organization creates the conditions for team performance. Groups become teams through disciplined action. They shape a common purpose, agree on goals, defining a working approach and develop complementary skills, and hold themselves accountable for results.

2.3.3.3 Team size

So, what is the right size for a team to form? A general recommendation is that a team be composed of representatives of the areas impacted by the problem and potential solution. More than 10 people is unwieldy – cannot even agree on a time and place to meet; less than 4 and team may not possess the diversity of thought and experience needed to avoid 'groupthink.' Larger groups (25-50) can theoretically become a team, but they usually break into sub-teams.

2.3.3.4 Key team disciplines

Eight key disciplines have been found to improve team effectiveness:

1. **Shared purpose**. Purpose gives a team focus and direction. When the purpose is something clearly important to the individuals, they are more likely to feel their time is well spent. If the purpose is given to the team by leadership, the team must still spend time building common understanding, ownership and commitment to the purpose. If not specified, the best teams invest time up front exploring, shaping and agreeing on a purpose that belongs to them individually and collectively.

2. **Commitment to team**. Team members buy-in to the decisions and standards of the team, where buy-in is honest emotional support, not consensus. Waiting for everyone to

agree is a recipe for mediocrity, delay and frustration. This requires clarity. Clarity is the removal of assumptions and ambiguity from a situation. Real clarity can only be achieved when team members can freely share ideas, thoughts and concerns in an unfiltered debate – productive conflict. Most people do not need to 'get their own way' in order to buy in; they simply need to be heard. This type of commitment extends beyond team meetings and to communications with the rest of the organization. Committed teams take the time to clarify their agreements and action plans so their communication is consistent when they interact with others outside the team.

3. **Leverage capabilities**. Teams develop and leverage a mix of skills, including:
   - Technical/functional expertise
   - Problem-solving skills
   - Interpersonal skills

   No one team will have all the needed skills at the outset but will have within it the capability to develop or obtain the skills needed through personal learning and development and reaching out to others in the organization for support.

4. **Communication.** Communication is critical to execution, change management and organization culture. Communications within the team and outside the team are critical to team success. Team communication includes:
   - Team charter – a document that describes in clear, measurable terms the task the team is to accomplish, scope, timelines and membership. The team operates within the framework of the charter.

Table 2.3 Example of a Team Charter

Project Title:			
Project Description:			
Business Linkage:			
Expected Results:			
Metric:			
Expected Financial Impact:	EVA$	IRR	NPV
Project Start Date		Est. Completion Date:	

Team Members

Employee Name	Expertise	Employee Name	Expertise

Required Signatures

Project Champion:		Date:	
Team Leader:		Date:	

   - Team meeting notes – document team discussions and agreements during formal meetings
   - Team action plans – document team plans for execution – what, who, when, how

The meeting notes and action plans form the team 'memory.'

5. **Meeting Management**. To continue to feel that time spent with the team is worth the effort, that time must be well spent through good meeting management. Keys are meeting roles and responsibilities as summarized in the following table.

Table 2.4 Team Position Responsibilities

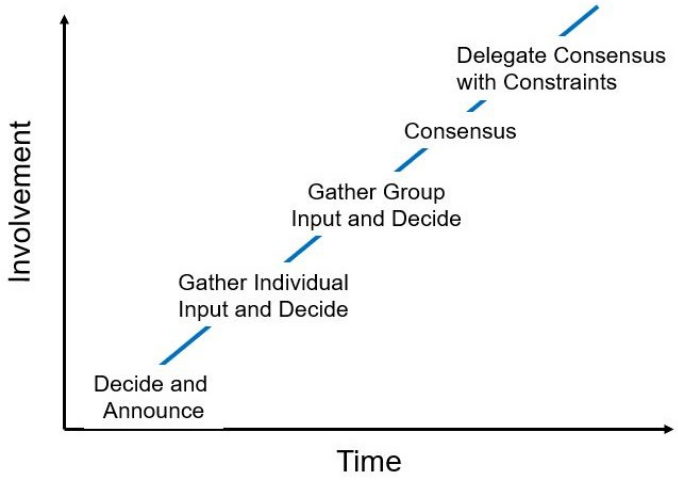
Position	Responsibilities
Leader	The leader determines if a meeting is really needed. If so, the leader plans and circulates a trial agenda prior to holding the meeting, decides upon the venue – best chosen as a safe, neutral location, allocates time and duration, assigns a facilitator, recorder and time keeper in advance, and takes full responsibility for the meeting outcome and evaluation. The leader must be careful of the cross disciplinary representativeness of participants and, for ease of communication, to keep meeting size as small as practicality will permit. At the meeting's close, the leader must assure that all involved buy into the outcome and are willing and able to support it.
Facilitator	The facilitator role, not as common in meetings as perhaps it should be, is critical to meeting success and is therefore highly recommended. A person taking on that responsibility should hold knowledge of the team's missions and objectives but should not necessarily be directly involved in carrying them out. Prior to the meeting, the facilitator coordinates with the leader; during it, the facilitator helps to provide laminar flow by recommending decision tools and methods, and helping the team maintain focus. To be effective, the facilitator must remain neutral, especially during moments of controversy. Tact and diplomacy are of the essence. This includes assuring the team members follow ground rules, that every voice gets heard and that discussions do not become heated.
Recorder	The recorder, a fully participating member, is the key to preserving team memory. Whatever the media, chalk board, flip chart or projected PowerPoint, notes are best taken in real time and displayed so everyone can see and offer immediate corrections and clarification where necessary. Every effort is made to preserve the team thinking accurately and in as few words as possible. The recorder aids in the distribution of minutes and notes following the meeting.
Time-Keeper	The timekeeper, also a fully participating member, keeps an eye on the clock and alerts the leader to approaching time limits. The timekeeper assists in assuring that meetings begin and end on time and that time allocated to each agenda item is kept.
Participant	The participant is a person with responsibility and capability for contributing. This might be someone with special expertise such as a hospital cardiologist or a microbiologist who supports R&D in a consumer goods company. It might also be someone who holds special authority such as a production line foreman. Effective teams will consist of a diversity of responsibilities and ranks within the organization.

It is usually helpful to initiate team efforts by spending time to establish meeting ground rules. The leader might begin by asking members for their ideas, and the facilitator may help coordinate. In addition, the recorder should summarize group thinking along these lines by posting suggested rules where all can see. Here is an example:

Guidance for brainstorming:
- No Yes-men, Yes-women allowed
- No grouches
- No judging
- Welcome "bad" ideas
- Aim for quantity (10 ideas each)
- Charge up! Caffeine, Sugar!
- No electronics
- Keep it short
- Revisit it tomorrow

6. **Decision-making.** Effective teams intentionally decide how they will make decisions and then consistently use that method, recognizing the more involvement required for the method, the longer it will take to make the decision:

Figure 2.4 Decision Making Continuum



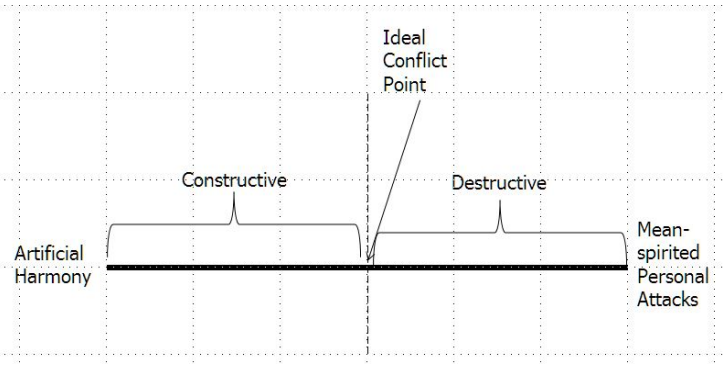
They also know when they are ready to make a decision, when they need to reach out to others for input, and to reconsider or to stress-test their ideas.

Five tips for better meetings:
1) Know and communicate the purpose of the meeting – is it tactical or strategic? Brainstorm, debate issues, explore alternatives or make recommendations?
2) Clarify the stake – why is the meeting important?
3) Add drama by putting the most controversial topics first.
4) Spend enough time to end with clarity and commitment; ending on time means little if the meeting ends without clarity and commitment.
5) Provoke conflict – hold productive debates to get to the bottom of issues.

7. **Conflict resolution**. Conflict is a fact of life in groups of people. It is simply a condition in which concerns of people appear incompatible; it is not good or bad in and of itself. Positive outcomes are possible when conflict creates deeper understanding of an issue; negative when it is not reconciled and results in poor decisions, deadlocks, wasted energy or apathy. It is therefore in the best interest of a team to learn to engage in productive debate – to find and hold the point between artificial harmony and mean-spirited attacks.

Conflict is a continuum:

Figure 2.5 The Conflict Continuum



See Lencioni (2005).

Theoretically, the best place on the continuum is close to the middle. This is a point where a group is having productive debate without slipping into destructive territory. Even the best teams will occasionally step over the line. This is a good thing as long as the team is committed to working through it.

"***Conflict cannot survive without your participation.***" - Wayne Dyer

In the heart of conflict automatic thoughts are put into our heads, no matter how irrational. These thoughts can lead to destructive responses such as arguing, gossiping/complaining about someone, belittling, being hypercritical, caving in, overpowering, defensiveness, passive aggression, dismissing others' opinions, revenge, being overly dramatic, sabotage, exaggeration, sarcasm, exclusion, stonewalling, finger-pointing, or withdrawal.

* from, *Everything DiSC Productive Conflict,* John Wiley & Sons, 2017.

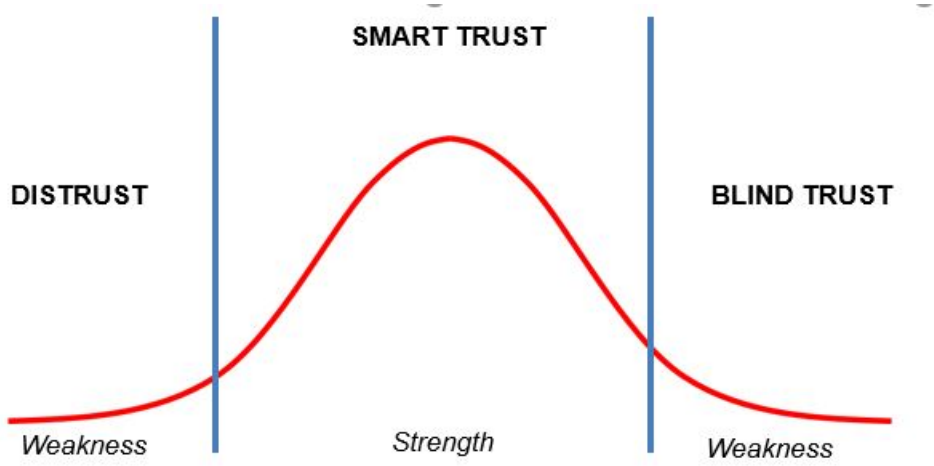
The key is to learn how to take a step back from the situation and make a productive response instead. Each time a team recovers from an incident of destructive conflict, it builds confidence that it can survive such an event, which builds trust. Trust is the foundation of creating a cohesive team that is able to have productive debate, make decisions and commitments to one another, and hold one another accountable to results.

Stephen M.R. Covey defines 13 trust behaviors for building relationship trust:

1. Talk Straight
2. Demonstrate Respect
3. Create Transparency
4. Right Wrongs
5. Show Loyalty
6. Deliver Results
7. Get Better
8. Confront Reality
9. Clarify Expectations
10. Accountability
11. Listen First
12. Keep Commitments
13. Extend Trust

All 13 require a combination of character and competence. The first five primarily flow from character, the second five from competence and the last three are an equal mix of character and competence. The 13 behaviors work together to create balance; these behaviors exist on a continuum. Too much results in blind trust; too little results in distrust. The "sweet spot" is **Smart trust as depicted in Figure 2.6.**

Figure 2.6 Smart Trust



*Continuum of trust behaviors from: Covey, Stephen M.R, The Speed of Trust: The One Thing That Changes Everything. Free Press, 2006

Creativity, trust, and higher performance are possible outcomes of productive conflict. In order to teach a team to engage in productive conflict, it is important to understand everyone's viewpoints on and comfort levels with conflict as they can be radically different. Some people are comfortable screaming and shouting while others shutdown.

A person's conflict style is determined by a number of factors – temperament, cultural background, and family norms.

One of the best ways to understand your and others' conflict profile is to use a profiling tool such as Myers-Biggs (MBTI) [need reference here and in reference list] or Everything DiSC (2017), both of which address how an individual's style reacts under stress. In addition, there is an instrument focused solely on identifying your conflict mode, the Thomas-Kilmann Instrument (TKI). [need reference here and in reference list]

8. **Planning for Results.** With smart trust, productive debate/conflict, commitment and accountability systems a team can identify actions required to accomplish its goals. This typically involves creating an action plan which identifies who does what, when and how in a way that everyone understands their role, the interdependence between roles and tasks and allows team members to put aside their own ego and focus on team success. The discipline is documenting the plan, making it visible, and using it to track progress and hold one another accountable.

2.3.3.5 Accountability

"Accountability is the glue that bonds commitment to results"
- Will Craig

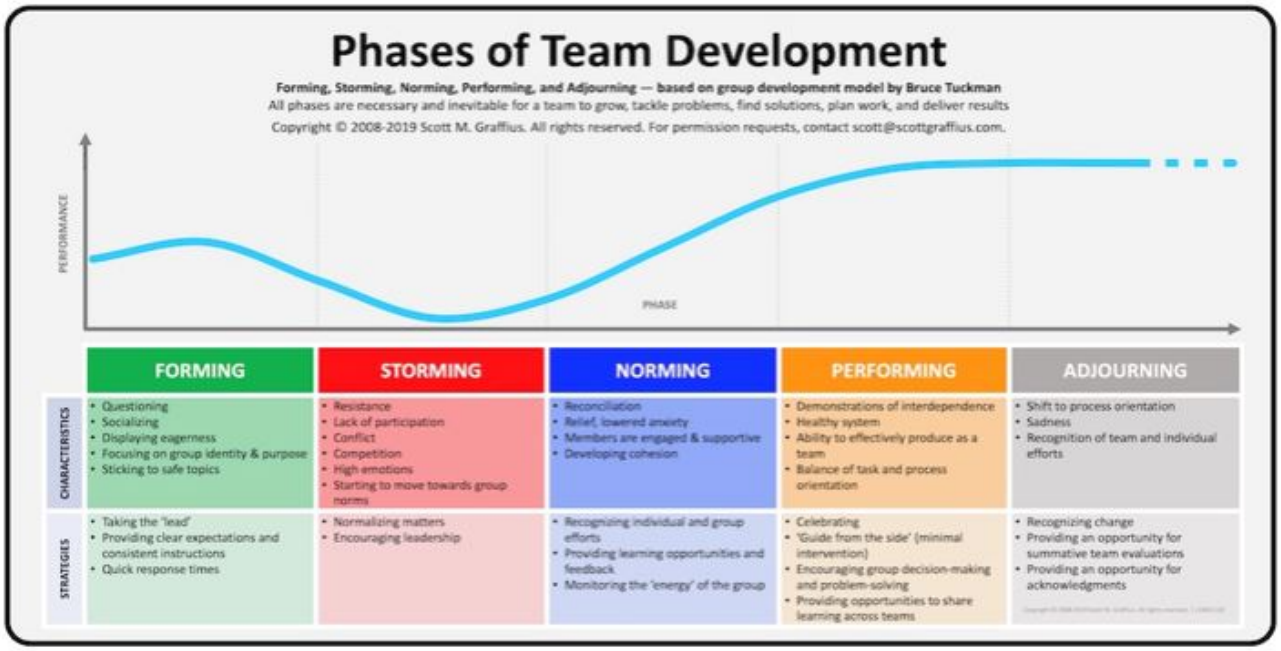
Accountability systems prompt and encourage people to keep promises and then monitor where those promises are kept. Accountability systems instill discipline to consistently repeat good practices. Elements of accountability systems that measure and communicate performance expectations include:

- a select few (12-15) indicators of overall performance; including measures used to improve the process and make daily performance decisions
- forums for two-way transfer of information
- long-term scheduling
- a formal problem-solving process
- clear, defined work processes

These system elements are building blocks to communicate performance expectations and results across the organization. Such a system organizes all the small things that allow your organization's teams to accomplish anything.

These disciplines will not appear overnight but can be built over time when the team leader and members intentionally work at it. Recognize that all teams will go through natural stages of forming, storming, norming, performing and adjourning:

Figure 2.7 Phases of Team Development



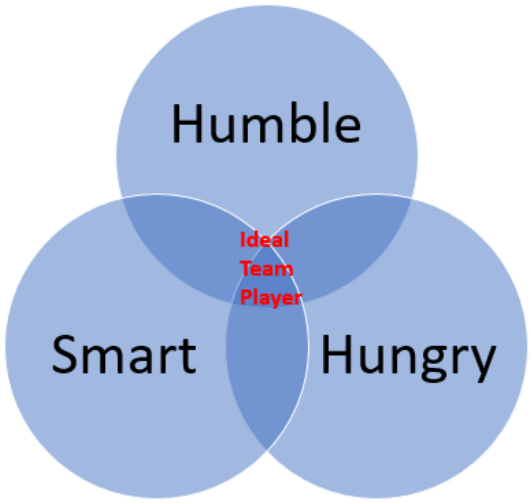
What makes a good team member? Effective teams require a mix of skills: technical/functional, problem-solving and interpersonal. But is it just skill that make a good team player?

"Coming together is a beginning, staying together is progress, and working together is success." -
Henry Ford

According to Pat Lencioni, an ideal team player exhibits three virtues:

1. **Humility** is the important of the three virtues. Humble team members lack excessive ego or concerns about status. They are quick to point out the contributions of others and slow to draw attention to their own. This virtue aligns with the Work of Leaders behavior Execution: Feedback: Offer More Praise.
2. **Hungry people are always looking for more** - more to do, more to learn, more responsibility. They almost never have to be pushed by their manager to work harder or longer; they are self-motivated. Healthy hunger is a manageable, sustainable commitment to doing a job well and going above and beyond when it is required. This is analogous to the Work of Leaders best practice behavior Execution: Momentum: Driven.
3. **People Smart.** Being people smart refers to a person's common sense about people – interpersonally appropriate and aware of what is going on within the group. They have good intuition and judgement about the subtleties of group dynamics. This virtue aligns with the work of Leaders best practice behavior Alignment: Dialog: Receptive.

As Lencioni (2016) admits, these virtues are not new or earth shattering taken one-at-a-time. It is the combination of the three that makes them powerful. If just one is missing in a team member, teamwork can be more difficult, if not impossible.



What happens if a team member only has one of the three?

1. Humble Only – this team member is pleasant, kind-hearted, unassuming, but does not feel or demonstrate a sense of urgency to get things done or have the ability to build relationships with others. They do not make waves but are left out of conversations and activities. Lencioni calls these team members "Pawns." The bottom line is that such team members will have little impact on team performance. Humble-only team member will survive long on teams that value artificial harmony and that do not demand performance from each member.

2. Hungry Only – this team member will be determined to get things done but with a focus on themselves and no understanding or concern for how they impact others. Lencioni calls these team members "Bulldozers." Hungry-only members can easily destroy a team but go unnoticed in organizations that place a high priority on results alone.

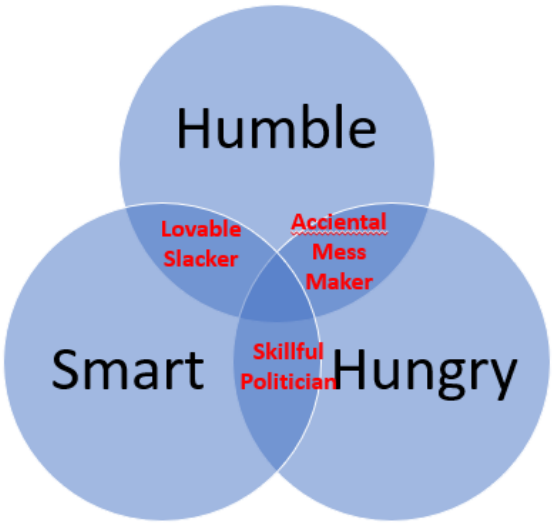
3. Smart-Only – this team member lacks humility and hunger but knows how to behave around others. They can be entertaining and likeable, but they have little interest in the well-being of the team or results. Lencioni calls these team members "Charmers." Bottom-line, they provide little contribution to the team.

What happens if a team member only has two of the three?

1. Humble and Hungry – Known as "Accidental Mess Makers" these team members generally want to serve he team and get results but lack any understanding of how they impact the rest of the team and create interpersonal problems within the team.

2. Humble and Smart - "Lovable Slackers," these team members are adept at working with others and not looking for attention, but only do what is asked of them, rarely seeking more. They have limited passion and commitment to the work of the team and need motivation and constant oversight or will put a drag on the team.

3. Hungry and Smart – "Skillful Politicians' these team members are ambitious and so skilled at team dynamics they often appear humble, but in reality they manipulate and scare other team members. These team members do well in organizations in which individual performance is valued over teamwork.



Lencioni (2016)

WARNING: It is not easy to identify these virtues and should not be done flippantly.

There are two areas where leaders should apply these three virtues:

1. **Hiring Team Members.** The most reliable way to ensure teamwork in your organization is to hire only ideal team players. Ask specific questions to tease out these virtues in candidates. Have a small group individually interview each candidate, share responses and observations. Do not ignore hunches as they will come back to haunt you later.

2. **Developing Team Members**.

While the three virtues are character behaviors versus competence behaviors, they can be developed/improved.

1. Developing Humility - This is the most nuanced of the three virtues and usually related to insecurity, maybe something rooted in childhood, family situation or a function of style. A manager may be able to help such an employee identify the root cause and admit the situation and then coach the employee to practice it. With practice, the employee may feel more comfortable with it.

2. Developing Hunger - A manage should give such an employee immediate, unambiguous feedback, repeatedly regarding their hunger behavior, or lack thereof. Praise the employee publicly when they exhibit signs of hunger.

3. Developing Smarts – Make it clear to the employee that it is not about intention that they do not recognize group dynamics or their impact on others, but quickly and lovingly get their attention to the situation so they can see and practice it appropriately.

In all three cases, it is most important for the leader to model the behaviors themself.

Once employees are hired with the three virtues and provided coaching to further develop them how does a leader embed these virtues into the culture and further ensure team effectiveness and results?

- Be explicit and bold with expectations for teamwork and the three virtues.
- Catch people doing it and hold them up as examples.
- Address any behaviors that violate these virtues, small and large. Provide opportunities for constructive learning.

2.3.4 Tools for Teams

Especially in the early stages of team formation, it is essential that the leader impart a clear statement of vision. What will things look like if everything goes right? What is the ideal state of the future? The best leaders maintain a strong sense of vision and an ability to communicate it.

The team should also work collectively on a mission statement. Why are we established? What is expected of us? Is our stated mission consistent with the vision?

Key strategies and tactics will be derived by the team from these considerations which must be revisited through the life of the team.

Initial team efforts may follow specific strategies, leading to tactics for problem resolution. For example, some strategies may be such things as:

- Improve throughput and finished product quality on Line 3.
- Stop going down blind alleys in R&D.
- Improve success rate of plant startups of new products and processes.
- Determine root causes of customer and consumer dissatisfaction.

- Identify market gaps as ripe areas for new product launches.

While these strategy examples and others like them may fall under an overall organizational mission, they may be best taken one at a time. Often, a useful first step is brainstorming.

2.3.4.1 Brainstorming

The leader may begin a meeting with big picture statements of vision and mission as necessary reminders to all. Then she or he may proceed to the immediate task at hand. Let us say the task is third on the list above, that is, improve the success rate of plant startups. There will probably be some questions for clarification and definition. Other questions may concern roles and responsibilities. These should be addressed, but it should also be recognized that many of their answers may come out, once participants put their minds together.

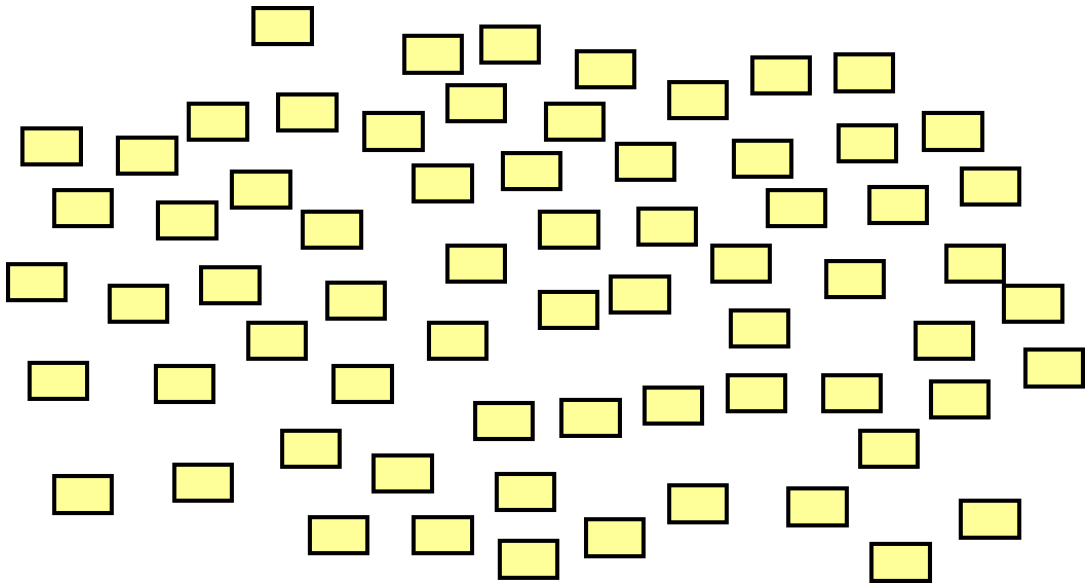
An effective group process is brainstorming. The goal is to get the ideas out independently – talking discouraged – without prejudice, judgement or ranking. Let the ideas emerge freely. They will be sorted and condensed later. It is helpful to obtain as many pertinent ideas as possible, but a number between 40 and 60 will usually serve well.

A technique to do this is to give each team member a set of 8 to 10 blank sticky notes and a dark, heavy pointed marking pen, asking each to state their ideas, one per note, as concisely as possible. Participants should hold onto the notes and remain quiet until others have finished.

Note that the facilitator may play a useful role here by assuring the ground rules are followed, by fielding questions, by reminding participants not to share ideas, and by generally monitoring the activity so everyone has an opportunity to generate ideas.

Next, the leader allocates an empty wall or white board and directs the participants to scatter their sticky notes randomly across that space. Notes should be intermingled so no one is aware of the sources. The outcome should look like Figure 2.8, except that each note should contain words.

Figure 2.8 A Scatter of Sticky Notes Derived from Brain Storming



2.3.4.2 Affinity Mapping

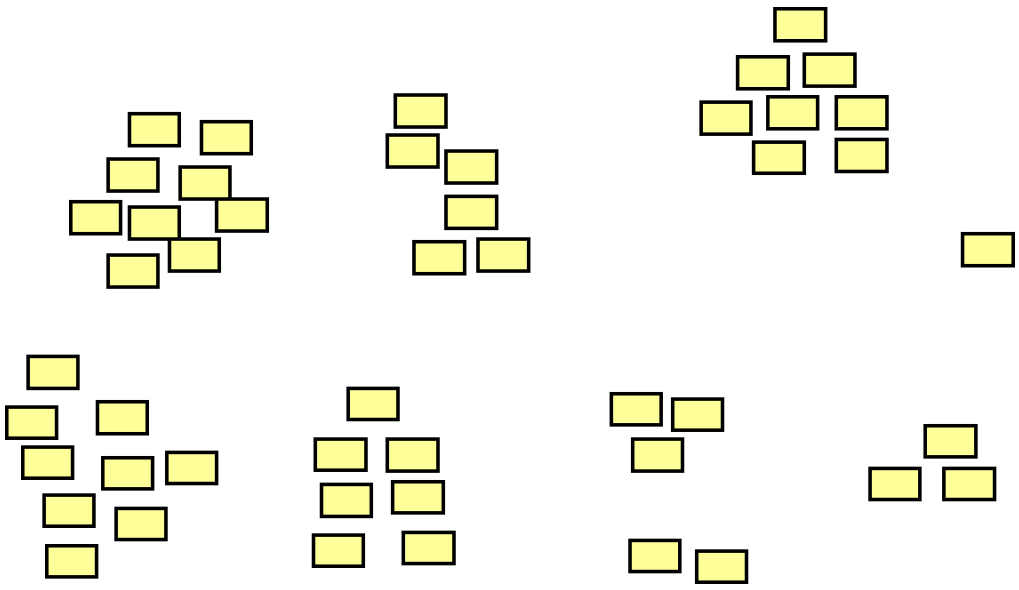
Once the ideas are out in the open, they can be consolidated into more manageable groupings. A popular tool for doing this is affinity mapping, a technique for aggregating ideas analogous to the way averages of numerical data represent group central locations.

Keying on the display of sticky notes, the leader invites 3 to 5 volunteers to approach the board and, without talking to each other, begin to move the notes into categories. These categories are undefined at this point; they are simply groupings that are intuitive to this subgroup of meeting participants.

Affinity mapping capitalizes on spatial thinking. An effort to shift from analytical thinking to spatial thinking is to use the other hand to move the notes – righties, use the left hands and vice versa. Another affinity mapping norm concerns note placement: if Person 1 moves a sticky note to a formed cluster, and Person 2 moves that same note to a different cluster, no violence is permitted. Differences may be resolved civilly following the mapping exercise. They are usually caused by differences of interpretation.

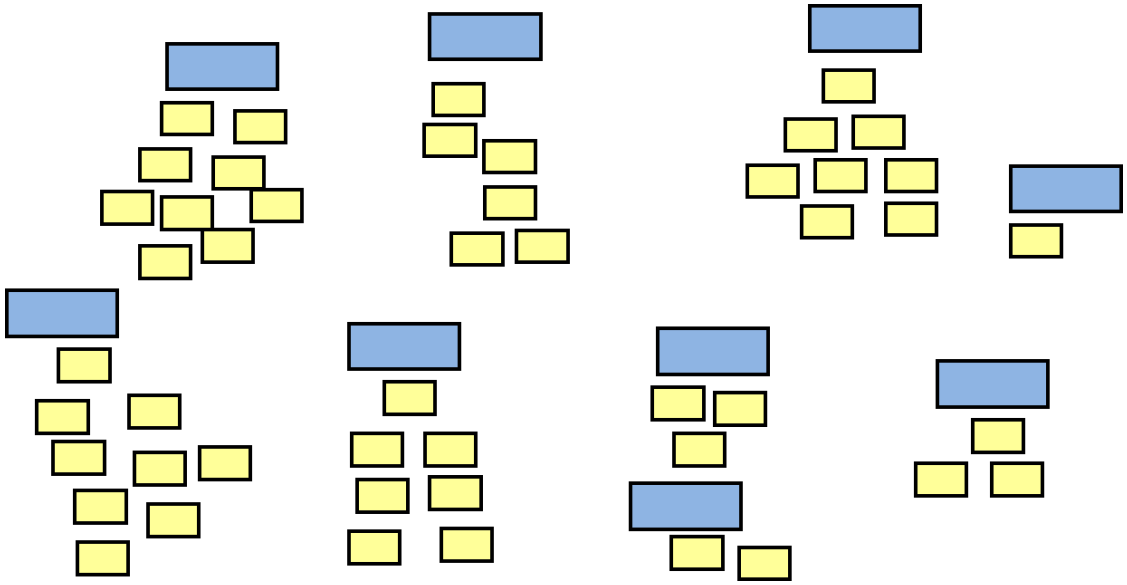
The leader can tell when the group is finished, even when group members cannot. The notes move more slowly, nearly to the point of halting entirely. At that point, the leader might decide to involve a second, independent subset of the participants to approach the board and refine the map. This process should only take a few minutes.

Figure 2.9 Affinity Map of Brain Storming Results



When this step is completed, the leader should ask this new group to place headers on each of the clusters that have been formed. Headers should summarize the cluster topic as succinctly as possible. Notice from the right side of Figure 2.10 that even an orphan idea deserves a header. It is often convenient to make duplicate headers using larger and different colored sticky notes than were used in the mapping exercise.

Figure 2.10 Affinity Map with Headers



Consolidation of ideas from brainstorming as made possible by affinity mapping has many uses depending on objectives. For example, in one situation the ideas were potential ingredients in pet food formulations. Affinity mapping categorized them into related components such as meats, cheeses, enhancers and flavorings. Representative items of each category were then used as factors in screening designs (See Chapter 2, Section 5) to learn of their influence on sales and market share.

2.3.4.3 Interrelationship digraphs

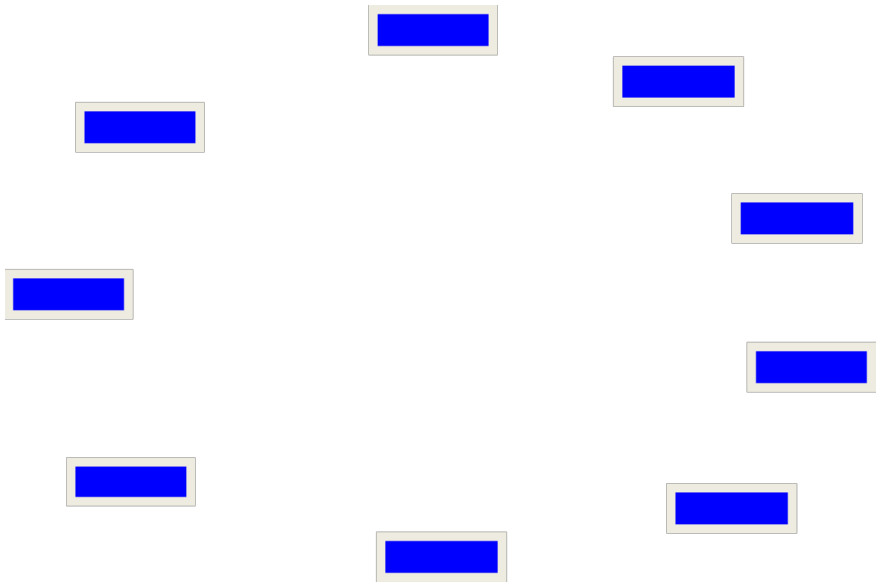
Suppose the group was formed and convened to assess reasons for a particular problem or behavior such as:

- Why doesn't R&D use designed experiments in their decision making?
- Why don't our plant managers and their staff take advantage of information from data to make improvement decisions?
- What is the cause of our consumer complaint handling delays and difficulties?
- Why are we losing sales of this popular brand?

Each group member is knowledgeable in the subject matter but may approach a solution from a different direction. Certainly, affinity mapping will get the ideas out, and the use of headers will provide summaries, consolidating them. Then what?

The interrelationships digraph is useful to provide causal relations among key drivers of the problem. This team tool is analytic, not spatial, so team members are encouraged to, and actually must, talk. Recall the recommendation in the precious section that duplicate headers be made. The leader of facilitator arranges the duplicate set in a circle as in Figure 2.11.

Figure 2.11 Affinity Map Headers Arranged in a Circle

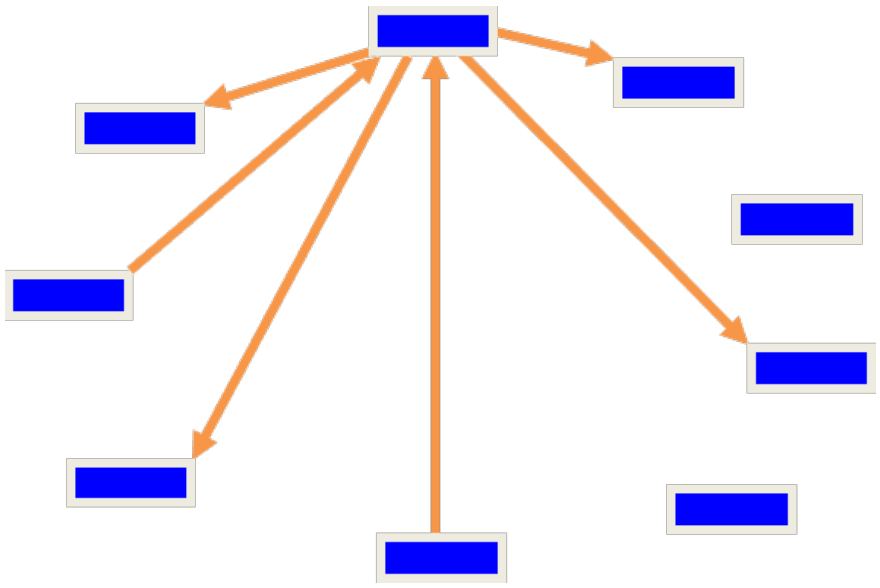


Notice that the original affinity map is preserved, headers and all, so it can be used as a reference for greater understanding of intent.

Next, the leader works to obtain group consensus regarding the relationship between the header at 12 o'clock and the header at 1 o'clock. Does the 12 o'clock header cause the 1 o'clock header, or does the 1 o'clock header cause the 12 o'clock header, or is there no relationship. The leader draws an arrow from cause to effect. Note that the quiver contains no two-headed arrows. The group must decide, and if no decision can be made, perhaps there is something amiss with the understanding of the headers – revert to the original affinity map for improved understanding and resolution.

Proceed from 12 o'clock to 2 o'clock, 12 o'clock to 3 o'clock, and so on until all the relations involving 12 o'clock are established. At the end of this first cycle, the interrelationships digraph might look like Figure 2.12.

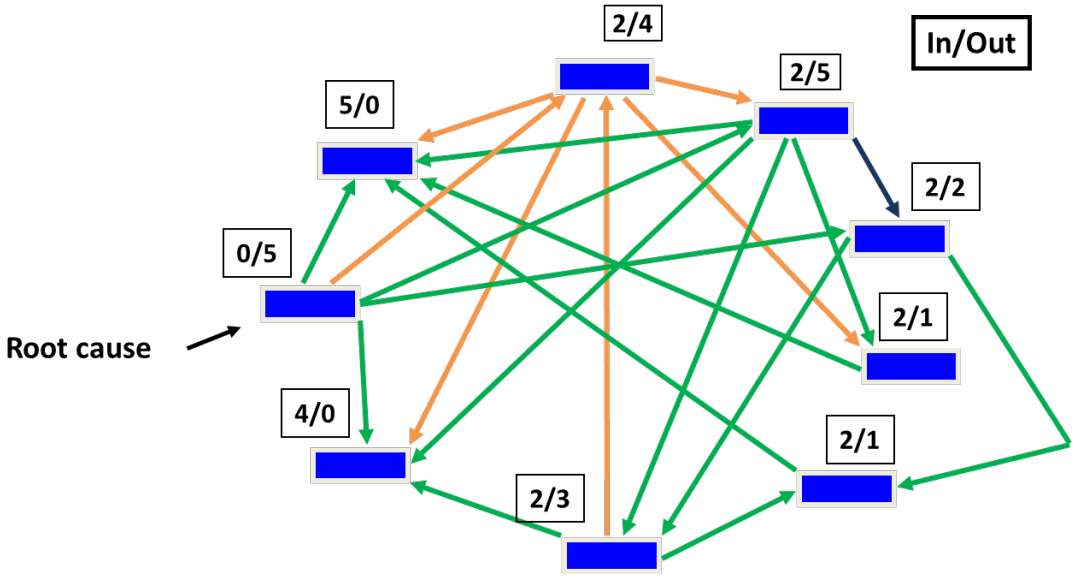
Figure 2.12 Interrelationships Digraph at the End of the First Cycle



Next, start with the header at the 1 o'clock position as the starting point and proceed to the relationships from there to 2 o'clock, then 3 o'clock, and so on. Then start with 3 o'clock, then 4 o'clock, and so on until all pairwise relationships are considered.

The final picture might look like that shown in Figure 2.13. Note that the orange arrows represent findings from the first cycle while the green arrows represent the combined findings from all subsequent cycles.

Figure 2.13 Completed Interrelationships Digraph Example



Next the facilitator counts arrows in and arrows out to and from each header. Two factors may serve as checks:

1. The number of arrows in must equal the number of arrows out.
2. If n is the number of headers, then the maximum number of arrows possible is n(n-1)/2.

Now, the headers with the most arrows out are the key drivers. Those with most arrows in are the effects. To make good progress, the team must focus on the key drivers first.

2.3.4.4 Multi-voting

Sometimes teams are formed to make decisions among categories or to set priorities among separate choices or alternatives. Multi-voting can come in handy in this situation.

As an example, a church music director retired, leaving an opening for new talent. A team consisting of administratively aware and musically talented members was asked to decide among candidates. The first step was to set criteria. This was accomplished using brainstorming, which produced 46 different ideas, and affinity mapping which consolidated the criteria to 12.

Each of 10 team members was given 4 sticky stars to assign to the criteria anyway they chose. They were asked to delay posting start to the board listing the criteria until all were ready. Results are listed in rank order in Table 2.14.

Table 2.14 Multi-Voting Church Music Director Criteria

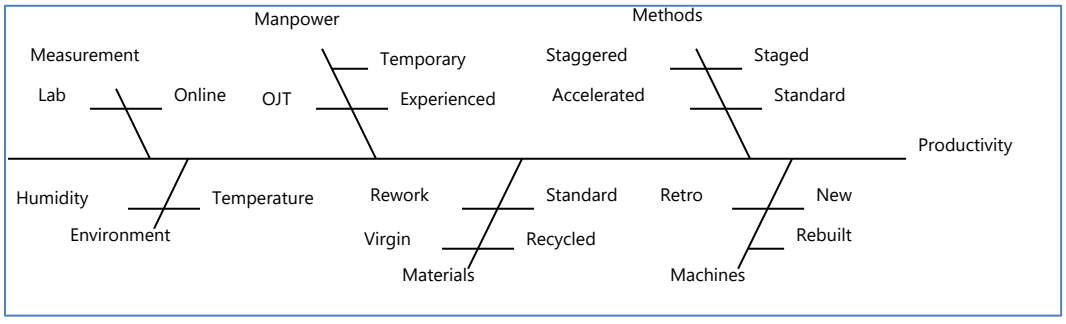
Criteria	Votes	Count
Team Builder	★★★★★★★★★	9
Multi-Talented	★★★★★	5
Spiritual	★★★★	4
Sense of Humor	★★★★	4
Sociable	★★★★	4
Inspirational	★★★	3
Joyful	★★★	3
Coach	★★	2
Caring	★★	2
Available	★	1
Experienced	★	1
Adventurer		0.2

As a surprise, it became clear that team building topped the list as a special need for this position and for the church as its internal culture stood. Candidate interviews were carried out with special attention to the criteria ranked by importance as designated by the number of votes.

2.3.4.5 Cause and effect diagrams

One early and very useful quality tool is the fishbone diagram (Ishikawa, 1968). Using it, problems sources are categorized by type such as those associated with methods, machines, manpower, materials, measurement and environment, with the major issue listed at the extreme right end – the head of the fish. A simple example is shown in Figure 2.15 where productivity issues are categorized into these, and sub-categories are added.

Figure 2.15 A Simple Fishbone Diagram



There are as many variations of the fishbone diagram theme as there are processes, and with careful attention to process detail, these diagrams can help organize process thinking, diagnosis and improvement.

An enhancement of the concept is the cause and effect diagram. It addresses situations where there are multiple issues that must be addressed. In figure 2.16, above, the issue is productivity, but in an increasingly complex world that and many others must be resolved.
To see how the cause and effect matrix works, consider the following example which follows the multi-voting story in the previous section.

Table 2.16 Cause and Effect Matrix for Church Music Director Search

	Team Builder	Multi-Talented	Spiritual	Sense of Humor	Sociable	Inspirational	Joyful	Coach	Caring	Available	Experienced	Adventurer	Weighted score
Weight:	9	5	4	4	4	3	3	2	2	1	1	0.2	
Candidate													
John	6	7.2	5.4	5	5	5.8	5.2	6.4	6.6	7	8.4	6.6	227.3
Mary	3.2	6.2	5.4	3.6	2.8	3.4	4	4.4	5.2	8	7.4	3.6	164.5
Lois	7	7.6	7.2	6	7	6.4	7.2	6.6	7.4	8.4	8.6	6	268.8
Frank	7.8	8	7	5	5.6	6	5.4	7.4	5.2	7.8	9	6.6	258.1

Criteria to be evaluated are listed across the top row. They are taken from the original affinity map (not shown) which generated the multi-voting categories and the weights shown in the second row. Note that weights may come from sources other than multi-voting.

In the body of the table are mean scores taken from grading sheets used by interviewers who attended the auditions and subsequent interviews. The final weighted score is the inner product of the candidate scores and the weights. For example, John's weighted score of 227.3 is his team building score of 6 weighted by 9, plus his multi-talented score of 7.2 weighted by 5, and so on. Following this logic puts Lois in first place followed by Frank. Moreover, it represents the combined thinking of all the team members.

As with fishbone diagrams, there are variations on the theme of cause and effect diagrams. Most situations are considerably more complex than the example above. For example, in many production systems, the input variables such as methods, machines, manpower and so on as shown on the above fishbone diagram have many, many subcategories. The same applies to healthcare organizations, education processes and every other process that can be named. Rows of the cause and effect matrix might number into the hundreds and might be divided hierarchically into categories and subcategories. Likewise, columns may be numerous well beyond the simple example shown here.

This means that there can be no hard and fast template for cause and effect matrices. Instead, each must be tailored to the situation and process at hand. That stated, there are some basic guidelines that can be useful. First, use a spreadsheet. Then:

1. List Output variables (Effects) across top of matrix
2. Rate Output variables on a 1-10 scale
3. List Input variables (Causes) down the left side
4. Rate the effect of each cause on each output variable (0,1,3,9)
5. Multiply across to calculate a ranking for each cause
6. List the causes in decreasing ranking order (construct a Pareto chart of the ranks)

Begin with a temporary spreadsheet similar to that shown in Table 2.5 and modify it to suit the situation.

Table 2.5 Cause and Effect Matrix Starter Template

		Process:											
		Date:											
			A	B	C	D	E	F	G	H	I	J	
		Output Variables											
		Weight:											
	Input Variables												Ranking
1													
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													

2.3.4.6 Additional tools for teams

Other tools for teams that should be in the facilitator's bag of tricks include process flow diagrams, Five Whys and Is–Is Not analyses. These are described briefly.

Flow diagrams are useful for creating team members' common understandings of the way materials, ideas and services move through a process. Before sensible action can be taken toward process improvement, all team members must hold a common understanding of it.

> Example: In one manufacturing facility, many batches were failing. A project leader began with the first stage of the process, working with a team of six employees responsible for running it. The leader asked the team members to each sketch a flow diagram of this first stage. He received seven different diagrams: one worker was not certain, so she drew two! Clearly, if workers do not agree on the process or do not understand it, they cannot possibly work to control it the same way or correctly. The plant engineer was called in to clarify the process flow so all workers had the same understanding. Process performance improved.

Five Whys is a simple group exercise for getting to root causes of well-defined problems. It lacks the thorough detail of cause and effect diagrams, but it applies group thinking systematically in successive steps to drill down to root causes of problems quickly. Of course, the exercise may require more than five steps.

Example: Production Line 3 stops more often than other lines.

1. Why does the line stop? Product is jammed at Stage 2.
2. Why is product jammed at Stage 2? The conveyor rollers do not spin properly.
3. Why don't they spin properly? They do not get lubricated.
4. Why don't they get lubricated? They are difficult for the maintenance staff to access.
5. Why are they difficult to access? There is no space between Line 3 and the adjacent line.

Is–Is Not analysis helps to define the problem more clearly by eliciting responses to questions such as the "who, what, why, where, when and how" may or may not have been involved. It can be useful to form a grid to help frame the problem.

Example: Finished product is sour.

	Is	Is not but could be
Who	Suppliers Y and Z	Any other suppliers
What	Sour taste	Other off flavors
Why	Shipping delays	Other out-of-specification measures
Where	Plant B	Plants A, C or D
When	January 1 – 14	Before or since
How	Lactic acid buildup	Accidental inclusions

For more on these tools see Hoerl and Snee (2020).

Section 2.3 References

Doyle, M. and D. Straus. How to Make Meetings Work, Jove Books, NY, 1982.

Everything DiSC. Productive Conflict training, John Wiley & Sons, 2017.

Hoerl, R.W., and Snee, R.D. Statistical Thinking: Improving Business Performance, 3rd ed., John Wiley & Sons, Hoboken, NJ, 2020.

Ishikawa, Kaoru. Guide to Quality Control. Tokyo: Asian Productivity Organization, 1968.

Karlgaard, Rich and Michael S. Malone. Team Genius: The New Science of High-Performing Organizations. New York, NY: Harper Collins, 2015.

Katzenbach, Jon R. and Douglas K. Smith. The Wisdom of Teams. Boston, MA: Harvard Business School Press, 1993.

Lencioni, Patrick. The Five Dysfunctions of a Team: A Leadership Fable. Jossey-Bass, 2002.

Lencioni, Patrick. Death by Meeting: A Leadership Fable. San Francisco, CA: Jossey Bass, 2004.

Lencioni, Patrick. Overcoming the Five Dysfunctions of a Team: A field Guide for Leaders, Managers and Facilitators. San Francisco, CA: Jossey-Bass, 2005.

Lencioni, Patrick. The Ideal Team Player: How to Recognize and Cultivate the Three Essential Virtues. San Francisco, CA: Jossey-Bass, 2016.

Scholtes, P.R., Joiner, B.L., and B. J. Streibel. The Team Handbook, 3rd ed., Joiner Associates, Madison, WI, 2003.

Senge, P. The Fifth Discipline, Doubleday, NY, 1990.

Zenger, John H and Joseph R. Folkman. The New Extraordinary Leader: Turning Good Managers into Great Leaders. New York, NY: McGraw-Hill, 2020.

Chapter 3 - Data Collection

Table of Contents

Preface

Here, we present principles and techniques for acquiring necessary and sufficient data for sensible, practical guidance with the ends being advancement of human well-being through the revelation of improvement opportunities and the identification of solutions to nagging problems. Discussions, presentations and examples focus on primary data issues including the theory of data acquisition, challenges inherent in the acquisition of data, the understanding and importance of data pedigree, the quantification of accuracy and precision under the heading of measurement systems analysis and finally, the essential planning of data acquisition, considering all the forgoing, for sound decision making. The latter is under the heading of the statistical design of experiments.

Section 3.1 – Theory of Data Collection

3.1.1 Objectives

The objectives of this section are: a) to define and discuss the critical terminology and concepts related to data and, b) to layout various broad themes concerning the use of empirical data in problem solving. The second objective might be described more generally as giving some guidance on what data can be used for.

3.1.2 Outline

This section develops key terminology and concepts relative to working with data including concepts around population, sample, processes, variable nomenclature, measurement scales, causation, statistical studies and analysis themes.

3.1.3 Variation

Variation is one of the fundamental characteristics of the empirical sciences. In all things measured, counted, or otherwise assessed in some way, there is variation among the items of interest and among several measurements or assessments of the same thing. There is also a great variety of items that people are interested in that can be measured or assessed in some way. The term "item" means anything (object, event, phenomena) upon which measurements may be made. Table 3.1 is a sample of such items.

Table 3.1 Examples of Variable Quantities

VARIABLE	EXAMPLES
item dimensions	length, width, volume, "size", thickness, diameter, depth, ovality
item properties	weight, color, density, tensile, elongation, breakage strength under a loading
time	between events, to completion, task/mission or other duration, turn-around, personal or sick time taken by employees in a year; days
money, value	cost, loss, sales receipts, income, expenditure, tax, net worth, numerous others
environmental conditions	temperature, humidity, rainfall, snowfall, storm size
characteristics of people	gender, race, religion, occupation, height, weight, education level, blood type, eye color, many others.
yes/no, binary (counting)	conforming/non-conforming, sale or no sale, agree/disagree, works/broken, has/has-not condition

event (counting)	destructive, accident, crime, environmental, citings or occurrences, defects

Variation may be defined as a measure of the extent to which items are different or changing from one time to the next, from one item or scenario to the next and from one measurement assessment to the next. In thinking about data and variation it is useful to have a set of terms and concepts, concisely defined, that help us put meaning around the concept of variation and the associated problem we are trying to solve.

3.1.4 Data Pedigree

Data pedigree (See Section 3 of this chapter) refers to the overall history of any data we have to work with. This includes:

- The origin or source that generated the items of inquiry (either from a process or a static set of items).
- The sampling process or how the sample of items was selected
- The measurement process that generated the resulting numbers
- The initial recording or "data entry" process
- Any possible editing (changes, additions and deletions) of the data along the value stream
- Technical knowledge of what the individual data items represent
- Any previous data analysis or other operations on the data.

Ignoring considerations for these elements can be the source for numerous types of errors. This is particularly true with large data sets or where multivariate data are concerned, as the larger the data set, the greater the likelihood for handling and other errors. Thus, whenever we are presented with a data set, we ought to ask questions that speak to these several elements. In addition, data checks in a database can be constructed to find numerical data errors, missing data and nonsense values that may inadvertently creep in.

3.1.5 Population and Sample

In statistical science, there are always two things going on. First, there is the data we have, and second, there is the source of that data. Here the word "source" means a *population of interest*. The *population* is just the set of all items of interest in any particular study. Again, the term "item" should be understood to mean virtually anything that can be measured, counted or assessed, numerically, in some way. An item is not necessarily an object. It may be understood to mean a property or other phenomena that can be measured in some way. Thus temperature, time, and color are not objects but properties that are connected with something. It is important that any population be well defined, by which is meant that if an individual item is presented, it should always be possible to say that the object either belongs to the population of interest or does not. Essentially, a *population of interes*t may be anything we define it to be so long as it is well defined. The older term *universe* also means the same thing as a population of interest.

Sometimes the population of interest is a simple set of existing items – a finite set of objects for example, large or small. In those cases, the concept of a *frame* can be useful in identifying the population.

A *frame* is an accounting in some way of the entire set of a finite population of interest. The frame provides a unique identification and listing of the objects in the population. It may be used in various ways to decide how a sample should be taken. In its most perfect form, a sample from a population of interest is determined by a random selection of numbers from the frame. This gives equal selection opportunity to all items in the frame. Note that this may not always be economically feasible, possible or even desirable. In many cases of objects, a rational sampling methodology may be preferable insofar as there may be important patterns in the population such as stratification, clustering or other unequal distribution of the population elements to consider. The term "rational" means we are thinking about how the sample is selected and selecting the sample in a judicious manner that allows a maximum of learning about the items of interest. The example following the section on processes will illustrate this concept.

3.1.6 Processes

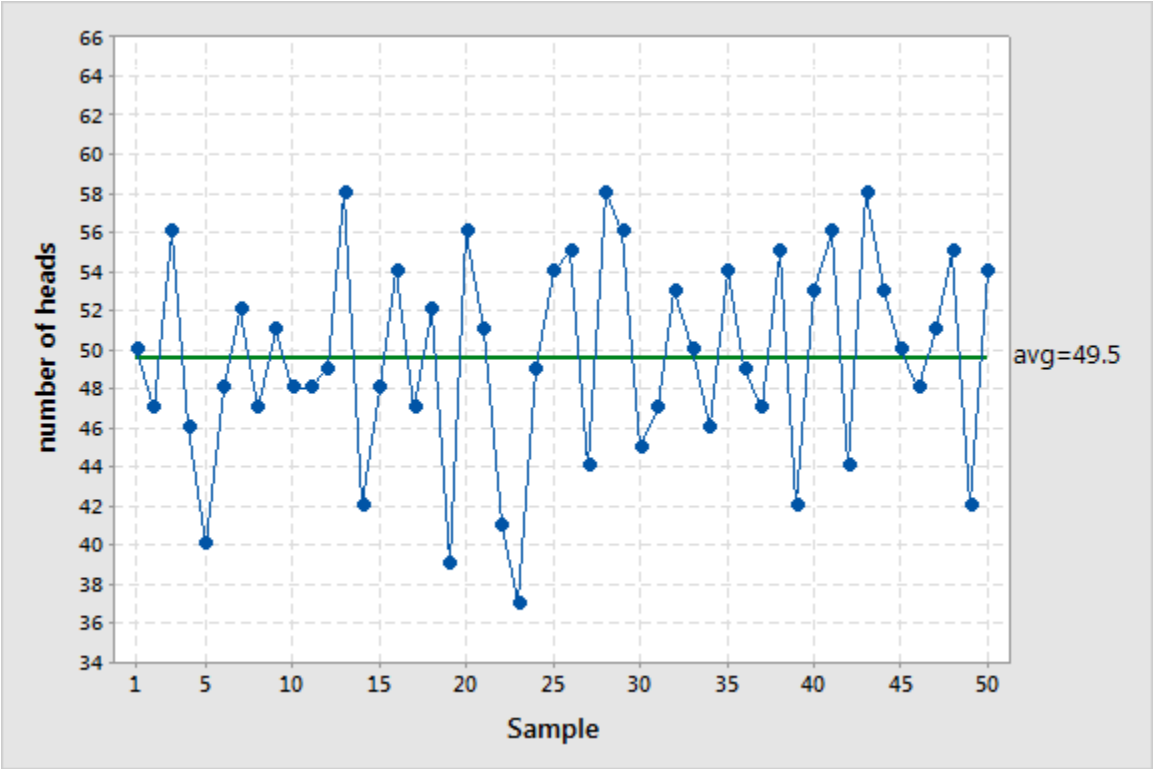
A *process* is a kind of dynamic or virtual population of interest as opposed to a static set of existing items. The "process" concept is a construct used to communicate that recurrent phenomena are continually generated in time order. That said, objects or items continue to emerge (and expire) in a process. More generally, a process may be thought of as a kind of "black box" that contains all of the mechanisms required for generating items – the *cause system*. The cause system is essentially a set of process variables that are random quantities in their own right and in combination give rise to the variation we see when items are generated, observed and measured. In a process we may think of the population as containing an infinite set of items – the entire set that could be generated by the process given enough time. It is important to note that a process may work slowly or fast, as far as generating the items of interest. For example, if the items of interest are US presidents, then we typically only get one observation every 4 or 8 years; if the items are "small molded" components from a high-speed manufacturing process we may get thousands of objects produced every day. Thus, speed of a process may be an important consideration in any effort. In some cases, a very slow process may be considered as the finite population that currently exists. This might be the case in many biological phenomena, for example, trees in a forest or fish in a pond at the present time.

A most important consideration in any kind of process is whether or not the process is undergoing some kind of change(s) over time. Such changes may be systematic, gradual, intermittent, erratic or otherwise occur in numerous ways. Theoretically, when a process is <u>not</u> undergoing such changes or upsets, we say that it is *stable* or in *statistical control*. We may further say, following Shewhart (1986), that if a process is in a state of statistical control, it is possible to predict, at least within some limits, and with an associated probability, what will happen in the near to medium term (as long as control is maintained). *Stability* is the more general term; the concept "Statistical control" grew out of manufacturing applications (i.e., Statistical Process Control). Process stability may be further characterized as having predictive merit insofar as we can predict what will happen in some future time using data from the current state of the process. When a process is stable, the random variation we see can usually be

modeled by some theoretical model or distributional form such as a normal distribution. That gives meaning to any kind of "prediction." For a stable process, any sample obtained from that process is considered a random sample, representing the population of interest. When a process is not stable, the task is always to first understand what is causing the upsets or changes and to bring the process into a stable state. In certain cases, an unstable process or distribution of items may be the natural state. In such cases, the information provided and the understanding of the upsets may be keys to future planning (see Example 3.1 below).

***Example 3.1 Stable Process*:** The following data represent 50 results, obtained one after the other, of tossing a fair coin randomly 100 times and recording the number of heads.

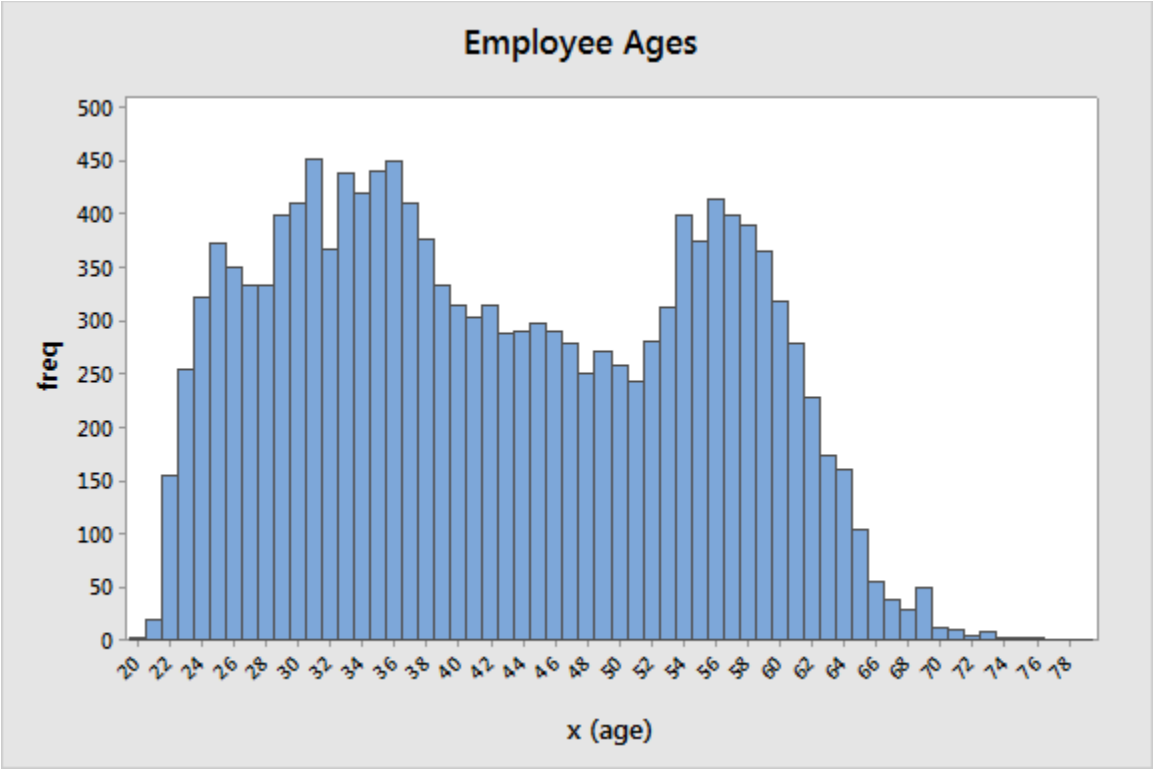
Figure 3.1 Number of heads appearing in each of 50 sets of 100 coin tosses



This data exhibits a stable pattern. Variation is coming from chance causes to be expected in similar games of chance. It can be shown that, in theory, the mean and standard deviation from tossing a coin 100 times and observing the numbers of heads are 50 and 5, respectively.

Example 3.2 Instability, Mixed Distribution: The following data represent the ages of 14,463 employees in the engineering organization of a large corporation.

Figure 3.2 Employee Ages



This data appears to represent two distinct distributions: a wave (on the right side) of "baby boomer" employees nearing retirement and a new wave of younger employees through mid-career age. This data may be construed as a kind of localized instability in time. That is, the wave we see may seem to be instability but is more likely a natural state for this phenomenon. The bi-modality reflecting the ages and number of employees may occur just as distinct in future representations but with possible shifts in the "wave" location. It can still provide key management information on the age distribution of employees and coupled with other data could be used as a planning tool to forecast future short-term retirements and hiring needs.

Example 3.3 In a large retail establishment, management wants to know something about the number and type of objects sold during normal operating hours (between 9 AM and 9 PM, open 7 days a week). It is hoped that such a study and the resulting information can be used to design a better sales system that would maximize sales. What data should they collect?

First, this is a process. Data emerge in time, and the volume of data collected is proportional to the duration length of the study. A completely random sample of sales receipts by day would be interesting but not terribly informative to the objectives of the study. Various causes of a sale may be listed: gender and age of the customer, time of day purchased, article purchased, on any given day, what was the weather like, day of the week the purchase was made, was the article

purchased a sale article, what department was the article purchased from, was this a repeat purchase, what was the article's cost, did the overall experience influence the purchase? Still other variables can be defined. Each of these variables will define a stratum in the larger shopping value stream. Some strata will be more important to capture than others. All of this information would provide answers to numerous questions leading to sales improvements. There is, however, the additional problem of data collection. How will this data be collected? What method will be used to gather and organize the data? Will it always be possible to obtain such data, or will there be some sales that result in missing some or all of the associated data? The former problem - what data is to be collected – is the study design; the later problem – how is the data to be collected – is the implementation of the study and often the more challenging part.

3.1.7 The Cause System: Special vs Common Causes

A system of causes is said to give rise to the variation among items originating from a process. The cause system may be described using two broad categories. On the one hand there are the *common causes* (Deming 1966). These causes are shared equally or held in common by each item produced. That is, the number of distinct causes and the degree to which these are operative (have impact) on the objects produced is identical from one object to another. All of the many small causes so classified are random variables in their own right, but the degree to which they are random is a stable form. The older terminology is "*random cause*" according to Shewhart (1986). A cause is said to be *special* if it affects some items in a different way. That is, one or more items have some variation component that others do not have. The older term is "*assignable cause"* according to Shewhart. So, any cause system of variation may be dissected into common causes, belonging to all objects, and special causes, belonging only to some objects.

There is a further distinction that some authors may sometimes make regarding a difference between special and assignable causes. Sometimes a special cause may not be assigned a reason, and some assignable causes may actually be predictable, even systematic. In the former case, extremes or anomalies of variation may sometime occur without a cause being identified or with many possible causes. This may just mean that further work needs to be done to reduce the probability of their further reoccurrence. The absence of a definitive assignable cause should not negate the fact that such causes are special. In the latter case, when a special cause is predictable, it may be the case that we can incorporate such causes into the common cause category until such time as the cause can be isolated and either removed or mitigated to minimal effect. Or it may be the natural state of affairs such as in Example 3.2.

3.1.8 Random Variables, Observations, Individuals

When a target population of interest is identified, it is useful to designate the variables of interest using some nomenclature (e.g., x, y, z, etc.). It is equally important to "know what any 'x' is", particularly when there is more than one variable of interest. It is very easy for people to confuse meanings about just what is being observed and measured. When a variable, x is identified, it is called a *random variable*. In any type of statistical study there will always be one or more so called random variables (RV) of interest. The notion of a random variable, in simplified

language, contains two key items: a) it is a variable quantity that can change and b) it takes its value at random. The later distinction is further complicated in that there is some distribution associated with the random variable and in practice we typically do not know what that is at the outset nor if it derives from a stable process. Indeed, for a process, major objectives are to determine how to model the random variable(s) of interest and to determine if the resulting data indicate stability. A random variable is really a theoretical construct, useful for purposes of description - of a process response or of a sampling methodology.

Once we have defined one or more random variables, when a measurement is made, we call that an *observation*. Thus, if x is a random variable defined as the weight of a package, if a single package is measured and x=12.23 lbs., the 12.23 is an observation of the random variable x. Observations are the real manifestations of the more abstract random variable concept. A further point of terminology concerns several observations. The several observed values we have are called *individuals* or *individual measurements*. For the package example, the 12.23 is further distinguished as an individual observation. This is so we do not confuse individual observations with any summary statistics we might compute using several observations (for example an average of several observations).

3.1.9 Statistics and Parameters

A *statistic* is any number determined on the basis of a random sample of individual observations. The simplest statistic would just be a single individual observation; but this would rarely occur in practice excepting for cases where observations were exceedingly rare. Many kinds of sample statistics are familiar, such as means, percentiles, rates, proportions, standard deviations, extremes, correlation coefficients and others. Some of these will be taken up in further chapters of this manuscript. The important thing to note here is that a statistic is also a random variable with its own distribution, generally distinct from the distribution of individuals. Another important concept is that of estimation. That is, any sample statistic estimates some aspect of the population or process of interest. Thus, the sample mean is an estimate of a population mean; the sample standard deviation is an estimate of the population standard deviation. The population value being estimated is referred to as a *parameter*. There are several types of parameters and several ways to estimate parameters.

3.1.10 Types of Statistical Studies

W. Edwards Deming (1966) defined two concepts that speak to the difference between an existing set of objects as the population of interest and a process generating objects in time. We call studies where the objectives and actions concern an existing set of objects *enumerative* studies. Studies about a process are called *analytic*. In an analytic study, action is directed either at bringing that process into a stable state and/or to improve the process in terms of its capability. Additional objectives in an analytic study might be to discover what associated process variables and their range of values cause excessive variation in that process. In an enumerative study, the action is directed to the existing state of affairs – the current population that exists. In some cases, both enumerative and analytic methods might apply.

A second classification, equally important, concerning statistical data is the concept of *observational* versus *experimental* studies and associated data. In an observational study, variables are just observed in the natural state of affairs – whether from a stable system or not – good, bad or indifferent! One purpose of such a study is to assess the state of a process/population at the present time and to determine a baseline from which further actions can be planned. In other cases, the observational study is the only feasible data that we can get. Think of weather phenomena or studies involving historical phenomena. The observational study is often a starting point in any particular effort.

Experimental studies aim to determine the effect, if any, of one or more independent variables on a response (dependent variable) and to discover what causes change in a response. The sub-field of statistics, experimental design, deals exclusively with this type of study (See Section 5). An important point in observational studies is the notion of *cause*. In general, we cannot precisely conclude that one thing causes another – we can only discover association or correlation. For example, shoe size is correlated with reading ability in the general population. People with large feet tend to read better than those with small feet. But we cannot say that large feet are the cause of good reading ability. The actual underlying cause is the fact that older people have larger feet than children do. People who have "large" feet are generally older and have been trained to read and have more experience in reading than children do. Thus, age and schooling are the real causes of reading ability and these just happened to be correlated with shoe size.

In addition, observational studies, by contrast to experimental studies, may be the only viable way to study something – due to practicality, politics, economics, ethics or some other reason. Observational studies generally provide information on the current state or capability of a system or a process or to inform us on real world cases. They are particularly important in medical science, psychology, social science and biology. The difference between experimental and observational has been summarized smartly by Sir John Herschel (1792-1871, English), inventor of photography (2009).

"*Experience may be acquired in two ways; either, first by noticing facts without any attempt to influence the frequency of their occurrence or to vary the circumstances under which they occur; this is observation; or, secondly, by putting in action causes or agents over which we have control, and purposely varying their combinations, and noticing what effects take place; this is experiment."*

3.1.11 Nomenclature

The two broad classifications of data are *variable* and *attribute* data. Variable data applies where X is a number on the real line continuum, such as weight, time, volume, dimensions, density, etc. There will always be the degree of resolution and other properties of measurement to contend with (See Section 4 on measurement systems).

Attribute data occur in two types. Type 1 attribute data occurs where the object inspected either has the attribute or does not. For each unit inspected define X=1 if the attribute is present and X=0 if it is not. The attribute can be anything we define it to be, but in many instances, the

occurrence of the attribute is a *non-conforming unit*. The older terminology is *defective unit*. In type 1 cases, the sum of the variable X over the n units inspected gives the number of units, r, in n having the attribute or number of non-conforming units. In that case the number of units, r, having the attribute (a random variable) in n units inspected is $0 \leq r \leq n$.

With this type of attribute, and when sampling a process or very large lots, the statistic that is appropriate is the proportion having the attribute in the sample. This is estimating the true but unknown proportion, p. For example, if in a sample of n=1142, r=13 were observed having the attribute, then the estimated true proportion would be r/n = 13/1142 or about 1.14%. Many types of statistical phenomena may be modeled using this type of variable. Table 3.2 gives several illustrations.

Table 3.2 Type I Attribute Examples

Attribute Description, Type I	Population Description
Number of defective units, in n	A large lot, or a Process
Number of heads, in n tosses of a fair coin	Process (fair coin flips)
Number of people favoring a proposition	The larger group of people being sampled
Number of customers purchasing something	A sample of 1200 customers
Number of passengers in n cancelling a flight	A given flight rout
Number of failed units in an active redundant system	A fleet of similar units having the system

In a type II attribute, X counts the number of *events* that occurs over an *observational region*. The event can be of various kinds and the observational region can be a single item, a group of discrete items or a single continuous region, including time or other spatially defined regions. In theory, any number of such events can occur within the region observed. The region may also be called the sample in this context. The defining characteristic of this type of attribute is that there is no theoretical upper limit to the number of events that can occur in the sample observed. It is also only possible to count the events – not non-events, unlike the type I attribute. It is assumed that events occur randomly over a homogeneous region, and the average number of events is proportional to the size of the region observed as long as these regions are homogeneous. The mean number of events, μ, occurring with the region, or the rate of occurrence, λ, completely governs the probabilistic behavior of the variable X. If the rate is given and we want to apply the model to an observation region of size t, then the mean in that region is λt. Care must be taken in that t and λ must carry the same units (hours, days, square feet, etc.). A great many phenomena follow this description. In this type of attribute, the appropriate statistic is generally a rate in the units of events per unit. Table 3.3 gives several examples.

Table 3.3 Type II Attributes, "event" and Observational Region

Attribute Description, Type II	Observational Region
Flaws on paper	12 square feet of paper
Surface blemishes on a new vehicle	A single vehicle or several vehicles
Lost time accidents in a quarter	Three months
Equipment breakdowns	1 work week
Surface blemishes on a bearing race	1 box containing 20 bearings
Automobile accidents	Between 7-8:30 AM at local intersection
Robberies on weekends	Saturday/Sunday in local city
Flue contractions in a city	The given city in a specified year
Emergency calls coming into a hospital	From Friday night to Saturday night
Attrition in a large company	Monthly
Aircraft rare events	Over an extended period for a large fleet
Flat tires by puncture	In a group of people in a 6 month period
Death by a kick from a mule	Prussian Army, 19th century
Bad credit card transactions	In a day
Fire alarms	A local city in a specific period
Truancy	In a specific school in a month

3.1.12 Scale Classifications

In addition to the broad classifications outlined above, there are four general types of scales that any variable can take on. These are: a) Nominal, b) Ordinal, c) Interval, and d) Ratio scales. The most primitive of these is the Nominal scale where the variable is simply a qualitative classification variable, name or label without any "good", "better", best" or "largest", smallest" connotation. In this type of scale, we can only classify and count the various classifications or categories and arrange these in some tabular or graphical form. Some simple examples where the variable X is in nominal scale include color, make of automobile, occupation, part nomenclature, city, state, country of origin, disease, race, religion, political affiliation. In some cases, the variable may be binary such as marital or single, retired or not, employed or unemployed, veteran or not.

A variable is said to have the ordinal scale if it has all of the characteristics of a nominal type scale and it possess an order, where it is not possible to determine any precise difference between adjacent ordered values. For example, in a survey where the variable might be cast as "excellent", "very good", "good", or "poor", there is definitely an order, but it is not possible to state that the difference between "excellent" and "very good" is the same as the difference

between “good” and “very good”. All such “good”, “better”, “best” variables work like this. In many cases, ordinal type data occurs with metrics involving satisfaction on some level. Many types of quantitative variables might be classified using an ordinal scale such as weight (light, heavy, very heavy), size (small, medium, large), net worth (low, middle, high, upper), depth (shallow, deep, very deep).

A variable is said to have the interval scale if it has all of the properties of the ordinal scale and it is possible to determine the quantitative difference between adjacent values. In addition, this scale does not possess a true 0. That is, X=0 does not mean absence of the quantity being measured. The prime example of this is temperature where the difference between 35 and 36 degrees F is the same as the difference between 101 and 102 degrees F, in terms of the physics. In addition, 0 degrees F does not mean the absence of heat. Variables on an interval scale can be discrete (3 years, 10 degrees, 4 weeks) or fractional (2.8 years, etc.)

The highest scale a variable can take on is the Ratio scale. Ratio variables are pure numbers that we can add, subtract, multiply and divide. All of these operations make sense for this type of variable. In this case X=0 means there is nothing. Measurements such as time, weight, height, dimensions, money and distance all have the ratio scale. The point in defining these scales is that some study designs and statistical methods and techniques depend on the type of scale being used.

3.1.13 Purpose and Themes for Data

Purposes for collecting data and several associated analysis themes are useful to discuss at the start of any project. This gives some guidance on how a study should procced – tasks, data and other activities. The objectives for any project should be stated and agreed to at the start. This could be very detailed or quite general, just stating something about the problem we are trying to solve. In some cases, the problem statement may not become clear at the start and may simply evolve as the study proceeds. Every case is somewhat different in this respect. To solve most problems, some data will be needed at the start and along the way to providing feasible solutions. The following list may help to categorize the general purposes of problem- solving efforts.

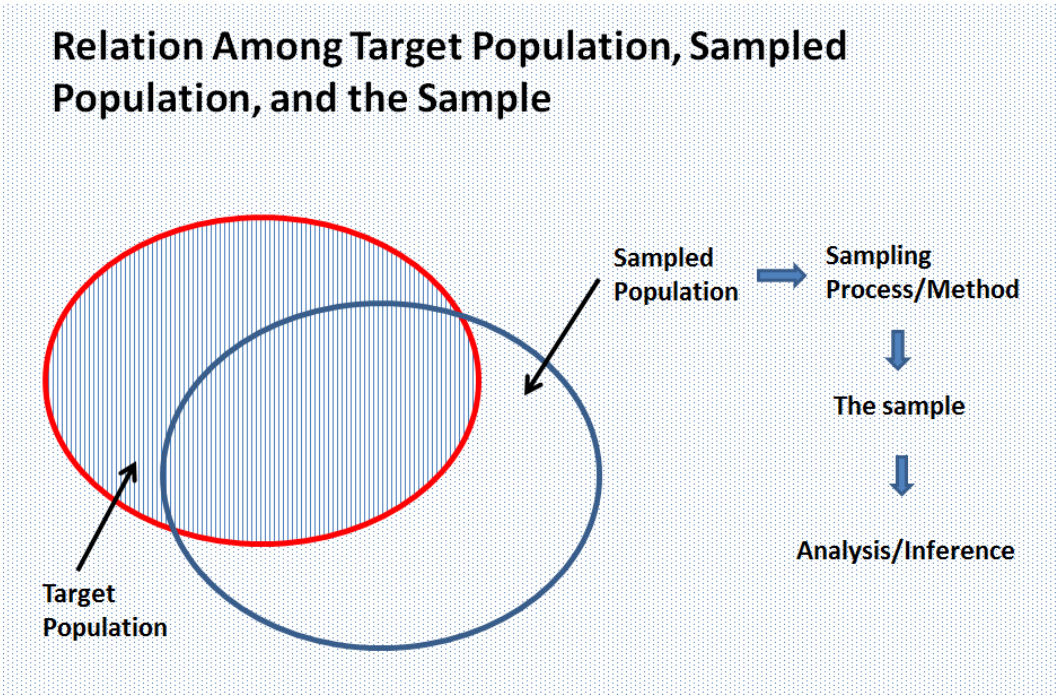
- Characterizing a present state of affairs
- Comparing two or more groups
- Meeting a requirement or comparison to a standard and capability
- Separating variation
- Optimizing, goal setting, modeling
- Predicting/planning

Each of these six categories has use at some phase of most projects, and most tasks will fall under one of these categories. Of particular importance are the first and last categories. In almost any process improvement effort there will be a current state of affairs and a record of past performance. To move forward with an improvement strategy, it is important that we know something about the present and/or past state of affairs. From a task centered point of view, the problem here is typically getting good data on the present and past states. In most organizations,

some data exists in some form, and the task centers in finding the appropriate data sources and getting that data into some good structural form. If the process is new there still may be some helpful data. We may find some similar process that has historical data available, or we may be able to benchmark several similar processes. Many present state characterizations are further complicated by multivariate conditions, numerous sub-processes and process outputs. As things can become complicated to the point where progress may be limited, it might be best to break up the larger process into smaller, less complex sub-tasks. In any event, some characterization of the present state will typically be needed, and this begins with getting reliable data.

During efforts to sample a present state, there may be problems associated with sampling the correct population of interest. There are the population of interest, the population to be sampled, and the sampling process, itself, to contend with. This is illustrated in the figure below.

Figure 3.3 The Complexities of Sampling.



Characterization will take on varying degrees of complexity, depending on the available data set. In many cases simple tables, summary statistics and graphical portraits may be sufficient. In other cases, multivariate data and many sub-processes will require more complex characterization. Characterizing the present state of a process can be regarded as a kind of process capability assessment (good, bad or otherwise). In such studies, data are observational, at least within the purview of specific portions of a process being studied. There is a tendency among some people to go directly to problem solving and experimentation prematurely during an initial data gathering and characterization process. This causes confusion, likely resulting in serious compromises to accurate characterization.

All statistics are about the past since they are based on data from processes or from existing populations that have already occurred. Managers and other decision makers usually want to know what will happen in the short- and long-term future. That could be very near term, such as three days, or longer term such as three to six months in a new product introduction, or very long term, such as 20 years or more in an aerospace risk analysis. It is important to keep management wants and needs in mind while planning data collection, analysis and reporting.

Between characterization of the present state and reporting there will likely be numerous other tasks involving data and data analysis in some form. Much of this, as stated above, can be categorized under the four additional areas: comparing two or more things, estimating compliance with a requirement or a standard, variation separation, and optimizing, goal setting or modeling. For example, an activity involving tolerance design can be considered as a form of optimization. Tolerance stacking, on the other hand, can be considered as a form of meeting a requirement. That is, in a tolerance stacking problem, the question is, "What is the final variation/capability of the stack if we consider the several components involved from a statistical point of view?" Consider this simple example. Suppose three components are assembled in a linear stack. Let x, y and z be the dimensions of these components. The final assembly is $h=x+y+z$. Assume further that the three components have a common normal distribution with some known mean, μ, and standard deviation, σ. Further, the components are selected and assembled randomly and independently from the manufacturing process. What will be the capability of the final dimension h? If the variance of each component is σ^2, the variance of h will be $3\sigma^2$. The standard deviation of the stack (h) will then be approximately 1.73σ. The mean of the distribution of h will be 3μ. The final distribution of the dimension h will be centered on 3μ with a four-sigma spread of $\pm 4(1.73\sigma)$ or $\pm 6.9\sigma$. From this, a realistic tolerance can be specified for h. This should be compared to the min/max method that uses the three components at their extremes in both directions. In that case, using $\pm 4\sigma$ for each component individually, the tolerance of h would be $3\mu \pm 12\sigma$. Then we have saved (reduced) the final tolerance by about 42%.

Comparing objects or other items and variation separation has to do with separating out the components of variation and locating possible special causes of variation, so the main tool will be experimental design and its many variations. For many projects, such activities will consume much of the time spent. Modeling is in large part a form of prediction. In modeling we are typically fitting a distribution to a data set or creating a multivariate regression model. In complex cases we might use Monte Carlo simulation as part of a model. In both cases, the aim is prediction – that is, what will likely happen in the future.

When data is at hand in any part of a project and where some kind of analysis is needed, we can think of the analysis as falling under several general themes. These themes are summarized below.

- Center
- Spread/variation
- Shape/distribution
- Outliers/extremes
- Time order
- Control/stability

For any single variable we can always ask for the center of the data, the variation in the data, the shape or distributional form of the data, whether there are any outliers or unusual values evident, is there a time element at play, and do the data suggest stability. Answering these questions will bring us a long way through any data analysis task. This summary is highly simplified, and there might be numerous methods (analytical and graphical) under each theme, but the general themes are relevant to any data analysis task. With multivariate cases, we can further add correlation structure to this list. Thus, with any type of project where there is empirical data, these purposes and analysis themes should be kept in mind as the project proceeds.

Another way to think about data broadly combines center with spread/variation. In the broadest sense, the mean or center of the data functions as the "signal" and the standard deviation represents noise. This would be the case in Example 1 where the mean and standard deviation are about 50 and 5, respectively. In that example the signal in n=100 tosses is x=50, and the variation around that is the noise that is due to purely random variation. The ratio "signal/noise" or SN ratio is sometimes used to assess the relative magnitude of the signal to the noise. In this example SN=50/5=10. If we were to increase the sample size (n>100) we would find that the SN ratio increases. For example, for n=500, SN = 22.4, for n=1000, SN=31.6. In both cases the SN has been calculated theoretically. In many cases of empirical investigations, the actual observation, y, might be related to an auxiliary variable, x, and may be thought about in terms of a model equation: $y = f(x) + \varepsilon$. In this model the *f(x)* represents the signal component of y and ε the unexplained random error component. There may also be several "x" variables.

If y represents an individual's lifetime earnings and x is the number of years of education achieved, although it is true that income over a lifetime increases on the average (signal), there is also a random component to this affecting the signal that may be due to any number of reasons. Thus, ε may be related to: a) the degree of motivation and diligence that people apply in finding good jobs; b) a health element in some people; c) a reluctance to re-locate to a better job; d) a skill debit despite the education; or e) the occupation. There are likely many other factors that create the random component.

Section 3.1 References

Deming, W. Edwards *Some Theory of sampling,* Dover Publication reprint, originally published by John Wiley & Sons, NY, 1966.

Hogg, Robert & Tanis, Elliot *Probability and Statistical Inference*, 7th edition, Pearson, Saddle River NJ, 2006.

Herschel, John Fredrick William *Preliminary Discourse on the Study of Natural Philosophy*, original publication, 1830, Cambridge University Press, 2009.

Shewhart, Walter *Economic Control of Quality of Manufactured Product*, ASQ Quality Press, 50th Anniversary Commemorative reissue, originally published, Van Nostrand Company, (1931) 1980.

Shewhart, Walter *Statistical Method from the Viewpoint of Quality Control*, Dover Publications reprint originally published 1939, Graduate School, Dept. of Agriculture, 1986.

Van Belle, Gerald *Statistical Rules of Thumb*, Wiley series in Probability and Statistics, NY, 2002.

Section 3.2 – Challenges of Data Collection

3.2.1 Objectives

Given the grounding in the theory of data collection presented in the previous section, we set the stage here for the practice of data collection together with common problems and challenges involved. The intent is that readers from all endeavors, regardless of statistical engineering application will find the contents of this section supporting prior to the data collection effort.

3.2.2 Outline

We begin with a discussion of what is meant by data collection as taken apart from acquiring data passively, or without question. This is followed by a discussion of some common problems experienced prior to and during data collection. Then, we conclude with a summary and discussion of major challenges faced during data collection.

3.2.3 What is data collection?

"Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer queries, stated research questions, test hypotheses, and evaluate outcomes," (Anastasia, 2017).

This definition may not be sufficient for all applications, but it serves well to emphasize the need for key elements of data collection strategies to:

- Acknowledge that data are gathered for the purpose of measuring, not proving, information
- Point out that data gathering must follow a logical, consistent system
- Emphasize that the intent of data gathering is to resolve issues. This may be to explore new possibilities or to find solutions to nagging problems

Alternatives for data collection are to either accumulate all the data generated by a process or to sample the data using a subset. In most situations, complete enumeration is either impossible or impractical. Sampling is the only solution, and in that case, we must decide how to sample in order to provide representative, unbiased estimates of the whole population.

Guidance is provided by the principle of "rational subgroup sampling." Following this principle, sampling is carried out under the same conditions for each process segment. An example is fill volumes taken from containers filled by a multiple same spout machine. Each spout should be sampled the same way so that resulting data might be combined to derive an unbiased estimate of all fills exiting the overall filling process composed of the full collection of spouts.

Of course, this principle applies regardless of the nature of the process – chemical/physical or sociological. Subsections of the "process" are sampled in the same manner.

The actual numbers of samples to be taken depends on the needed precision of the population estimate to be derived, and it depends on the nature of the data. There are two broad types of data:

- Quantitative Data. These are data that describe quantities, values or numbers, making them measurable. They are usually expressed in numerical form, such as length, size, amount, price, percent and duration. They are spaced along a continuous scale so that between any two numbers, there is another number.

- Qualitative Data. These data, on the other hand, are categorical rather than numerical in nature. Generally, they take a counting form as the number of defects or defectives, the number of voters in a district, and the number of phone calls received in a fixed time period. Unlike quantitative data, they are not measurable along a continuous scale and are gained mostly through observation. Narratives often make use of adjectives and other descriptive words to refer to data describing appearance, acceptability, defective or non-defective, "yes" votes, and other qualities.

3.2.4 Common Problems of Data Collection

An understanding of common problems found during past efforts to extract meaning from large data sets is helpful to the formulation of plans for avoiding future data collection obstacles. Some problems experienced include the following:

- A lack of management planning for data integrity, including the necessary investment in computer equipment and software for high quality data maintenance, including backup and security
- Inappropriate database structure to meet the needs of data analysis and interpretation
- Failure to ruggedize data collection systems to language and cultural differences around the globe
- The absence of a written and communicated plan for data collection
- Conflicting information due to lack of database planning and resulting in inconsistent and uncoordinated data systems
- The routine data collection of data inadequate in amount to be informative of product or process problems
- Instances of large quantities of missing data due to undetected measurement equipment failure
- Time pressure on data collectors due to overwhelming workloads and responsibilities, resulting in data omissions, illegible records and entry errors

Consequences of improperly collected data include the inability to obtain accurate information, resulting in incorrect organizational decisions.

3.2.5 Challenges of Data Collection

In this section, we re-visit some of the common problems of data collection and discuss their challenges to the researcher or statistical engineer. Challenges of data collection have been discussed by Kwadamah & Brobbey (2015) and include, but are not limited to, the following factors:

3.2.5.1 Subjective Data

Researcher fatigue. The process of conducting focus groups and interviews, the process can be stressful, and exhaustion is a key element to the success of such groups and interviews. Fatigue can degrade the quality of the data obtained. Thus, the researcher must manage their own and their subject's fatigue. The researcher must:

- be a good listener
- be able to handle diverse personality types
- be able to get quiet participants during focus groups engaged
- enable everyone to contribute

Interviewer fatigue can also be an issue. Haste must be avoided so thoroughness and accuracy can be assured.

Location of interviews. Interview location can be important because it can influence a person's responses. In fact, the location of an interview is an indispensable constituent of data collection to which the researcher must be aware. As an example, suppose that the interview was scheduled in an administrator's office. The interviewee may associate this office with meetings held there to discuss behavior or academic-related matters. So, anyone asked to come to this office for an interview may be expecting a different type of discussion, resulting in their hesitation to release any information. The interviewee may even provide erroneous responses for fear of being victimized for unpopular answers. Studies have shown that interviews should be conducted at a neutral site that is conducive for both the researcher and respondent. Interviewers should talk with operators as close to the process as possible and where interviewees are most comfortable.

Literacy constraints. Limitations of an interviewee's reading and writing ability can negatively affect the data gathering process. A researcher who conducts an interview with questions that are verbose can place the respondent in an awkward situation. Some respondents may feel humiliated by their inability to understand key words in the interview question, leading to a negative influence on the quality of the response. In advance of the interview process, the researcher must study the literacy levels of the respondents he/she wants to survey and adjust the level of the survey to the appropriate level of understanding. Be aware that an inferiority complex can set in when respondents begin to ask for clarification of interview question words. They may even avoid eye contact with the researcher at this point. Findings have shown that data quality depends on the literacy level of the respondents. Respondents with low literacy faced challenges in comprehending differences among the survey's responses, such as strongly

agree, moderately agree, neither agree nor disagree, moderately disagree and strongly disagree. The design of surveys must factor in the literacy level of the respondents and adjust wording to avoid ambiguity. This can improve data quality.

Lengthy Data Collection. The data collection process can be negatively affected by the time span of completing a survey or by how much time it takes to gather the surveys. If the survey is too long or the interview takes too much time respondents may become uncomfortable, resulting in inappropriate responses to survey or interview questions. If respondents are in a hurry to complete a survey or to end the interview process, they may provide useless information. For example, in extended surveys, respondent hunger of thirst may set in. Interviewers will be wise to anticipate these and other comfort needs during the data gathering process.

Interviewer time constraints. Deadlines press on everyone, including those carrying out interviews. Questionnaires should be designed succinctly to save time all around.

Insufficient data. Care should be taken during the interview process to determine if the frequency and lengths of responses are diminishing with surveys taken over time. A researcher who relies on survey data may see a decline in response over time.

3.2.5.2 Objective Data

Measurement validity. The lack of validity of the measurement instrument is a threat to the data collection process. An ineffective measurement process may have an error large enough to require much more data to make conclusions than may be practical to obtain. A measurement system analysis (MSA) study is often a requirement when any method is introduced to collect data on process response. Such studies are addressed later in this chapter for both quantitative (variable) and qualitative (observable) data. The statistical engineer should be knowledgeable of how to design and analyze MSA studies to obtain estimates of both repeatability and reproducibility error.

Representativeness of the data. It is not enough to simply know how large of a sample size is required to make decisions. The representativeness of the sample is equally important. For example, if a statistical engineer determines that a sample of size of 250 widgets should be collected from a production lot in a warehouse, the method to be used to identify the sampled widgets is also important. If a representative sample of the lot is desired, it would seem logical to take as random a sample as practicable. Would it make sense to take all 250 widgets from the same skid of widgets in the warehouse? Of course not! That skid represents just a snapshot of the time it took production to make the entire lot! Should we take a random sample from each skid? Doing so may take more time than is available because it would require opening most or all the boxes of widgets! What can be done? It may make more sense to take random samples from a corner box of each layer of a skid for all skids. This would be a more systematic sample as the skids are typically constructed in layers in order of production. So, taking samples from the corner of the skid by layer, means sampling periodically in order of production.

3.2.5.3 Other problems associated with data collection

Irrelevant or duplicate data collected. Collecting accurate and pertinent data should be the objective of every data researcher. In some situations, data are collected that have nothing to do with the problem at hand. In others, the carelessness of collecting data results in duplicate data records, requiring extensive clean-up prior to data analysis. A statistical engineer must be sure to collect data which serves to address the problem and that duplicate records are avoided.

Pertinent data is omitted. Variables essential to the study at hand may not have been considered or may have inadvertently been eliminated from the data set. The statistical engineer should work with IT and other collaborating disciplines to assure that data are collected on variables which may have been excluded from the current database. Recommendations for data collection improvement should include acquisition of systems or sensors, as appropriate.

Erroneous or misinterpreted data are collected. In situations where data may be relevant and clean, they may still be devoid of information. Measurement systems may lack accuracy or precision or both. Two following sections of this chapter address these issues. Section 3.3, The Importance of Data Pedigree, addresses some consequences of erroneous and misinterpreted data and presents methods of pitfall avoidance. And Section 3.4 on Measurement Systems Analysis provides specific methods to assure both the accuracy and the precision of methods used to guide improvement.

Database format is poorly organized. Scientists and engineers typically report their data in user friendly tables with rows and columns marked by headings and other explanatory information. Simple spreadsheet software used for reporting purposes permits generation of simple graphs and charts which can be imbedded in written reports and presentations. By contrast, most statistical software packages expect data one record at a time. For example, a report friendly table with r rows and c columns requires r times c records to be statistical software friendly. These formatting differences can cause data processing headaches. While some statistical software provides convenient workarounds such as routines for data stacking, the statistical analysis still requires manual intervention prior to statistical analysis. Careful planning of data base construction is to meet the needs of reporting and analysis alike is essential to quick and efficient data analysis.

3.2.5.4 Large Data Bases (Big Data)

The emergence of very large data bases has created the need for more sophisticated analytical approaches such as machine learning. There are some particular big data challenges to data collection that mirror some of what has been discussed but which also include some new challenges.

Identifying the right data to collect. All businesses have access to tremendous quantities of data from hundreds of customer, operational and financial sources. Given the cost of data collection, it is essential that businesses take a discerning approach to Big Data. Big Data strategies need to be backed by a clear understanding of the desired outcomes, cost and return on investment. For example, the most successful marketing companies can identify which data sources can turn

insight into action as a means of helping to build predictive customer models to drive more business opportunities. First-time strategies need to begin with a handful of key objectives against which data collection and analysis can be regularly reviewed. Taking on too much data from too many sources can overwhelm a business and cloud the effectiveness of a strategy.

Integrating Big Data into multiple departments. Big Data has a value that lies in its practical applications, such as improving results within sales, marketing and other departments. Statistical engineers can plan for and achieve results that allow multiple departments to understand relationships among process variables and product attributes as a means of optimizing process performance.

Data regulation and compliance. Superhero movies like to use the adage "with great power comes great responsibility." Likewise, with Big Data come big issues of compliance. One of the by-products of the sheer volume of data being collected is that many businesses are accessing and storing data in relatively unstructured environments. The more progressive companies have discovered the advantages of environments, such as Hadoop, which allows for the storage of both structured (typical databases) and unstructured (images and text) data. However, it should be noted that information can be collected and acted upon without sensitivity of data security compliance. This neglect puts both data and the business at risk! In fact, if the database covers several locations, it is necessary to consider the ever-increasing number of data laws, as such as those found in the EU. Also, there are also more commercial reasons to keep data clean and compliant. It is said that the average customer database contains between 25% and 30% useless records, and if the data is only three years old that only 10% of the records will likely be of value. While these estimates most likely will not apply to a process engineering database there is always the possibility that parts such as thermocouples may fail, and their data will of no value. It is also likely that products will change, and new measurements will be needed and old ones discarded. The statistical engineer should consider the cost of acquiring the data and the return on the investment.

Choosing relevant data analysis partnerships. Big Data presents such a large challenge that it often makes business sense to partner with a service provider that specializes in data insight and analysis. With conventional data analysis, the analyst collected the data and performed the analysis on their own computer using either a programming language, such as FORTRAN (in years past) or R (popular for statistics in recent years), or a commercial statistics package, such as JMP or Minitab (among many others). By contrast, the application of recently developed software enables massive datasets to be properly analyzed, promoting improved understanding which can be incorporated into business strategies. In that manner the real return of Big Data may be clearly achieved.

Section 3.2 References

www.cleverism.com/qualitative-and-quantitative-data-collection-methods/

Cowan "Design Education Based on an Expressed Statement of the Design Process," Proc. Inst. Civ. Engrs., vol. 70, part 1, Nov., (1981): pp. 743-753.

Jewel, T.K. "A Systems Approach to Civil Engineering Planning and Design," Harper and Row, New York, 1986.

Kwadamah, Forgive and Brobbey, Priscilla. "Challenges of Data Collection Process: A Review of Existing Literature," *Global Academy of Finance and Management*, (2015): https://www.coursehero.com/file/30151995/ForgiveAFAarticleGAFMpdf/

Lewis, T.K. "Observations of Problem-Solving by Engineering Students," *Australia Journal of Education*, vol. 18, no. 2, (1986): pp. 172-183.

Sharp, James J. "Methodologies for Problem Solving: An Engineering Approach," *The Vocational Aspect of Education*, 43:114, (1991): pp. 147-157.

Walker, E.A. et al. "Goals of Engineering Education," American Society for Engineering Education, Washington, D.C., 1968.

Section 3.3 – The Importance of Data Pedigree

3.3.1 Objectives

The purpose of this section is to explain the critical importance of data quality, and how this can be evaluated through documentation of the data pedigree. The term "data pedigree," its core elements, and how to practically utilize a data pedigree, are explained in detail in Sections 3.3.5 and 3.3.6.

3.3.2 Outline

We initially discuss the issues that can occur with analysis of poor-quality data. We then review the related concepts of "information quality," the US Food and Drug Administration's (FDA's) "data integrity," and the legal profession's "chain of custody." Based on these comparisons, we define data pedigree and the key elements that should be documented in a complete pedigree. We then provide practical advice on utilizing a data pedigree when analyzing data.

3.3.3 Data Quality

Analysis of data is obviously critical to statistical engineering. While there is obviously much more to statistical engineering than data and statistics, devoid of analytics, the discipline would not consist of much more than hype, slogans and fanfare. However, as data sets have gotten larger and larger, and easier and easier to download from the Internet, information about the quality of the data is often lost, assumed or overlooked. This situation should not come as a surprise, because the vast majority of statistics textbooks present data sets as "random samples" of impeccable quality. It is therefore understandable that some statisticians and engineers assume that data are "innocent until proven guilty." Too often, they are not even under suspicion.

Beneath the exciting sound bites on Twitter and Instagram about analytics, there is also growing awareness of a dark side of analytics, one in which models frequently fail with potentially disastrous consequences. For example, many have written about the disaster at the Duke Genomics Center, in which four cancer gene signature papers were retracted because the results turned out to be invalid, in part because of discrepancies in the data the Duke researchers analyzed. Unfortunately, it is likely that people died because oncologists utilized the results of these papers in prescribing treatment to those battling cancer (Kolata 2014).

Similarly, numerous articles have reviewed the Challenger space shuttle disaster of January 28, 1986 (Dalal et al. 1989). In this case, NASA scientists held a conference call to decide whether it was safe to launch the shuttle, given the abnormally low temperatures at Cape Canaveral that day (31 degrees F). The team reviewed available data on the relationship between temperature and O-ring failures, but someone had deleted observations with no failures. As shown by Doganaksoy, et al. (2006), this omission led to a decision to launch, when any reasonable analysis of the full data set would have revealed that launching at that temperature would be extremely dangerous. The entire crew of seven died.

Even the flagship scientific journal Science is not above publishing faulty research due to questionable data. In 2015, Nature retracted a 2014 paper on attitudes about same-sex marriage, in part because of "certain statistical irregularities in the responses" that were analyzed in the paper. (http://www.sciencemag.org/news/2015/05/science-retracts-gay-marriage-paper-without-agreement-lead-author-lacour).

An example illustrating the potential errors than can result from questionable text data is the recent article in the journal Science (Vosoughi et al. 2018), in which researchers from MIT analyzed over 4.5 million tweets on 126,000 topics. They ultimately concluded: "Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information..." In fact, the spread of false information reached "critical mass" (defined as reaching 1,500 people) *six times faster* than true information.

We argue that these incidents reflect one root cause of the ongoing "reproducibility crisis" in science (Ince 2012). While much of the effort to address this crisis has focused on standardizing analyses and providing original source code used by authors, we feel that a major source of problems has been overlooked: data quality.

In our view, data quality has often been overlooked in textbooks because it is not easy to mathematize. Conversely, the impact of data quantity is easy to mathematize, hence most statistics textbooks incorporate formulas for sample sizes needed to estimate parameters with a desired degree of precision, as well as "power curves" showing how large a sample size is needed to detect an effect of a given magnitude with a specific probability. Such formulas are of course, technically valid and potentially useful. But what if the original data is faulty? How useful are these formulas and power curves then? Unfortunately, many practitioners of analytics have no formal training in data quality, hence they are likely to overlook this potential root cause.

3.3.4 Benchmarking Other Disciplines

Taking this concept further, Kenett and Shmueli (2014) define the concept of information quality (InfoQ) as the potential of a dataset to achieve a specific (scientific or practical) goal using a given empirical analysis method. That is, the InfoQ depends on the goals of the analysis, and the specific methods employed. In Kenett and Shmueli (2017), these authors note the critical role of data and information quality in addressing the reproducibility crisis. In addition, they provide a formal, quantitative framework for evaluating information quality.

InfoQ is determined by eight dimensions that are intended to be assessed by consideration of both the data and the goals of the analysis:

1. Data Resolution: the measurement scale and uncertainty
2. Data Structure: the degree to which the data are comprehensive with respect to the goal
3. Data Integration: how disparate data sources have been integrated
4. Temporal Relevance: the time frame of data versus the goals

5. Generalizability: the degree of relevance of the analysis results beyond the immediate data
6. Chronology of Data and Goal: the degree to which the analyses (versus data) are synched with the needs of the decision maker
7. Operationalization: the ability to act based on the analyses – actionability
8. Communication: presentation of the results in the right way at the right time to decision makers

These dimensions can then be incorporated into the following formula to produce a number, the InfoQ, defined as the utility (U) for a specific analysis (f), on a given data set (X), with respect to a given goal (g):

$$\text{InfoQ}(U,f,X,g) = U(f(X|g)).$$

See Kenett and Shmueli (2017) for details on conducting such quantification.

Note that InfoQ is dependent on the specific goal of the analysis intended. Therefore, it is not an inherent attribute of the data set itself. We argue that a related but distinct concept, data pedigree, which is an inherent attribute of the data set, is needed to accurately evaluate data and information quality. Before defining data pedigree, we first benchmark a related concept from the legal professions of many countries.

Interestingly these legal professions seem to be way ahead of the scientific and analytics communities when it comes to understanding the importance of "data quality." In legal circles, data quality means documenting the integrity of evidence. The legal professions of English speaking countries typically use the term "chain of custody" for evidence, as opposed to data quality, but of course evidence is the "data" analyzed in a courtroom. Basically, the chain of custody refers to documentation of how the evidence was originally obtained (e.g., legal versus illegal search), and its movement and location from that point on, until presented in court.

The Legal Dictionary (https://legal-dictionary.thefreedictionary.com/chain+of+custody) states the following concerning the importance of documenting chain of custody for legal evidence: "Proving chain of custody is necessary to 'lay a foundation' for evidence in question, by showing the absence of alteration, substitution, or change of condition." We argue that documenting the pedigree of data, ensuring freedom from alteration, substitution, or change in condition, is equally required to "lay a foundation" for any statistical analysis.

The Legal Dictionary goes on to state: "Court-rendered judgments and jury verdicts that are based on tainted, unreliable, or compromised evidence would *undermine the integrity of the entire legal system*…" (emphasis ours). We argue that statistical analyses based on "tainted, unreliable, or compromised" data sets have, in fact, undermined the integrity of the entire scientific system. In science we refer to this undermined integrity as "the reproducibility crisis."

The US Food and Drug Administration (FDA) utilizes a similar concept, *data integrity*, which they define as the completeness, consistency, and accuracy of data. The FDA states (FDA 2016, p. 2): "Complete, consistent, and accurate data should be attributable, legible,

contemporaneously recorded, original or a true copy, and accurate (ALCOA)." These concepts bare obvious similarities to the concepts of chain of custody, information quality, and existence of a "gold standard."

The FDA points out that data integrity is a foundational component of current good manufacturing practices (CGMP), which forms basic requirements for organizations that manufacture pharmaceuticals in the US. Further, these manufacturers are expected to maintain "metadata", which provide context for properly interpreting the data and evaluating integrity. The FDA notes (FDA 2016, p. 3): "A data value is by itself meaningless without additional information about the data." For example, units of measurement and time stamps would be simple examples of required metadata.

The FDA also requires an "audit trail" to document the "who, what, when, and why" of data. For example (FDA 2016, p.3): "Electronic audit trails include those that track creation, modification, or deletion of data (such as processing parameters and results) and those that track actions at the record or system level (such as attempts to access the system or rename or delete a file)." Clearly, there are strong similarities between the FDA concept of an audit trail and the legal profession's concept of chain of custody, highlighting the importance of understanding what happens to data over time.

In our view, analytics, and scientific inquiry in general, could benefit significantly by adopting the same rigor in data quality as the legal profession has in its concept of chain of custody for evidence, and the FDA has with the concepts of data integrity and audit trail. Hoerl and Snee (2019) refer to such an approach as documenting the data pedigree, borrowing the term "pedigree" from animal husbandry, that is, show dogs, race horses, etc. Obviously, the value of a yearling racehorse depends significantly on the quality of its pedigree.

3.3.5 The Need for Documentation of the Data Pedigree

Data pedigree is defined as:

> Documentation of the origins and history of a data set, including its technical meaning, background on the process that produced it, the original collection of samples, measurement processes utilized, and the subsequent handling of the data, including any modifications or deletions made, through the present. (Hoerl and Snee 2019).

A proper data pedigree should include each of the elements listed in Table 3.4.

Table 3.4 Core Elements of a Data Pedigree

- What the data represent; that is, a basic explanation of the underlying subject matter knowledge of the phenomenon being measured, including units of measurement

- Description of the process that produced the data, such as a financial process, healthcare process, manufacturing process, etc.

- The number of samples that were subsequently measured, and a description of how they were obtained from this process, including the timeframe

- The specific measurement process used to assign numbers or attributes to the "samples"

- The existence (or lack) of recent analyses of the said measurement system, such as gage R&R studies, calibration studies, etc.

- The history of the data, documenting the chain of custody - who has had access to the original data, what if any changes or deletions have been made - and access to the "gold standard," i.e., access to a copy of the original data that can be verified

Common sense needs to be applied to the data pedigree. If someone is gathering data to improve his or her golf game, obviously the degree of rigor required in documenting the data pedigree is minimal. However, in medical research, or when public safety is involved, the pedigree should be rigorous, to enable evaluation of whether it meets the high standards of the FDA definition of data integrity, for example. Note that the data pedigree captures the metadata, or "data about data" (FDA 2016) concerning a particular data set, no matter how good or bad it is. That is, data pedigree is not a standard, but rather an objective depiction of the meaning, condition and history of the data. Data integrity, on the other hand, defines a standard for acceptance by the FDA.

The first element of the data pedigree is an explanation of what the data represent, i.e., the underlying subject matter knowledge involved, including units of measurement. For example, if the data involve measurements of color using the CIELAB color space (https://en.wikipedia.org/wiki/Lab_color_space), it is naïve to think that someone could analyze this data, and properly interpret the results, without an understanding what L*, a*, and b* represent in this color space.

Even such mundane things as the units of measurement are important. Sadly, in 1999 NASA lost a $125 million Mars orbiter when two teams working together on the project utilized different units of measurement; one metric and the other English (http://edition.cnn.com/TECH/space/9909/30/mars.metric.02/).

The next element is a description of the process that produced the data. There is an old saying in statistics: "You can't understand the data unless you first understand the process that produced it." This not only refers to the measurement process, but also understanding of the process that produced the samples subsequently measured. Process understanding, perhaps obtained through a SIPOC (supplier, input, process, output, customer) diagram (Hoerl and Snee 2012), provides the context that enables one to draw actionable conclusions from data analyses.

For example, suppose you are handed data on the viscosity of an industrial chemical. You could certainly begin analyzing the numbers without knowing anything about the chemical process involved. However, if the specific chemical in question ages rapidly, it would be hard to draw actionable conclusions without knowing how old the samples were. Of course, without knowing

anything about the chemical process, you would not even know that this was a relevant question to ask.

Next, it is important to document how the samples that were measured were originally selected. In textbooks, it is easy to state that a sample was randomly selected. However, in practice random sampling is an ideal that is rarely accomplished. In evaluating an election poll, for example, were registered votes sampled, likely voters, or some other group? If likely voters, how was a "likely voter" defined? What timeframe was used? How were voters who refused to answer questions handled? We argue that the devil is indeed in the details! Without understanding the sampling approach, it is impossible to know how broadly the results of the analysis might be inferred.

Measurement system evaluation is a traditional strength of the quality profession and is critical in evaluating data pedigree. For example, having an operational definition of a given measurement, including units of measurement, provides a foundation. We have, unfortunately, seen numerous data sets presented with no explanation of the definition of the measurement, or how it was made. There is a reason that documentation of the measurement system is a key element of ISO 9000 quality standards. In addition to documenting the measurement system, it is important to know if, how, and when this measurement system has been formally evaluated or calibrated.

If we benchmark the improvement methodology Lean Six Sigma, we see that the Measure step is one of the five phases of a Lean Six Sigma project and typically involves formal evaluation of the measurement system (Antony et al. 2018). Similarly, one could argue that the core purpose of the accounting profession is to ensure accuracy and consistency of financial statements (measurements). "Tainted, unreliable, or compromised" measurement systems are, unfortunately, all too common.

Lastly, we feel it is important to document the chain of custody (history) of the data set, in terms of who has had access to the data and could have made modifications, including eliminating data points. Ideally, a "gold standard" of the original data set should be maintained, i.e., what the FDA calls an "original or true copy." To understand why, consider the case of economists Carmen Reinhart and Kenneth Rogoff. In 2010 they published research on a large data set including 44 countries and spanning over 200 years of history (Reinhart and Rogoff 2010). Their analysis demonstrated a negative growth rate for countries with a high debt to gross domestic product ratio, which had obvious implications for economic policy. However, another set of economists sought to replicate these results, but could not (Herndon et al. 2013).

Upon further investigation, Herndon et al. determined: "We…find that coding errors, selective exclusion of available data, and unconventional weighting…led to serious errors…Our finding is that when properly calculated, the average real GDP growth rate for countries carrying a public-debt-to-GDP ratio of 90 percent is actually 2.2 percent, not -0.1 percent as published in Reinhart and Rogoff." In other words, when the data issues were addressed, the conclusions were exactly opposite of those originally published, i.e., that high debt leads to increased growth, not decreased growth.

In an article ironically entitled "Why AI is Still Waiting for its Ethics Transplant," Rosenberg (2017) quotes Kate Crawford on the criticality of understanding data pedigree in the context of artificial intelligence (AI):

> Data will always bear the marks of its history. That is human history, held in those data sets. So if we're going to try to use that to train a system, to make recommendations or to make autonomous decisions, we need to be deeply aware of how that history has worked.

Knowing the data pedigree provides additional important information: insight regarding how to analyze the data set. The core elements of a data pedigree, summarized in Table 3.4, help identify the sources of variation in a data set. Recall that understanding variation is a core element of statistical thinking (Hoerl and Snee 2012). The sources of variation define the appropriate models that could be used to analyze the data. Only after you know the potential sources of variation in a data set can you effectively create an appropriate model.

Statistical tools such as analysis of variance, regression, and multivariate analysis, are all based on knowing the potential sources of variation. Some sources of variation may not be of interest, and therefore are nuisance variables, but ignoring them in modeling is likely to produce bad results. Knowledge of the data pedigree also helps determine whether the data set is adequate for answering the question being asked in the first place. In some cases, unfortunately, it will not be, and time can be saved by not developing a model that will ultimately be inadequate for the questions of interest.

3.3.6 Utilizing a Data Pedigree in Practice

It may not be obvious what one does with a data pedigree. Recall that the purpose of documenting it in the first place is to understand the limitations of the data. Therefore, it should be used to determine if the current data are appropriate to achieve the objectives of the study, what types of analyses would be reasonable, and how broadly the results of these analyses might be inferred or generalized. A great deal of time could be saved if researchers understood from the very beginning that the data they have will not enable them to achieve their objectives. This could lead to acquisition of more appropriate data, perhaps through a designed experiment. As a simplistic example of a pedigree guiding analyses, if the data were collected over time, then some evaluation of the temporal behavior of the process would be reasonable. However, if the data were collected at one point in time, temporal analysis would make no sense.

We will use the corporate default case study from Hoerl and Snee (2017) to illustrate what a data pedigree looks like, and how it is utilized. In this case, a team from GE Global Research and GE Capital worked together to develop a model to predict defaults of corporations that were in GE Capital's financial portfolio. Predicting corporate defaults is obviously, a large, complex problem! It is also unstructured, in the sense that there is no commonly accepted definition of the word "default" in finance.

The team acquired data to work on this problem, as GE Capital did not possess data appropriate to addressing the problem. The model utilized probabilities of default (PD) as well as a slope or

momentum metric to map corporations to a risk space indicated as buy, hold, or sell. An abbreviated version of the pedigree of this data is given in Table 3.5. See Hoerl and Snee (2017) for further details on this statistical engineering case study.

Table 3.5 Abbreviated Data Pedigree from Default Prediction Case

Data Representation: While the model used Probabilities of Default (PD), these were calculated statistics. The original data consisted of market capitalization (stock prices times number of shares) and both long- and short-term debt metrics, as of the closing of each month, in US dollars. PD's were then calculated using the Black-Scholes-Merton methodology (Merton 1996) and a proprietary quantitative definition of "default," which was based on impaired payment of debt, using the metrics noted above. In laymen's terms, market capitalization estimates the overall economic value of a corporation's tangible (e.g., equipment) and intangible (e.g., brand) assets.

Process Description: The process of interest was the North American economic system, particularly as it relates to equity (stock) and debt (loan) markets for public corporations. This is obviously too large and complex of a system to document in detail here, but it was critical to study and understand this system well prior to analyzing the data.

Sampling: The original data set consisted of 1,986 Public, North American, Non-Ninancial (PNANF) corporations. The timeframe of the sampling was from January 1997 until December 2001 (5 years). The 1,986 sampled corporations were a *judgement sampl*e (i.e., not random) from the population of all possible PNANF corporations, selected to resemble the GE Capital portfolio. GE Capital representatives negotiated with the vendor, who we will refer to as "Smith and Company" to protect confidentiality, to identify this sample. Two actual GE Capital portfolios, one containing 1,106 PNANF corporations, and the other 553, were used for validation of the final model. These portfolios were also selected via judgement by GE Capital.

Measurement: The specific measurement processes belonged to Smith and Company and the individual companies that published their own financial metrics. In theory, each company utilized generally accepted accounting principles (GAAP) to calculate (measure) these. Publicly available market websites and publications (New York Stock Exchange, NASDAQ, etc.) were used to obtain monthly equity prices.

Measurement Evaluation: The team found it impossible to verify the data accuracy in an absolute sense. As noted below, no true gold standard existed. However, several comparisons were made of the data purchased from Smith and Company with GE Capital data, and no discrepancies were found. Smith and Company did have a data quality system in place, including auditing of the

numbers by an external accounting firm. However, there was no reasonable means of validating this data quality system.

Chain of Custody: The data had been maintained on internal (not cloud based) servers by Smith and Company's IT organization. To the best of anyone's knowledge, no modifications had been made, other than the correction of some outliers, i.e., invalid data points that had been identified through the data quality system. However, considerable turnover had occurred in the IT department, and it was impossible to verify the data relative to a true gold standard.

How was this pedigree utilized in practice? First, in this case the data were selected based on the intended analysis. For example, the specific corporations were sampled via judgment to reflect the GE Capital portfolio. Therefore, it was clear that this data would enable the team to attack the problem. Often, the data are given, and then the analyses determined. So, in this case, the data pedigree did not suggest obtaining additional data, or performing specific analyses, but in many cases it will. Here, the primary utilization of the pedigree was to ensure that the resulting model was applied appropriately.

For example, Table 3.5 shows that only Public, North American, Non-Financial (PNANF) corporations were included. This was a conscious choice based on the team's structuring of the problem (Phase 2 of the statistical engineering framework). Obviously, it would not be appropriate to apply the model to financial organizations, such as banks, or to private or European organizations. If the pedigree were not carefully documented, it would be easy to make such a mistake.

Further, the timeframe of the data collection was important (1997-2001), in that those with subject matter knowledge could understand how these years might have been unique in the financial markets, relative to other time frames. This helped them understand when it would and would ot be appropriate to utilize this model going forward.

Interestingly, financial data are often assumed to be 100% accurate, with little effort made to evaluate the measurement system. However, financial statements, such as commercial invoices, credit card or social security statements, and inventory lists, frequently contain errors. Third-party auditing generally does not validate the numbers *per se*, but rather states that the numbers have been arrived at using generally accepted accounting principles (GAAP). This is an important distinction when considering measurement accuracy in financial applications.

3.3.7 Summary

The scientific community, particularly the quality profession, has long understood the importance of measurement system evaluation. However, this is only one aspect of a data pedigree. Similarly, we are witnessing a growing body of failures of analytics in technical literature. Some of these blunders are humorous, but others are deadly. Data quality has been identified as one cause of these problems, and rightly so. The FDA, among many other federal agencies, and the legal profession, have long recognized the critical importance of this issue.

We argue that to properly evaluate data quality for a specific analysis, or determine the data integrity, we first need to document the pedigree of the data. If professional journals and agencies offering research grants were insistent on formal documentation of data pedigree, we feel that analytical studies would be much more reproducible, and that blunders would be significantly reduced. Similarly, those teaching analytics should emphasize the importance of documentation of the data pedigree before conducting formal data analyses, and for utilizing these pedigrees during analysis, and when generalizing (drawing inference from) the results of the analysis.

3.3.8 Standards of Practice

The key standard of practice discussed in this section is the importance of formally documenting the data pedigree, and then referring back to it during statistical analyses, and when determining and communicating the path forward from such analyses.

Section 3.3 References

Antony, J., Hoerl, R.W., and Snee, R.D. "Lean Six Sigma: Yesterday, Today, and Tomorrow," The International Journal of Quality & Reliability Management, 34,7, (2017): 1073-1093.

Food and Drug Administration Data Integrity and Compliance With CGMP: Guidance for Industry, (2016): https://www.fda.gov/downloads/drugs/guidances/ucm495891.pdf.

Dalal, D.R., Fowlkes, E.B., and Hoadley, B. "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure," Journal of the American Statistical Association, 84, (1989): 945-957.

Doganaksoy, N., Hahn, G.J., and Meeker, W.Q. "Assuring Product Reliability and Safety," Statistics, A Guide to the Unknown, 4th edition, Duxbury Press, Pacific Grove, CA, 2006.

Herndon T., Ash M., and Pollin R. "Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff," Working Paper Series 322, Political Commentary Research Institute, 2013.

Hoerl, R.W. and Snee, R.D. Statistical Thinking: Improving Business Performance, 2nd edition, John Wiley & Sons, Hoboken, NJ, 2012.

Hoerl, R.W. and Snee, R.D. "Statistical Engineering: An Idea Whose Time Has Come?", The American Statistician, 71, 3, (2017): 209-219.

Hoerl, R.W. and Snee, R.D. "Show Me the Pedigree," Quality Progress, January 16-23, 2019.

Kenett, R.S. and Shmueli, G. "On Information Quality," Journal of the Royal Statistical Society, Series A, 177, (2014): 3-38.

Kenett, R.S. and Shmueli, G. Information Quality: The Potential of Data and Analytics to Generate Knowledge, John Wiley and Sons, Chichester, UK, 2017.

Kolata, G. "How bright promise in cancer testing fell apart", The New York Times, July 7, 2012. (2014): https://www.nytimes.com/2011/07/08/health/research/08genes.html.

Merton, R.C. Continuous-Time Finance (revised ed.), Blackwell, Malden, MA, 1996.

Ince D. “The problem of reproducibility,” Chance, 25, (2012): 4–7.

Reinhart, C.M. and Rogoff, K.S. “Growth in time of debt,” American Economic Review, 100, 2, (2010): pp. 573–578.

Rosenberg, S. “Why AI is Still Waiting for its Ethics Transplant,” (2017): https://www.wired.com/story/why-ai-is-still-waiting-for-its-ethics-transplant/.

Vosoughi, S., Roy, D., and Aral, S. “The Spread of True and False News Online,” Science, 359, 6380, (2018): pp. 1146-1151.

Section 3.4 – Measurement Systems Analysis

3.4.1 Measurement Systems Analysis – Introduction

Measurement Systems Analysis (MSA) refers to the study of the properties of measurement systems. The "system" may be as simple as a single device or tool used for either variable (quantitative) or attributes (qualitative) inspection or may be a far more complicated "system" encompassing multiple gaging, people, software, hardware, environmental factors, facilities and other variables deemed important for the system to work properly. In most cases of manufacturing process control, the main interest is in people (appraiser variation or AV) and the gage (equipment variation or EV) used for the measurement. Together these two components of MSA constitute the classical "gage R&R" analysis found in many quarters. In addition, other MSA properties such as resolution, stability, bias, linearity, consistency and stability are important.

Essentially, anywhere a measurement is taken and for whatever purpose, we can always ask what the error in the measurement produced is. That is, if y is the measurement of a variable quantity, and if there exists in any sense some true value, x, for the phenomena measured, then the random variable y may be cast in simple form as $y = x + \varepsilon$ where ε is the random error term that can also have several components. One goal of the MSA study is to characterize the error distribution of ε. What is its mean and variance, is it stable, does the variance shift around as a function of the object measured, etc.

It is useful to have several working definitions associated with the properties of a measurement system at the start of any MSA study. This makes the execution of the study as well as the final communication of results clear to the user of these types of results.

3.4.2 Defined Terms of Measurement Systems

Table 3.6 Terms, Definitions and Comments

Accepted reference value, a value that serves as an agreed-upon reference for comparison, and which is derived as either: (1) a theoretical or established value, based on scientific principles, (2) an assigned or certified value, based on experimental work of some national or international organization, or (3) a consensus or certified value, based on collaborative experimental work under the auspices of a scientific or engineering group.

Accuracy, the closeness of agreement between a test result or measurement result and the true value. NOTE 1: In practice, the accepted reference value is substituted for the true value. NOTE 2: The term "accuracy", when applied to a set of test or measurement results, involves how close the values are to their true value.

Appraiser, the person who uses a gage or measurement system.

Bias, the difference between the expectation of measurement results and either the true value (if known) or an accepted reference value.

Calibration, the process of establishing a relationship between a measurement device and a known standard value(s).

Confusion matrix, in an attribute MSA study, in its simplest form, a 2x2 matrix that shows the number of correct and incorrect binary classifications of n objects being classified by a single appraiser or two appraisers against each other. The four cells within the matrix give: a) the correct classifications of objects having the attribute; b) the incorrect classifications of objects having the attribute; c) the correct classification of objects not having the attribute; and d) the incorrect classification of objects not having the attribute.

Discrimination ratio, a statistical ratio calculated from the statistics from a gage R&R study that measures the number of 97% confidence intervals, constructed from the gage R&R variation that fits within six standard deviations of true values.

Distinct product categories, an alternate term for the discrimination ratio.

Gage, a device used as part of the measurement process to obtain a measurement result.

Gage consistency, the constancy of repeatability variance over a period of time. Consistency means that the variation within measurements of the same object (or group of objects) under the same conditions by the same appraiser behaves in a state of statistical control as judged, for example, using a control chart.

Gage performance curve, a curve that shows the probability of gage acceptance of an object given its real value or the probability that an object's real measure meets a requirement given the measurement of the object.

Gage R&R, the combined effect of gage repeatability and reproducibility.

Gage resolution, the ability of the gage to detect changes in the characteristic being measured and discriminate between measurement values.

Gage stability, the absence of a change, drift, or erratic behavior in bias over a period of time, Stability means that repeated measurements of the same object (or average of a set of objects) under the same conditions by the same appraiser behave in a state of statistical control as judged for example by using a control chart technique.

Linearity, An assessment of accuracy through the defined range of expected measurements in any inspection system.

Measurement process, process used to assign a value to a property of an object or other physical entity.

Measurement result, a value assigned to a property of an object or other physical entity being measured.

Measurement system, the collection of hardware, software, procedures and methods, human effort, environmental conditions, associated devices, and the objects that are measured for the purpose of producing a measurement.

Measurement Systems Analysis (MSA), any number of specialized methods useful for studying a measurement system and its properties.

Precision, the closeness of agreement between independent test/measurement results obtained under stipulated conditions. NOTE 1: Precision depends only on the distribution of random errors and does not relate to the true value or the specified value. NOTE 2: The measure of precision is usually expressed in terms of imprecision and computed as a standard deviation of the test results or measurement results. Less precision is reflected by a larger standard deviation. NOTE 3: Quantitative measures of precision depend critically on the stipulated conditions. Repeatability conditions and reproducibility conditions are particular sets of extreme stipulated conditions.

Repeatability, the variation resulting when a single object is measured multiple times, independently, under stable conditions, by a single appraiser using the same equipment. The phrase *repeatability conditions* is used to describe these conditions. In a basic gage R&R study repeatability or repeatability conditions is also referred to as Equipment Variation or EV.

Reproducibility, the variation resulting among the average values determined independently by several appraisers, where each appraiser measures the same group of objects using the same equipment and under stable conditions. The phrase *reproducibility conditions* is used to describe these conditions. In a basic gage R&R study reproducibility or reproducibility conditions is also referred to as Appraiser Variation or AV.

3.4.3 Simple Attribute MSA

If y is a binary attribute indicator variable, say $y=1$ if a condition or attribute is present in an object inspected and 0 if not, then there can still be the error of declaring the attribute present when it is not, or declaring the attribute not present when it is. There are four possible classification cases in any binary type of measurement:

- *True positives* (TP) – the object has the condition and is classified as having the condition
- *True negatives* (TN) – the object does not have the condition and is classified as not having the condition.
- *False Positives* (FP) – the object does not have the condition and is classified as having the condition
- *False negatives* (FN) – the object has the condition and is classified as not having the condition.

A simple attribute MSA study with a single appraiser would start with a set of *n* objects of which *r* truly have the condition and *n-r* do not. The appraiser then proceeds to classify the *n* objects – usually in some randomized order. In the total of *n* exactly *s* is classified as having the attribute and *n-s* as not having the condition. The resulting data can then be arranged in a 2 by 2 contingency table. Figure 3.4a below illustrates.

Figure 3.4a – Simple Attribute Inspection Template

		True State, *X*		
		0	1	
Measured, *Y*	0	TN	FN	*n-s*
	1	FP	TP	*s*
		n-r	*r*	*n*

In this figure, the TN, FN, FP and TP represent the count of these conditions – as defined above. The table is often called a *confusion matrix*. Once the table is completed, several estimates of inspection error may be calculated and are typically cast as probabilities. Dividing every number in the table in Figure 3.4b by *n* converts the counts to probability estimates. To do the appropriate calculations it is also useful to denote *x* as the binary indicator variable for units that truly have the attribute and *y* as the binary indicator variable for units that are classified as having the attribute. Using this nomenclature, all of the appropriate metrics may be calculated. For example, the misclassification rate of classifying a true positive as a negative is cast as the conditional probability *P(y=0|x=1)*, called the False Negative Rate (FNR). Using the nomenclature in Figure 3.4a, we can express the FNR as

$$FNR = \frac{FN}{FN + TP}$$

The following example illustrates these ideas.

A certain type of attribute inspection process for the presence of foreign material in an aircraft component uses a probe for the inspection process. The inspection area is concealed by other aircraft components, and the probe can move around during the inspection making detection

ambiguous. To determine the properties of the inspection process, a simple experiment was set up in a simulated laboratory setting using n=150 components where 70 of the components were known positives for the attribute and 80 of the components were known negatives. A single inspector was used. The following raw count data were obtained.

Figure 3.4b Raw Inspection Count Data

		True State, X		
		0	1	
Measured, Y	0	73	13	86
	1	7	57	64
		80	70	150

The data in Figure 3.4b are turned into probabilities by dividing every number in the table by n=150. Figure 3.4c shows the results.

Figure 3.4c Probability Matrix

		True State, X		
		0	1	
Measured, Y	0	0.487	0.087	0.573
	1	0.047	0.380	0.427
		0.533	0.467	1.000

The four cells in the center of the matrix are estimates of the four cases of the joint probabilities *P(x,y)* where *x* and *y* each take on the binary classification values of 0 and 1. The far-right column and bottom row are estimates of the marginal probabilities of *P(y) and P(x)*. The overall relative agreement or *accuracy rate* is estimated as (TN + TP)/*n*. In this example, the accuracy rate is 0.487+0.380 = 0.867 or 86.7%. The overall *miss-classification rate* is based on the opposite diagonal cells (FN + FP)/*n* = 20/150 = 0.133 or 13.3%. Several other rate calculations are of interest. The *False Positive Rate* (FPR) is the rate of occurrence indicating a condition exists when it does not. This is a conditional probability calculation:

$$FPR = \frac{FP}{FP + TN} = \frac{0.047}{0.533} = 0.0875$$

The *False Negative Rate* (FNR) is calculated in a similar way:

$$FNR = \frac{FN}{FN + TP} = \frac{0.087}{0.467} = 0.186$$

In the language of classical statistics, the FPR is the type I error rate (α) and the FNR is the type II error rate (β). "Power" or *sensitivity* is the quantity 1-β and represents the probability of detecting a meaningful difference that exists[1]. In a quality control sense, it is the probability that you reject a lot that is "bad". In the language of quality control, FPR and FNR are referred to as producer's and consumer's risks respectively. Thus, in the example, α=0.085 and β=0.186 making the power or sensitivity 1-β=0.814.

It is of additional interest to the consumer to ask for the true condition of any unit, given its inspection result where these are different. Thus, we have the rates called *False Omission Rate* (FOR) and *False Discovery Rate* (FDR) expressed as:

$$FOR = \frac{FN}{FN + TN} = \frac{0.087}{0.573} = 0.151$$

$$FDR = \frac{FP}{FP + TP} = \frac{0.087}{0.427} = 0.110$$

The power, is, in some applications, referred to as the detection probability or POD (*Probability Of Detection*). So, POD = 1-□ = 1 – FNR. In the example, POD = 0.814. In some applications, we can increase POD by doing redundant inspections, assuming the several inspections give independent results. In this example, suppose that three independent inspections are made. Since each has a POD of about p=0.814, and we want at least one of the three inspections to capture the condition, the overall POD with three inspections (POD_3) is increased to:

$$POD_3 = 1 - (1 - p)^3 = 1 - (1 - 0.814)^3 = 0.994$$

Doing the three redundant inspections will also increase the FPR and decrease the FNR. In this case the FPR increases to about $0.24 = 1 - (1-0.088)^3$. That is a tradeoff that is sometimes made, particularly where safety is of concern. In some applications, we might be given a POD and the FPR. It may also be known from empirical experience that the true proportion of objects having the undesirable condition in the overall population is q. What can we say about the probability the condition exists in an object given the test says it has the condition? We analyze this scenario using a simple case of Bayes Theorem. Define the following quantities: POD = $P(y=1|x=1)$, FPR = $P(y=1|x=0)$ and $q=P(x=1)$ [the marginal probability of x being equal to 1 calculated as s/n in Figure 3.4a]. We want to calculate $P(x=1|y=1)$.

$$P\left(x=1\middle|y=1\right) = \frac{P\left(x=1, y=1\right)}{P\left(y=1\right)} = \frac{P\left(y=1\middle|x=1\right)P\left(x=1\right)}{P\left(y=1\middle|x=1\right)P\left(x=1\right) + P\left(y=1\middle|x=0\right)P\left(x=0\right)} \qquad (3.1)$$

[1] Also known as the True Positive Rate or TPR, as well as recall or hit rate.

Substituting for POD, FPR and q, the following simplified formula results.

$$P\left(x=1|y=1\right)=\frac{\left(POD\right)q}{\left(POD\right)q+\left(FPR\right)\left(1-q\right)} \tag{3.2}$$

Suppose POD=0.99 and FPR=0.05; further suppose that q=0.002. That is, the detection rate is reasonable high at 99% but the actual proportion of objects that truly have the condition in the at large population is relatively small, here 0.002. Using these numbers, we find:

$$P(x=1|y=1)=\frac{(0.99)(0.002)}{(0.99)0.002+(0.05)(1-0.002)}=0.038$$

Then only about 3.8% of units classified as having the condition actually have the condition despite the fact that the test accuracy may be shown to be about 95% and the POD is 99%. To improve on P(x=1|y=1) we need a much lower FPR.

In using this simple case, we can also give a statistic called Cohen's Kappa, κ, which measures a degree of goodness in the overall inspection process. Generally, the kappa statistic is used where there are two appraisers, and the objective is to compare the two. We can also use it in principle in this simple case to compare the true quantities with the measured quantities. The kappa formula is:

$$\kappa=\frac{p_o-p_e}{1-p_e}=1-\frac{1-p_o}{1-p_e} \tag{3.3}$$

In the formulation for κ, p_0 is identical with the test accuracy and p_e is measuring the degree of agreement due to chance alone. Formally the calculations are done as p_0 = P(x=1,y=1) + P(x=0,y=0) = (TN + TP)/n = 0.867, and p_e =P(x=1)P(y=0)+P(x=0)P(y=1) = 0.495. When p_0=1, there is perfect accuracy and κ=1. When $p_0 = p_e$ there is no agreement other than what we would expect by pure chance. In that case κ=0. It is sometimes possible for κ to be negative. In that case $p_0 < p_e$ and there is worse than chance agreement. In the example, kappa is calculated as:

$$\kappa=\frac{0.867-0.495}{1-0.505}=0.745$$

AS13003 (2015) gives some guidance on interpreting point estimates of κ and also discuss the statistical properties of this statistic. The authors suggest that for the range $0.61 \leq \kappa \leq 0.8$ there is substantial agreement; however, cases do vary and are somewhat unique. Sound context related technical judgment should be exercised in every case.

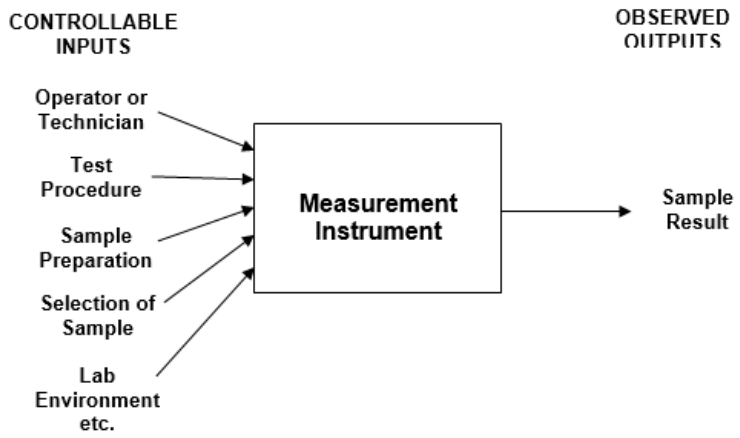
3.4.4 Simple Variable Measurement MSA

3.4.4.1 Basic Concepts

Data are too often simply taken at face value and nothing of their origin is considered. In other words, data taken from a manufacturing process are more often than not used to control it with no regard given to whether or not the measurement process that generated the data was in control. Shewhart (1931) once said that, "In any program of control we must start with observed data; yet data may either be good, bad, or indifferent. Of what value is the theory of control if the observed data going into that theory are bad? This is the question raised again and again by the practical man."

Consider the simple model of a process as seen in Figure 3.5. Inputs to the measurement process are varied but assumed controllable.

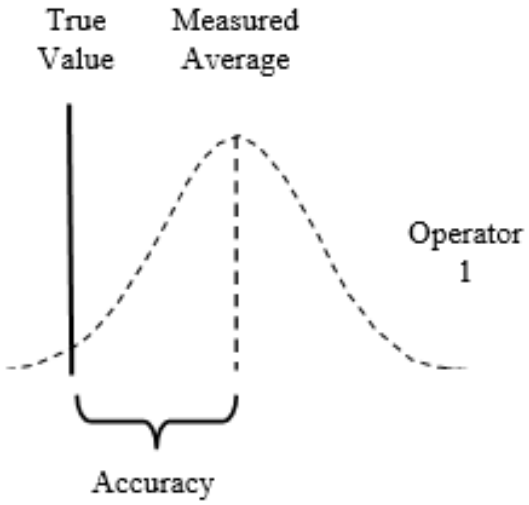
Figure 3.5 Measurement data are a result of a process involving several inputs, most of them controllable.



Each sample result is the direct output of the manner in which it was created by the combination of the person making the measurement, the procedure used, the quality of the sample preparation, where the sample came from, the temperature and humidity in the lab at that time, etc. Of course, if the measurement instrument is designed to be robust to operator and environment effects then the sample result will be less affected. However, it is unlikely that the instrument can avoid use of an improper procedure, poor sample preparation, and a sample that is deficient in some respect.

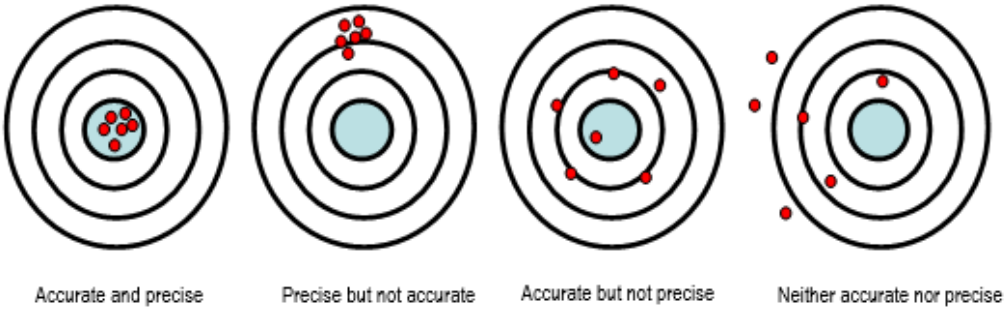
If the sample measured were a "standard sample", or control sample, with a specified *reference target* value such as a NIST-traceable standard, then we can evaluate each result against the target to determine whether the measurement was correct or not. A standard sample that reads on average the same as the target value, or *true value*, means that the measurement process is accurate, and the average is considered the true average if the measurements were made with a precision calibrated instrument. If the standard sample fails to agree with the target value on average, the measurement process is inaccurate, or not accurate, and calibration is necessary. Accuracy is often called the bias in the measurement and is illustrated in Figure 3.6.

Figure 3.6 Gage accuracy is the difference between the measured average of the gage and the true value, which is defined with the most accurate measurement equipment available.



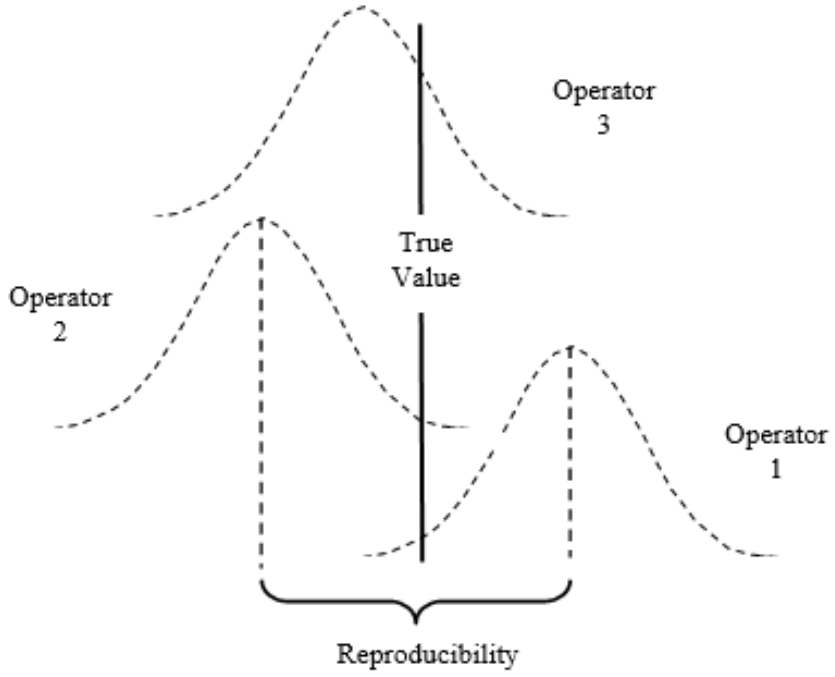
The variability of the sample measurements is also considered. When the variability of the sample data is small, the measurement is said to have *precision*. If the sample variation is large, i.e., scattered, then the measurement process is *imprecise*, or not precise. Figure 3.7 shows the four scenarios that relate the combinations of data that are either accurate, precise, neither or both.

Figure 3.7 Measurement data can be represented by one of four possible scenarios.



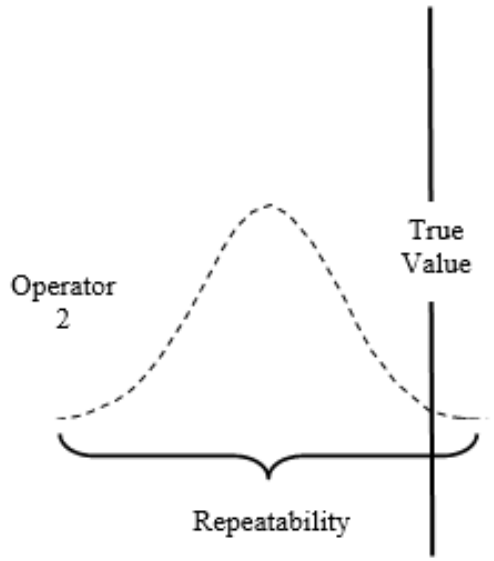
Gages, as measurement equipment, are subject to variation. They must be accurate and precise. If a gage is not properly calibrated for accuracy, a bias may be present. We could experience the same result if different people use the calibrated gage and get different results. This is referred to as the *reproducibility* of the gage and is illustrated in Figure 3.8.

Figure 3.8 Gage reproducibility can be represented as the variation in the average of measurements made by multiple operators using the same gage and measuring the same parts.



On the other hand, if a single person uses the gages and takes repeat readings of a single sample there will be variation in the results. This is referred to as the *repeatability* of the gage and is illustrated in Figure 3.9.

Figure 3.9 Gage repeatability can be represented as the variation in the measurements made by a single operator using the same gage and measuring the same parts.



3.4.4.2 What Can Affect the Measurement Process?

A measurement process contains any or all of the following:

- Machine(s) or device(s)
- Operator(s) or appraiser(s)
- Sample preparation
- Environmental factors

Multiple machines or devices are often used since it is unrealistic to expect that all process measurements can be done by a single machine or device. For this reason, it is important to assess whether the use of multiple measurement machines or devices is contributing to the error of the measurement system. Likewise, it is typical that more than one operator or appraiser will be needed to make the measurements. Since not everyone has the same attention to detail, it is not uncommon for there to be a potential contribution to measurement error due to differences in results among operators or appraisers—measuring the same sample.

A frequent omission in measurement studies is the consideration of any sample preparation that could affect the measurement result. Samples that are prepared in a lab can see their measurements affected due to improper polishing, insufficient material, poor environmental conditions, incorrect chemical solutions, and many other reasons. Often these problems can be resolved through adequate training of laboratory personnel.

Bishop, Hill and Lindsay (1987) offer some useful questions to ask when investigating a measurement system:

- Does the technician know that the test is very subjective?
- Are technicians influenced by knowledge of the specification and/or control limits of the process attribute?
- If more than one technician and/or instrument is used to collect the measurements, are there any statistically significant differences among them? We do not want to make changes to the process when the data are really representing measurement differences!
- Is the measurement instrument the source of the problem? Perhaps it is in need of calibration, or its settings are not correct, so an adjustment is needed.

These authors present three examples of how problems associated with the measurement system can mislead the engineer who is investigating the production process. In each example, a measurement problem would result in an unnecessary, or a lack of a needed, process adjustment.

- **When things are worse than they appear.** This situation can occur when the test method is very subjective, and there are multiple technicians doing the measurements. Statistical differences will no doubt be present among the technicians based on measurements of the same samples. If the technician knows the target value for the product, as well as the specification limits for the response, the data may become tainted with readings closer to the target than they really are. Thus, the measurement data will make the process look better controlled than it actually is. The solution is a

combination of a new, less subjective test method, new control limits based on the test method, and further training of the technicians.

- **When things are better than they appear.** This situation can occur when the analytical method being used is out of statistical control from time to time. In response to this, engineers may feel compelled to "do something" to bring the process back into control. Unfortunately, the engineers will often fail to discover any assignable cause for the apparent process change. Likewise, it may not be apparent what the assignable cause is for the test to be out of control. The solution is to institute the use of a "standard sample", or control sample, which is submitted along with the production samples for measurement. If the "standard sample" continues to read within its control limits, the test method is deemed to be correct and any out of control production measurements should be a cause for action. On the other hand, in the situation described here, the "standard sample" will often indicate that the test is out of control, and that process changes should not be made based on a faulty measurement. As a rule, no process control change should be made when the results of the production samples are correlated with the results of the control sample.

- **When the specimen tested is not representative of the product.** This situation can occur if a test specimen is taken from the wrong production line for which the data are being used for control purposes. It can also occur if the test method is not consistent with a proper recommended technique, such as that prescribed by an ASTM standard. In addition, this situation can occur if the test specimen was taken from a production lot other than the one intended. These scenarios are only examples of how a test specimen can fail to be representative of the product being evaluated. The reader can probably cite other examples based on their knowledge of other processes.

In each of the above examples, the authors used a type of nested design discussed in the next section. Typically, these designs are used for investigating measurement systems involving multiple technicians making multiple sample preparations and multiple measurements on each sample preparation. The technicians, preparations and repeated measurements are sources of variation that need to be quantified.

Such designs should be performed in conjunction with a process investigation. In this manner, you can judge how much of the variation seen in the data is attributed to the production process and how much to the measurement process. If the measurement process accounts for the larger portion of the total variation, then efforts should be directed towards this area as an opportunity for making the overall process more consistent.

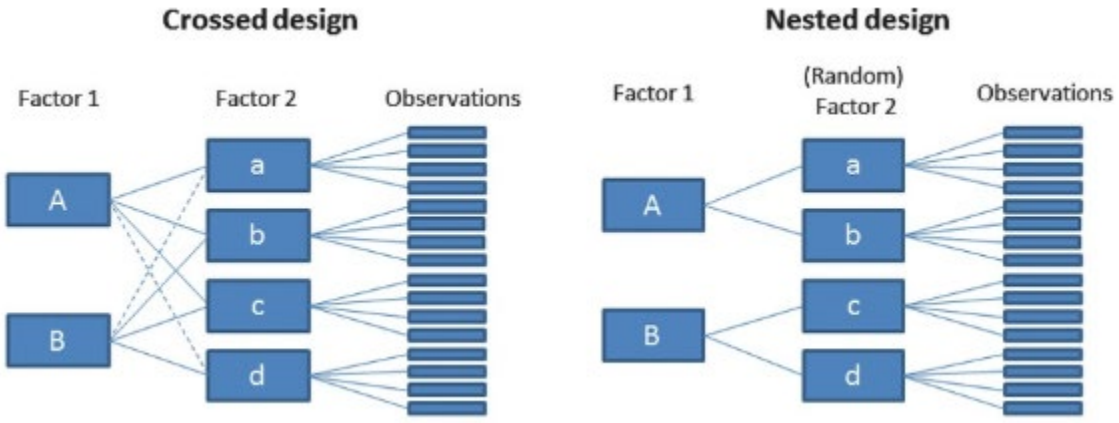
Samples should be submitted in a “blind” fashion to the technician so that person is not aware of what its reading should be, i.e., knowledge of its target value. These samples should be part of the typical workload and they should be tested in a random sequence (not the order in which they come from the production process).

3.4.4.3 Crossed vs. Nested Designs

A natural extension of the nested design occurs when the experimenter wishes to partition sources of variability due to differences in parts, operators, periods of time, etc. to provide some direction for identifying opportunities to reduce measurement variation. Nested designs of this nature are typically referred to as variance component designs. Variance component designs treat operators and parts as random effects such that their contribution to total variation is additive in nature.

Oftentimes, a crossed design may be more appropriate than a nested design. In the case of a crossed design, each sample, or part, is measured repeatedly by each operator on each day, etc. in such a way that the factors are crossed with each other. This can be seen in Figure 3.10. The operator-part interaction is often of the most interest in these designs. A significant interaction will indicate that the operators were not able to reproduce their results for all of the measured samples, or parts. In other words, one or more operators may have had difficulty measuring a part whereas the others did not. Crossed designs treat operators and parts as fixed effects such that we are looking for statistical differences among the levels of each factor and their interaction.

Figure 3.10 Crossed design structure compared to a nested design.



In the case of a nested design, each sample, or part, cannot be measured by another operator or on another day, etc. such that the factors become nested within the other factors.

Nested designs are necessary if the testing is destructive in some way, or if some or all of one or more factors are isolated from the others in a matter that makes a crossed design impractical to conduct, or it is not cost efficient to run.

For example, if plant location is a factor, it may not be cost effective to send each operator to the other plants just to collect data using a particular type of gage, but it may be possible to carry a set of samples between locations (as long as they do not become broken in transit). Of course, we would have to assume that the gages at each location agree which is a big and often unwarranted assumption. If the gages are regularly calibrated to NIST-traceable (or other organization-traceable) standards, then it may be safe to assume that the gages do not contribute

much to the reproducibility of the measurement between locations. Both designs can be used in assessing gage measurement capability.

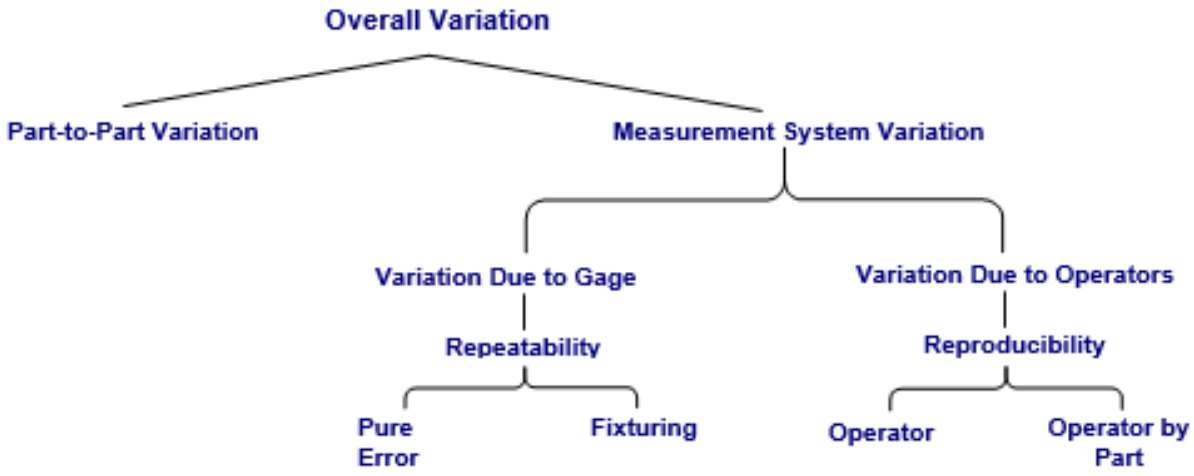
3.4.4.4 Gage R&R

In "gage R&R" one R stands for repeatability and the other R for reproducibility. Repeatability is measuring the variation in the gaging system or the closeness from one measurement to the next in measuring the same object by a single person repeatedly. Reproducibility is a kind of personal bias that offsets the average measurement by an appraiser by a fixed amount but differing and random for each appraiser. The people who participate in the study are considered a random sample from an otherwise infinite population of people who might use the gage. In theory, each such participant has a random reproducibility component that offsets the average measurement from the true value. The gage R&R study provides an estimate of the variance of the population of such offsets and judges if the effect we see (variance estimate) is significantly different from 0. The basic model incorporating both R&R is:

$$y_{ijk} = x_i + v_j + \varepsilon_{ijk} \tag{3.4}$$

In (3.4), x is the true value, v is the random reproducibility term and ε the random repeatability term. An interaction term is not being used in this model. y_{ijk} is the k^{th} repeat measurement of the i^{th} part, by the j^{th} operator. The terms are considered independent, so the variance of the measurements (y) is a sum of the three component variances – true value, reproducibility and repeatability variances. The last two constitute the gage R&R variance. More generally, the variance of the measurements represents the total process variation. "Process" includes true part variation and measurement system variation. The later includes gage R&R and possible bias effects. We can call these two components precision and accuracy, respectively. Figure 3.11 shows the breakdown.

Figure 3.11 Components of overall variation into part-to-part and measurement variation (and its components).



In another variation of simple repeatability, there are two or measurements of the same part by a single appraiser, using several parts. The following example uses two measurements per part.

3.4.4.5 Gage R&R Study (Long Method)

The American Society for Quality (ASQ) Automotive Division SPC Manual (1986) defines the long method as determining repeatability and reproducibility of a gage separately. Comparing these estimates can give insight into causes of gage error. If reproducibility is large compared to repeatability, then possible causes could be:

- Operator that is not properly trained to use and read the gage
- Calibration markings on the gage are not clear to the operator

If repeatability is large compared to reproducibility, then possible causes could be:

- Gage needs maintenance
- Gage should be redesigned to be more rigid
- Clamping of and location for gauging needs improvement

In preparation for running the study, establish the purpose of the study and determine the kind of information needed to satisfy its purpose. Answer these questions:

- How many operators will be involved?
- How many sample parts will be needed?
- What number of repeat readings will be needed for each part?

Next, collect the parts needed for the study. These parts should represent the range of possible values the gage is expected to see in practice. Finally, choose the operators needed to conduct the study. Again, you will want to choose people who represent the range of skill within the pool of inspectors available. Measurements should be taken in random order to reduce the possibility of any bias.

The study is typically conducted using the following steps involving multiple operators (use 2-3, preferably 3), multiple parts (use 5-10, preferably 10), and repeat number of trials (use 2-5, will depend on cost and time constraints).

1. Refer to operators as A, B, etc., and to parts as 1, 2, etc. (number parts so the markings are not visible to the operators).
2. Calibrate the gage to be evaluated.
3. Operator A measures the parts in random order and enters the data into the 1st column of the form shown in Table 3.7.
4. Repeat step 3 for the other operator(s) and appropriate columns.
5. Repeat steps 1 to 4, with the parts measured in another random order, as many times as the number of trials specified. After each trial, enter the data on the form for each part and operator.

6. Steps 3 to 5 can be modified for large size parts, when parts are unavailable, or when operators are on different shifts.
7. Using the data collection form in Table 3.7 and the calculations form in Table 3.8, compute the gage R&R statistics.

Repeatability, also referred to as Equipment Variation (EV), estimates the spread that encompasses 99% of the measurement variation due to the same operator measuring the same part with the same gage, and is calculated as

$$EV = 5.15\hat{\sigma}_{EV} = 5.15\hat{\sigma}_e = 5.15\left(\frac{\overline{\overline{R}}}{d_2^*}\right) = \overline{\overline{R}}\ x\ K_1$$

$$\hat{\sigma}_{EV} = \hat{\sigma}_e = \frac{EV}{5.15}$$

where $\overline{\overline{R}}$ is the average range of the operator ranges $\overline{R}_A$, $\overline{R}_B$, etc., and K_1[2] is a tabulated constant which is given in Table 3.7. The factor 5.15 represents the overall number of standard deviations ($\pm$2.575) about the mean within which 99% of the observations are expected to lie under the assumption of a normal distribution. Reproducibility, also referred to as *appraiser variation (AV)*, estimates the spread that encompasses 99% of the measurement variation due to different operators measuring the same part with the same gage, and is calculated as

$$AV = 5.15\hat{\sigma}_{AV} = 5.15\hat{\sigma}_o = 5.15\sqrt{\left(\overline{X}_{diff}\ x\ K_2\right)^2 - \left[\left(EV\right)^2 / \left(n\ x\ r\right)\right]}$$

$$\hat{\sigma}_{AV} = \hat{\sigma}_o = \frac{AV}{5.15}$$

---

[2] K_1 is $\frac{5.15}{d_2^*}$, where d_2^* is tabulated in Appendix Table 3.A.1 and is based on k = (# operators)x(# parts) and n = # trials. For example, if three operators are used with four parts for three trials, then k=(3)(4)=12 and n=3 which yields a value of d_2^*=1.713 from Appendix Table 3.A.2. Thus, K_1 will be

$$K_1 = \frac{5.15}{d_2^*} = \frac{5.15}{1.71} = 3.01$$

NOTE: The values in Table 3.A.1 for K_1 are based on a value of d_2^* for infinite degrees of freedom (last row in Appendix Table 3.A.2.

where $\overline{X}_{diff}$ is the range of the operator averages $\overline{X}_A$, $\overline{X}_B$, etc., K_2[3] is a tabulated constant which is given in Table 3.8 and based on a d_2^* factor for k=1 found in Appendix Table 3.A.2, n is the number of parts measured, and r is the number of trials.

Repeatability and reproducibility, also referred to as *gage R&R*, estimates the spread that encompasses 99% of the variation due to both sources and is calculated as follows and illustrated in Figure 3.9

$$R\,\&\,R = 5.15\hat{\sigma}_m = 5.15\sqrt{\left(\hat{\sigma}_{EV}\right)^2 + \left(\hat{\sigma}_{AV}\right)^2} = 5.15\sqrt{\hat{\sigma}_e^2 + \hat{\sigma}_o^2}$$

$$\hat{\sigma}_m = \frac{R\,\&\,R}{5.15}$$

[3] K_2 is $\frac{5.15}{d_2^*}$, where d_2^* is tabulated in Appendix Table 3.A.2 and is based on k = 1 and n = # operators. For example, if three operators are used, then k=1 and n=3 which yields a value of d_2^*=1.910 from Appendix Table 3.A.2. Thus, K_2 will be

$$K_2 = \frac{5.15}{d_2^*} = \frac{5.15}{1.91} = 2.70$$

Table 3.7 Gage Repeatability and Reproducibility Data Collection Sheet (long method).

Operator	A						B						C						
Part #	1st Trial	2nd Trial	3rd Trial	4th Trial	5th Trial	Range	1st Trial	2nd Trial	3rd Trial	4th Trial	5th Trial	Range	1st Trial	2nd Trial	3rd Trial	4th Trial	5th Trial	Range	Average
1																			
2																			
3																			
4																			
5																			
6																			
7																			
8																			
9																			
10																			
Average																			

$\bar{R}_A$ $\bar{R}_B$ $\bar{R}_C$

Sum $\bar{X}_A$ Sum $\bar{X}_B$ Sum $\bar{X}_C$

$\bar{R}_A$	
$\bar{R}_B$	
$\bar{R}_C$	
Sum	
$\bar{\bar{R}}$	

# Trials	D4
2	3.27
3	2.58
4	2.28
5	2.11

$\left(\bar{\bar{R}}\right) x \left(D_4\right) = UCL_R$ *

( ) x ( ) = _____

Max Oper $\bar{X}$	
Min Oper $\bar{X}$	
$\bar{X}_{diff}$	

Max Part $\bar{X}$	
Min Part $\bar{X}$	
Rp**	

* Limit of the individual ranges (Rs). Ranges beyond the limit are circled, and the cause should be identified and corrected. Repeat the readings with the same operator and same sample, or discard values and reaverage and recompute R and UCL_R from rest of data.

** Range of all the sample averages

Table 3.8 Gage Repeatability and Reproducibility Calculations Sheet (long method).

GAUGE REPEATABILITY & REPRODUCIBILITY REPORT

Gauge Name: _____ Date: / / Characteristic: _____

Gauge No.: Study Done By: Part No. & Name:

Gauge Type: _____ Specification: _____ Spec. Tolerance (if 2-sided): _____ **(TOL)**

From Data Collection Sheet: $\bar{\bar{R}}$ = [ ] $\bar{X}$ Diff. = [ ] R_p = [ ]

Parts = [ ] # Trials = [ ]

MEASUREMENT ANALYSIS

Repeatability (a.k.a. Equipment Variation, EV)

$EV = (\bar{\bar{R}}) x (K_1)$

= ( ) x ( )

= [ ]

Reproducibility (a.k.a. Appraiser Variation, AV*)

$AV = \sqrt{[(\bar{X}_{diff}) x (K_2)]^2 - [(EV)^2 / (n x r)]}$

= $\sqrt{[(\quad) x (\quad)]^2 - [(\quad)^2 / (\quad)]}$

= [ ]

Repeatability and Reproducibility (R & R)

$R \& R = \sqrt{(EV)^2 + (AV)^2}$

= $\sqrt{(\quad)^2 + (\quad)^2}$

= [ ]

Part Variation (PV)

$PV = (R_p) x (K_3)$ = ( ) x ( ) = [ ]

Total Variation (TV)

$TV = \sqrt{(R \& R)^2 x (PV)^2}$

= $\sqrt{(\quad)^2 x (\quad)^2}$ = [ ]

# Trials	K_1	# Operators	K_2
2	4.57	2	3.65
3	3.04	3	2.70
4	2.50	4	2.30
5	2.21	5	2.08

# Parts	K_3
2	3.65
3	2.70
4	2.30
5	2.08
6	1.93
7	1.82
8	1.74
9	1.67
10	1.62

* If AV is a negative value within the square root sign, the appraiser variation will default to zero (0).

% TOLERANCE ANALYSIS

% EV = 100 [(EV)/(TOL)]

= 100[( ) / ( )]

= [ ] %

% AV = 100 [(AV)/(TOL)]

= 100[( ) / ( )]

= [ ] %

% R&R = 100 [(R&R)/(TOL)]

= 100[( ) / ( )]

= [ ] %

% TOTAL VARIATION ANALYSIS

% EV = 100 [(EV)/(TV)]

= 100[( ) / ( )]

= [ ] %

% AV = 100 [(AV)/(TV)]

= 100[( ) / ( )]

= [ ] %

% R&R = 100 [(R&R)/(TV)]

= 100[( ) / ( )]

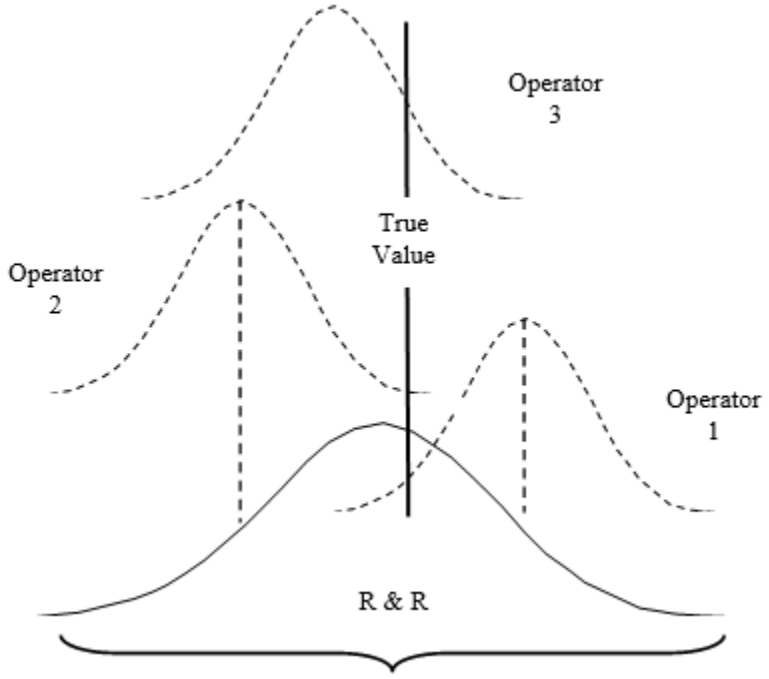
= [ ] %

% PV = 100 [(PV)/(TV)]

= 100[( ) / ( )]

= [ ] %

Figure 3.12 Gage R&R can be represented as the total variation due to measurements made by multiple operators using the same gage and measuring the same parts.



Part-to-part variation, also referred to as *PV*, estimates the spread that encompasses 99% of the measurements from a normal distribution and is calculated as

$$PV = 5.15\hat{\sigma}_p = 5.15\left(\frac{R_p}{d_2^*}\right) = R_p \; x \; K_3$$

$$\hat{\sigma}_p = \frac{PV}{5.15}$$

where R_p is the range of the part averages, and K34 is a tabulated constant which is given in Table 3.A.1 and based on a d_2^* factor for k=1 found in Appendix Table 3.A.2.
The total process variation, also referred to as TV, is calculated from the measurement study as

[4] K_3 is $\frac{5.15}{d_2^*}$, where d_2^* is tabulated in Appendix Table 3.A.2 and is based on *k* = 1 and *n* = # parts. For example, if four parts are used, then *k*=1 and *n*=4 which yields a value of d_2^*=2.239 from Appendix Table 3.A.2. Thus, K_3 will be

$$K_3 = \frac{5.15}{d_2^*} = \frac{5.15}{2.24} = 2.30$$

$$TV = 5.15\hat{\sigma}_t = 5.15\sqrt{\hat{\sigma}_p^2 + \hat{\sigma}_m^2}$$

$$\hat{\sigma}_t = \frac{TV}{5.15}$$

The *Number of Distinct Categories*, also referred to as *NDC*, that can be obtained (estimated) from the data and is calculated as

$$NDC = \frac{1.41\hat{\sigma}_p}{\hat{\sigma}_m}$$

For more information on this metric, see Wheeler and Lyday (1989) and the AIAG *Measurement Systems Analysis Reference Manua*l (1990).

Some guidelines in the interpretation of NDC are:

- If NDC =1, the measurement system cannot be used to control the process since the gage cannot tell one part from another, i.e., the data are 100% noise
- If NDC = 2, the data fall into two groups, like attribute data
- If NDC = 3, the variable data are considered to be of a low-grade quality which will produce insensitive control charts
- If NDC = 4, the variable data are improved
- If NDC = 5, the variable data are even better (minimum acceptability)
- The NDC should be > 5, and the larger the better, in order for the measurement system to be deemed truly acceptable

The *Discrimination Ratio*, also referred to as *DR*, estimates the degree to which the observed variation is beyond that characterized by the control limits of an $\overline{X}$ chart of the data (discussed in the next section). Note that control limits are based on short term variation, i.e., repeatability in a measurement sense, and the observed variation contains this variation as well as the product variation. Thus, the discrimination ratio shows the relative usefulness of the measurement system for the product being measured. The ratio estimate yields the number of non-overlapping categories within the control limits, or natural process limits, that the product could be sorted into if operator bias can be eliminated. The discrimination ratio is calculated as

$$DR = \sqrt{\frac{2\sigma_p^2}{\sigma_e^2} - 1}$$

Since operator bias is often present, it is useful to recalculate the discrimination ratio incorporating this bias and then comparing the two ratios. While the formula for the ratio remains the same, the estimates for σ_p^2 and σ_e^2 become

$$\sigma'^2_e = \sigma_m^2 = \sigma_e^2 + \sigma_o^2 \quad \text{and} \quad \sigma'^2_p = \sigma_p^2 + \sigma_o^2$$

so that

$$DR = \sqrt{\frac{2\sigma'^{2}_{p}}{\sigma'^{2}_{e}} - 1}$$

A *percent tolerance analysis* is sometimes preferred as a means of evaluating a measurement system. Values of % EV, % AV, and % R&R are calculated using the value of the specification tolerance (TOL) in the numerator as follows:

$$\% \text{ EV} = 100\ [(\text{EV})/(\text{TOL})]$$
$$\% \text{ AV} = 100\ [(\text{AV})/(\text{TOL})]$$
$$\% \text{ R\&R} = 100\ [(\text{R\&R})/(\text{TOL})]$$

Common guidelines for the interpretation of the % R&R are:

- % R&R < 10%, the measurement system is OK for use
- 10% < % R&R < 30%, the measurement system may be acceptable contingent upon its importance in application, cost of its replacement, cost of its repair, etc.
- % R&R > 30%, the measurement system is not to be used, and effort is needed to identify sources of excess variation and correct them

Another common evaluation is a *percent total variation analysis*. The computations are similar to the percent tolerance analysis with the exception that the denominator of the ratios is the total variation (TV).

$$\% \text{ EV} = 100\ [(\text{EV})/(\text{TV})]$$
$$\% \text{ AV} = 100\ [(\text{AV})/(\text{TV})]$$
$$\% \text{ R\&R} = 100\ [(\text{R\&R})/(\text{TV})]$$

Unfortunately, these are poor statistical metrics as they represent ratios of standard deviations. A more appropriate method is to express them as ratios of variances. In this manner, the ratios become variance components which sum to 100% when the % PV ratio is factored in as follows

$$\% \text{ EV} = 100\ [(\sigma_{EV})^2/(\sigma_t)^2]$$
$$\% \text{ AV} = 100\ [(\sigma_{AV})^2/(\sigma_t)^2]$$
$$\% \text{ R\&R} = 100\ [(\sigma_{R\&R})^2/(\sigma_t)^2]$$
$$\% \text{ PV} = 100\ [(\sigma_p)^2/(\sigma_t)^2]$$

Thus, the variance components can be graphically portrayed with a simple pie chart, or in a breakdown diagram as shown in Figure 3.8. Pure error, which is a component of repeatability, is the variability of repeated measurements without removing and re-fixturing the part. It is the smallest possible measurement error.

Gage accuracy is defined as the difference between the observed average of sample measurements and the true (master) average of the same parts using precision instruments. *Gage*

linearity is defined as the difference in the accuracy values of the gage over its expected operating range. Gage stability is defined as the total variation seen in the measurements obtained with the gage using the same master or master parts when measuring a given characteristic over an extended time frame. *Gage system error* is defined as the combination of gage accuracy, repeatability, reproducibility, stability and linearity.

3.4.4.6 Example - Gasket Thickness

A plant that manufactures sheets in the production of gaskets was concerned about the measurement of thickness. The engineer designed a gage R&R study to evaluate the measurement system. Three operators were chosen for the study and five different parts (gaskets) were chosen to represent the expected range of variation seen in production. Each operator measured each gasket a total of two times. The specification for thickness is 76±20 mm. The data are shown in Table 3.9.

Table 3.9 Gasket thicknesses for a gage R&R study

	Operator					
	A		B		C	
Part	1st Trial	2nd Trial	1st Trial	2nd Trial	1st Trial	2nd Trial
1	67	62	55	57	52	55
2	110	113	106	99	106	103
3	87	83	82	79	80	81
4	89	96	84	78	80	82
5	56	47	43	42	46	54

The data were entered in the data collection form in Table 3.10 and the summary statistics were computed for use in the calculations form in Table 3.11. The results of the gage R&R analysis are as follows:

$$EV = 5.15\sigma_{EV} = 4.267\ x\ 4.56 = 19.456$$

$$\sigma_{EV} = \frac{19.456}{5.15} = 3.778$$

$$AV = 5.15\sigma_{AV} = 5.15\sqrt{\left(8.5\ x\ 2.70\right)^2 - \left[\left(19.456\right)^2 / \left(10\right)\right]} = 22.078$$

$$\sigma_{AV} = \frac{22.075}{5.15} = 4.287$$

$$R\,\&\,R = 5.15\sigma_m = 5.15\sqrt{(3.778)^2 + (4.287)^2} = 29.427$$

$$\sigma_m = \frac{29.427}{5.15} = 5.714$$

$$PV = 5.15\sigma_p = 58.167\ x\ 2.08 = 120.790$$

$$\sigma_p = \frac{120.790}{5.15} = 23.454$$

$$TV = 5.15\sigma_t = 5.15\sqrt{(23.454)^2 + (5.714)^2} = 124.323$$

$$\sigma_t = \frac{124.323}{5.15} = 24.140$$

% EV = 100 [(19.456)/(40)] = 48.64%
% AV = 100 [(22.078)/(40)] = 55.19%
% R&R = 100 [(29.427)/(40)] = 73.57%

The % R&R value of 73.57% indicated that the measurement system was not acceptable. The engineer recommended to management that the measurement system should be investigated further to identify sources of variation that can be eliminated.

It was possible to compute a % PV value as part of the tolerance analysis, but the result was not very meaningful. For this study, the % PV is calculated to be

$$\% \text{ PV} = 100\ [(120.790)/(40)] = 301.97\%$$

The variance components were also calculated from the study results. These components gave the investigator some direction on where to focus efforts to reduce variation.

Using the calculations in Table 5, the variance component analysis is as follows:

$$\% \text{ EV} = 100\ [(\sigma_{EV})^2/(\sigma_t)^2] = 100\ [(3.778)^2/(24.140)^2] = 2.45\%$$

$$\% \text{ AV} = 100\ [\sigma_{(AV)}^2/\sigma_{(t)}^2] = 100\ [(4.287)^2/(24.140)^2] = 3.15\%$$

$$\% \text{ R\&R} = 100\ [(\sigma_{R\&R})^2/(\sigma_t)^2] = 100\ [(5.714)^2/(24.140)^2] = 5.60\%$$

$$\% \text{ PV} = 100\ [(\sigma_p)^2/(\sigma_t)^2] = 100\ [(23.454)^2/(24.140)^2] = 94.40\%$$

As expected, the values of % EV (2.45%) and % AV (3.15%) sum to the contribution of the gauge % R&R value of 5.60%. Most of the variation seen in the data (94.40%) was due to part-to-part differences.

Table 3.10 Gage Repeatability and Reproducibility Data Collection Sheet (long method) for example.

Operator	A						B						C						
Part #	1st Trial	2nd Trial	3rd Trial	4th Trial	5th Trial	Range	1st Trial	2nd Trial	3rd Trial	4th Trial	5th Trial	Range	1st Trial	2nd Trial	3rd Trial	4th Trial	5th Trial	Range	Average
1	67	62				5	55	57				2	52	55				3	58.000
2	110	113				3	106	99				7	106	103				3	106.167
3	87	83				4	82	79				3	80	81				1	82.000
4	89	96				7	84	78				6	80	82				2	84.833
5	56	47				9	43	42				1	46	54				8	48.000
6																			
7																			
8																			
9																			
10																			
Average	81.8	80.2				5.6	74.0	71.0				3.8	72.8	75.0				3.4	
		81.8				$\bar{R}_A$		74.0				$\bar{R}_B$		72.8				$\bar{R}_C$	
	Sum	162.0					Sum	145.0					Sum	147.8					
	$\bar{X}_A$	81.0					$\bar{X}_B$	72.5					$\bar{X}_C$	73.9					

$\bar{R}_A$	5.6
$\bar{R}_B$	3.8
$\bar{R}_C$	3.4
Sum	12.8
$\bar{\bar{R}}$	4.267

# Trials	D_4
2	3.27
3	2.58
4	2.28
5	2.11

$\left(\bar{\bar{R}}\right) x \left(D_4\right) = UCL_R$ *
(4.267) x (3.27) = 13.939

Max Oper $\bar{X}$	81.0
Min Oper $\bar{X}$	72.5
$\bar{X}_{diff}$	8.5

Max Part $\bar{X}$	106.167
Min Part $\bar{X}$	48.000
R_p**	58.167

* Limit of the individual ranges (Rs). Ranges beyond the limit are circled, and the cause should be identified and corrected. Repeat the readings with the same operator and same sample, or discard values and reaverage and recompute R and UCL_R from rest of data.

** Range of all the sample averages

Table 3.11 Gage Repeatability and Reproducibility Data Collection Sheet (long method) for example.

GAUGE REPEATABILITY & REPRODUCIBILITY REPORT

Gauge Name: Gasket Inspection
Gauge No.: C123
Gauge Type: calipers

Date: 10 / 08/ 04
Study Done By: Alan
Specification: 76 ± 20

Characteristic: Gasket Thickness
Part No. & Name: GW123
Spec. Tolerance (if 2-sided): 40 (TOL)

From Data Collection Sheet: $\bar{\bar{R}}$ = 4.267 $\bar{X}$ Diff. = 8.5 R_p = 58.167

Parts = 5 # Trials = 2

MEASUREMENT ANALYSIS

Repeatability (a.k.a. Equipment Variation, EV)

$EV = (\bar{\bar{R}}) x (K_1)$

= (4.267) x (4.56)
= 19.456

Reproducibility (a.k.a. Appraiser Variation, AV*)

$AV = \sqrt{[(\bar{X}_{diff}) x (K_2)]^2 - [(EV)^2 / (n x r)]}$

$= \sqrt{[(8.5) x (2.70)]^2 - [(19.456)^2 / (10)]}$
= 22.078

Repeatability and Reproducibility (R & R)

$R \& R = \sqrt{(EV)^2 + (AV)^2}$

$= \sqrt{(19.456)^2 + (22.078)^2}$
= 29.427

Part Variation (PV)

$PV = (R_p) x (K_3)$ = (58.167) x (2.08) = 120.790

Total Variation (TV)

$TV = \sqrt{(R \& R)^2 x (PV)^2}$

$= \sqrt{(29.427)^2 x (120.790)^2}$ = 124.323

# Trials	K_1	# Operators	K_2
2	4.57	2	3.65
3	3.04	3	2.70
4	2.50	4	2.30
5	2.21	5	2.08

# Parts	K_3
2	3.65
3	2.70
4	2.30
5	2.08
6	1.93
7	1.82
8	1.74
9	1.67
10	1.62

* If AV is a negative value within the square root sign, the appraiser variation will default to zero (0).

% TOLERANCE ANALYSIS

% EV = 100 [(EV)/(TOL)]
= 100[(19.456) / (40)]
= 48.64 %

% AV = 100 [(AV)/(TOL)]
= 100[(22.078) / (40)]
= 55.19 %

% R&R = 100 [(R&R)/(TOL)]
= 100[(29.427) / (40)]
= 73.57 %

% TOTAL VARIATION ANALYSIS

% EV = 100 [(EV)/(TV)]
= 100[(19.456)/(124.323)]
= 15.65 %

% AV = 100 [(AV)/(TV)]
= 100[(22.078)/(124.323)]
= 17.76 %

% R&R = 100 [(R&R)/(TV)]
= 100[(29.427)/(124.323)]
= 23.67 %

% PV = 100 [(PV)/(TV)]
=100[(120.790)/(124.323)]
= 97.16 %

The number of distinct categories was also be computed from the study results as

$$NDC = \frac{1.41\sigma_p}{\sigma_m} = \frac{(1.41)(23.454)}{5.714} = 5.8 \rightarrow 6$$

Since the value of categories was 6, the measurement system was acceptable. The discrimination ratio for this study incorporating operator bias was computed as

$$DR = \sqrt{\frac{2\left[(23.454)^2 + (4.287)^2\right]}{(5.714)^2} - 1} = 5.8 \rightarrow 6$$

which agreed with the number of distinct categories estimate. The fact that part-to-part variance component accounts for such a large portion of the total variation is consistent with a larger value of distinct categories the gauge can distinguish. The engineer recomputed the discrimination ratio under the assumption that the operator bias could be eliminated and found that

$$DR = \sqrt{\frac{2(23.454)^2}{(3.778)^2} - 1} = 8.7 \rightarrow 9$$

Thus, the engineer discovered that the measurement system could be improved from distinguishing six quality levels to nine quality levels by eliminating the operator bias which was possible through certification and training.

The analysis of gauge R&R studies is available in a wide array of software programs. If the gasket thickness example is treated as a crossed design with the operators and parts considered as fixed effects, the following analysis from Minitab is typical. The estimates shown here are consistent with those shown previously (minor differences due to rounding error). Note that the "VarComp" column represents the square of the standard deviation estimates σ_m, σ_e, σ_o, σ_p, and σ_t, respectively, which are shown in the "StdDev (SD)" column in the fourth table. Note that the number of distinct categories reflects a conservative estimate, i.e., 5.8 is rounded down to 5 instead of up to 6.

Two-Way ANOVA Table With Interaction

Source	DF	SS	MS	F	P
Part	4	12791.1	3197.78	247.730	0.000
Appraiser	2	415.4	207.70	16.090	0.002
Part * Appraiser	8	103.3	12.91	1.058	0.439 (not significant)
Repeatability	15	183.0	12.20		
Total	29	13492.8			

Two-Way ANOVA Table Without Interaction

Source	DF	SS	MS	F	P
Part	4	12791.1	3197.78	256.925	0.000
Appraiser	2	415.4	207.70	16.688	0.000
Repeatability	23	286.3	12.45		
Total	29	13492.8			

Gage R&R

Source	VarComp	%Contribution (of VarComp)
Total Gage R&R	31.972	5.68
Repeatability	12.446	2.21
Reproducibility	19.525	3.47
Appraiser	19.525	3.47
Part-To-Part	530.889	94.32
Total Variation	562.861	100.00

Source	StdDev (SD)	Study Var (5.15 * SD)	%Study Var (%SV)	%Tolerance (SV/Toler)
Total Gage R&R	5.6544	29.120	23.83	72.80
Repeatability	3.5279	18.169	14.87	45.42
Reproducibility	4.4188	22.757	18.63	56.89
Appraiser	4.4188	22.757	18.63	56.89
Part-To-Part	23.0410	118.661	97.12	296.65
Total Variation	23.7247	122.182	100.00	305.46

Number of Distinct Categories = 5

Alternatively, this gauge study could be treated as a nested design with parts and operators nested within parts. Note that the Minitab analysis for this model produces similar variance components compared to the crossed design model. It is also seen in this analysis that statistical differences still exist among the operators.

Nested ANOVA: Gasket Thickness versus Part, Appraiser

Analysis of Variance for Gasket Thickness

Source	DF	SS	MS	F	P
Part	4	12791.1333	3197.7833	61.654	0.000
Appraiser	10	518.6667	51.8667	4.251	0.006
Error	15	183.0000	12.2000		
Total	29	13492.8000			

Variance Components

Source	% of Var Comp.	Total	StDev
Part	524.319	94.24	22.898 (vs. 23.041 in crossed design)
Appraiser	19.833	3.56	4.453 (vs. 4.419)
Error	12.200	2.19	3.493 (vs. 3.528)
Total	556.353		23.587

3.4.5 The Simple Measurement Model

In moving to MSA applications involving variable data, much of what is done will concern calculation of variances under several scenario conditions. The simplest such assessment of a measurement system is the case where there is a single appraiser measures a single object. The model for this simple case is:

$$y = x + \varepsilon \tag{3.5}$$

In (3.5), y is the measurement result of an object whose true measure is x. The quantity x may also be considered as a standard value. The quantity ε represents the random error component assumed with mean 0 and some unknown variance σ^2. The ε variable is often further assumed to be normally distributed, but that is not a mandatory assumption. The quantity σ represents the standard deviation of simple repeatability. By stipulating that the mean of the distribution of ε is 0 we are assuming the measurement of x is unbiased. If the system were biased in a systematic way, we would have to add a constant $B \neq 0$ to (3.5) giving the enhanced model:

$$y = x + B + \varepsilon \tag{3.6}$$

If x is considered a constant single value, then the expected or average value of y is $x+B$. The model in (6) can be further expanded to accommodate a linearity effect by adding a slope term $m \neq 1$ to the model. This gives the more general linear model:

$$y = mx + B + \varepsilon \tag{3.7}$$

The effect of the linearity term m is to change the expectation of y as the true value, x, changes - since the expectation of y is now mx. In an ideal measurement system, we want m=1 and B=0. Then we would say the system is unbiased with perfect linearity. One simple analysis methodology that can be used to estimate values for m and B and to test the hypothesis of m=1 and B=0 is to use a simple linear regression of y on x using several values of x. Using this simple technique one can then determine confidence intervals for the model parameters m and B. Should we find that the confidence intervals for m and B include 1 and 0 respectively, we could then conclude that there is not enough evidence that the system is biased and/or contains a linearity effect. In the regression output, the *standard error of the estimate* is estimating σ and can then be taken as the estimate of the repeatability standard deviation. In this type of analysis, we can also check the normality and stability of the residuals using a probability plotting technique and a control chart. We use the variable d=y-x for each (x,y) pair in our data. d is in theory equal to ε, since y=x+ε under an assumption that m=1 and B=0 in model (3.7).

To illustrate with an example, a set of ten standard weights used in the calibration process of a certain type of scale were used in an MSA study. Four measurements were taken for each of the ten standard weights. The results are shown in Figure 3.13. Below the figure in Table 3.12 is shown the regression analysis summary. There we see the estimated model coefficients and their standard errors. For both coefficients the values m=1 and B=0 are well within one standard error of each estimate. There is no statistical reason to reject the system for being biased or for having a significant linearity effect. Using this method, the estimated standard deviation of repeatability is the regression standard error of the estimate or $\hat{\sigma} = 0.203$. There are several other ways that this simple analysis could be performed, but the regression methodology neatly fits this scenario. More details can be found in references in the Appendix to this section.

Figure 3.13 Scatter plot of the Regression Analysis of Scale weights.

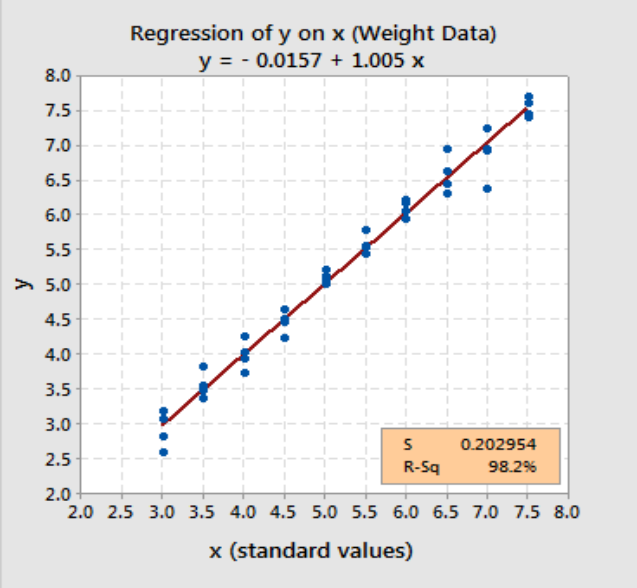
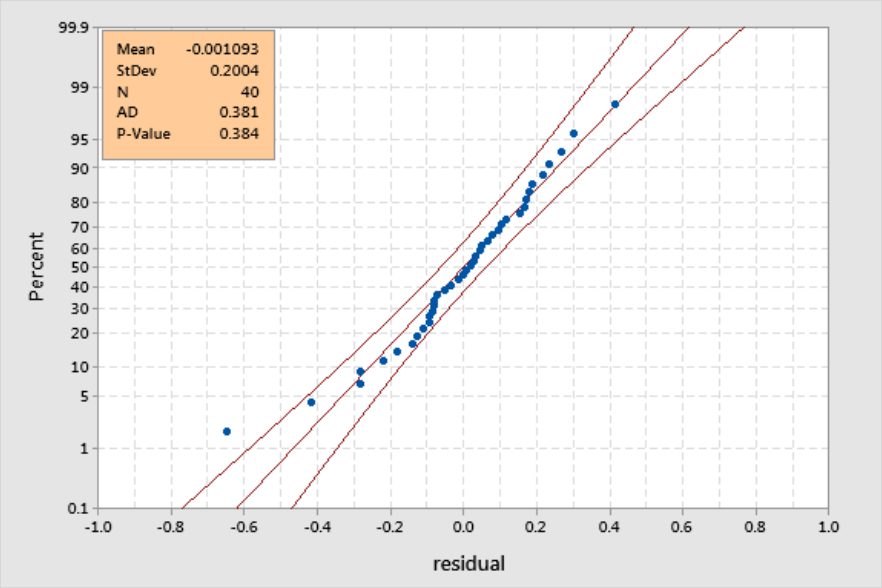


Table 3.12 Coefficient Estimates and Associated Standard Errors

Term	estimate	interpretation	SE
B	-0.0157	bias	0.1220
m	1.0052	slope (linearity)	0.0223
S	0.203	repeatability std. dev.	

The normal plot or the residuals for this example is shown in Figure 3.13. The residuals are well behaved and conform to the normal distribution. Other enhancement to the general model (3) for this case could have been used such as adding in a quadratic effect.

Figure 3.14 Normal Probability Plot of Residuals



Section 3.4 References

AS13003, SAE International, *Measurement Systems Analysis*, Warrendale, PA, 2015.

ASTM E2332, *Guide to Measurement Systems Analysis*, American Society for Testing Materials (ASTM), West Conshohocken, PA, 2004.

ASQC Automotive Division SPC Manual, American Society for Quality Control, Milwaukee, WI, 1986.

Automotive Industry Action Group, *Measurement Systems Analysis (MSA) Reference Manual, 4th Ed*, 2010.

Bishop, L., Hill, W. J. and Lindsay, W. S. , "Don't be Fooled by the Measurement System," *Quality Progress*, December, pp. 35-38, (1987).

Fleiss, J. L., Levin, B., Paik, M. C., *Statistical Methods for Rates and Proportions*, 3rd ed., Wiley Series in Probability and Statistics, John Wiley & Sons., NY, 2003.

Ott, E. R., Schilling, E. G., and Neubauer, D. V., *Process Quality Control*, 4th edition, ASQ Quality Press, Milwaukee, WI, 2005.

Shewhart, W.A., *Economic Control of Quality of Manufactured Product*, D. Van Nostrand Company, New York, p. 376, (1931).

Wheeler, D. J. and Lyday, R. W., *Evaluating the Measurement Process*, SPC Press Inc., 1989.

Section 3.4 Appendix

Table 3.A.1 Control Chart Constants d_2 for Samples of n_g

n_g	d_2
2	1.128
3	1.693
4	2.059
5	2.326
6	2.534
7	2.704
8	2.847
9	2.970
10	3.078
11	3.173
12	3.258
13	3.336
14	3.407
15	3.472
16	3.532
17	3.588
18	3.640
19	3.689
20	3.735
21	3.778
22	3.819
23	3.858
24	3.895
25	3.931

Table 3.A.2 Expanded Table of the Adjusted d_2 factor (d_2*) and df for Estimating the Standard Deviation from the Average Range

To be used with estimates of s based on k independent sample ranges of n_g each. (Unbiased estimate of σ^2 is (R-bar/d_2*)^2; unbiased estimate of σ is R-bar/d_2, where d_2 is from Table 3.A.1.)

	Subgroup size, n_g															
k	**n_g = 2**		**n_g = 3**		**n_g = 4**		**n_g = 5**		**n_g = 6**		**n_g = 7**		**n_g = 8**		**n_g = 9**	
Number of Samples	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*
1	1.400	1.0	1.910	2.0	2.239	2.9	2.481	3.8	2.672	4.7	2.830	5.5	2.963	6.3	3.078	*7.0*
2	1.278	1.8	1.805	3.6	2.151	5.5	2.405	7.2	2.604	8.9	2.768	10.5	2.906	12.1	3.024	13.5
3	1.231	2.6	1.769	5.4	2.120	8.2	2.379	10.9	2.581	13.4	2.747	15.8	2.886	18.1	3.006	20.3
4	1.206	3.5	1.750	7.3	2.105	11.0	2.366	14.5	2.570	17.9	2.736	21.1	2.877	24.1	2.997	27.0
5	1.191	4.4	1.739	9.1	2.096	13.7	2.358	18.1	2.563	22.3	2.730	26.3	2.871	30.2	2.992	33.8
6	1.180	5.3	1.732	10.9	2.091	16.4	2.353	21.7	2.557	26.8	2.725	31.6	2.866	36.2	2.987	40.6
7	1.173	6.1	1.727	12.7	2.086	19.2	2.349	25.4	2.554	31.3	2.722	36.9	2.863	42.2	2.985	47.3
8	1.167	7.0	1.722	14.5	2.083	21.9	2.346	29.0	2.552	35.7	2.720	42.1	2.861	48.2	2.983	54.1
9	1.163	7.9	1.719	16.3	2.080	24.6	2.344	32.6	2.550	40.2	2.718	47.4	2.860	54.3	2.981	60.8
10	1.159	8.8	1.717	18.2	2.078	27.4	2.342	36.2	2.548	44.7	2.717	52.7	2.858	60.3	2.980	67.6
11	1.156	9.6	1.714	20.0	2.076	30.1	2.341	39.9	2.547	49.1	2.715	57.9	2.857	66.3	2.979	74.3
12	1.154	10.5	1.713	21.8	2.075	32.9	2.339	43.5	2.546	53.6	2.714	63.2	2.856	72.4	2.979	81.1
13	1.152	11.4	1.711	23.6	2.074	35.6	2.338	47.1	2.545	58.1	2.714	68.5	2.856	78.4	2.978	87.9
14	1.150	12.3	1.710	25.4	2.073	38.3	2.338	50.7	2.544	62.5	2.713	73.7	2.855	84.4	2.977	94.6
15	1.149	13.1	1.709	27.2	2.072	41.1	2.337	54.3	2.543	67.0	2.712	79.0	2.855	90.5	2.977	101.4
20	1.144	17.5	1.705	36.3	2.069	54.8	2.334	72.5	2.541	89.3	2.710	105.3	2.853	120.6	2.975	135.2
25	1.141	21.9	1.702	45.4	2.066	68.4	2.332	90.6	2.540	111.6	2.709	131.7	2.852	150.8	2.974	169.0
30	1.138	26.3	1.701	54.5	2.065	82.1	2.331	108.7	2.539	134.0	2.708	158.0	2.851	180.9	2.973	202.8
40	1.136	35.0	1.699	72.6	2.064	109.5	2.330	144.9	2.538	178.6	2.707	210.7	2.850	241.2	2.973	270.3
60	1.133	52.6	1.697	108.9	2.062	164.3	2.329	217.4	2.536	267.9	2.706	316.0	2.849	361.8	2.972	405.5
∞	1.128	∞	1.693	∞	2.059	∞	2.326	∞	2.534	∞	2.704	∞	2.847	∞	2.970	∞

SOURCE: The approximation for d_2* is based on the approximation given by P. B. Patnaik in the paper, "The Use of Mean Range as an Estimator of Variance in Statistical Tests", Biometrika, Vol. 37, pp. 78-87, 1950. The calculation for the degrees of freedom is based on an extension to the approximation given by P. B. Patnaik, which was presented by H. A. David in the paper, "Further Applications of Range to the Analysis of Variance", Biometrika, Vol. 38, pp. 393-407, 1951, to improve the accuracy for $k > 5$.

Table 3.A.2 Expanded Table of the Adjusted d_2 factor (d_2*) and df for Estimating the Standard Deviation from the Average Range (continued) To be used with estimates of s based on k independent sample ranges of n_g each. (Unbiased estimate of σ^2 is (R-bar/d_2*)^2; unbiased estimate of σ is R-bar/d_2, where d_2 is from Table 3.A.1.)

	Subgroup size, n_g															
k	**n_g = 10**		**n_g = 11**		**n_g = 12**		**n_g = 13**		**n_g = 14**		**n_g = 15**		**n_g = 16**		**n_g = 17**	
Number of Samples	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*
1	3.179	7.7	3.269	8.3	3.350	9.0	3.424	9.6	3.491	10.2	3.553	10.8	3.611	11.3	3.664	11.9
2	3.129	14.9	3.221	16.2	3.305	17.5	3.380	18.7	3.449	19.9	3.513	21.1	3.572	22.2	3.626	23.3
3	3.112	22.4	3.205	24.4	3.289	26.3	3.366	28.1	3.435	29.9	3.499	31.6	3.558	33.3	3.614	34.9
4	3.103	29.8	3.197	32.5	3.282	35.0	3.358	37.5	3.428	39.9	3.492	42.2	3.552	44.4	3.607	46.5
5	3.098	37.3	3.192	40.6	3.277	43.8	3.354	46.9	3.424	49.8	3.488	52.7	3.548	55.5	3.603	58.1
6	3.095	44.7	3.189	48.7	3.274	52.6	3.351	56.2	3.421	59.8	3.486	63.2	3.545	66.5	3.601	69.8
7	3.092	52.2	3.187	56.8	3.272	61.3	3.349	65.6	3.419	69.8	3.484	73.8	3.543	77.6	3.599	81.4
8	3.090	59.6	3.185	65.0	3.270	70.1	3.347	75.0	3.417	79.7	3.482	84.3	3.542	88.7	3.598	93.0
9	3.089	67.1	3.184	73.1	3.269	78.8	3.346	84.4	3.416	89.7	3.481	94.9	3.541	99.8	3.596	104.6
10	3.088	74.5	3.183	81.2	3.268	87.6	3.345	93.7	3.415	99.7	3.480	105.4	3.540	110.9	3.596	116.3
11	3.087	82.0	3.182	89.3	3.267	96.4	3.344	103.1	3.415	109.6	3.479	115.9	3.539	122.0	3.595	127.9
12	3.086	89.4	3.181	97.4	3.266	105.1	3.343	112.5	3.414	119.6	3.479	126.5	3.539	133.1	3.594	139.5
13	3.085	96.9	3.180	105.6	3.266	113.9	3.343	121.9	3.413	129.6	3.478	137.0	3.538	144.2	3.594	151.1
14	3.085	104.4	3.180	113.7	3.265	122.6	3.342	131.2	3.413	139.5	3.478	147.5	3.538	155.3	3.593	162.8
15	3.084	111.8	3.179	121.8	3.265	131.4	3.342	140.6	3.412	149.5	3.477	158.1	3.537	166.4	3.593	174.4
20	3.083	149.1	3.178	162.4	3.263	175.2	3.340	187.5	3.411	199.3	3.476	210.8	3.536	221.8	3.592	232.5
25	3.082	186.4	3.177	203.0	3.262	219.0	3.340	234.4	3.410	249.2	3.475	263.5	3.535	277.3	3.591	290.7
30	3.081	223.6	3.176	243.6	3.262	262.8	3.339	281.2	3.410	299.0	3.475	316.2	3.535	332.7	3.590	348.8
40	3.080	298.2	3.175	324.8	3.261	350.4	3.338	375.0	3.409	398.7	3.474	421.6	3.534	443.7	3.590	465.1
60	3.079	447.2	3.175	487.2	3.260	525.6	3.337	562.5	3.408	598.0	3.473	632.3	3.533	665.5	3.589	697.6
∞	3.078	∞	3.173	∞	3.258	∞	3.336	∞	3.407	∞	3.472	∞	3.532	∞	3.588	∞

SOURCE: The approximation for d_2* is based on the approximation given by P. B. Patnaik in the paper, "The Use of Mean Range as an Estimator of Variance in Statistical Tests", Biometrika, Vol. 37, pp. 78-87, 1950. The calculation for the degrees of freedom is based on an extension to the approximation given by P. B. Patnaik, which was presented by H. A. David in the paper, "Further Applications of Range to the Analysis of Variance", Biometrika, Vol. 38, pp. 393-407, 1951, to improve the accuracy for $k > 5$.

Table 3.A.2 Expanded Table of the Adjusted d_2 factor (d_2*) and df for Estimating the Standard Deviation from the Average Range (continued) To be used with estimates of s based on k independent sample ranges of n_g each. (Unbiased estimate of σ^2 is (R-bar/d_2*)^2; unbiased estimate of σ is R-bar/d_2, where d_2 is from Table 3.A.1.)

	Subgroup size, n_g															
k	n_g = 18		n_g = 19		n_g = 20		n_g = 21		n_g = 22		n_g = 23		n_g = 24		n_g = 25	
Number of Samples	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*	d_2*	*df*
1	3.714	12.4	3.761	12.9	3.805	13.4	3.847	13.8	3.887	14.3	3.924	14.8	3.960	15.2	3.994	15.6
2	3.677	24.3	3.725	25.3	3.770	26.3	3.813	27.2	3.853	28.1	3.891	29.0	3.928	29.9	3.962	30.8
3	3.665	36.4	3.713	37.9	3.759	39.4	3.801	40.8	3.842	42.2	3.880	43.6	3.917	44.9	3.952	46.2
4	3.659	48.6	3.707	50.6	3.753	52.5	3.796	54.4	3.836	56.3	3.875	58.1	3.912	59.9	3.947	61.6
5	3.655	60.7	3.704	63.2	3.749	65.7	3.792	68.1	3.833	70.4	3.872	72.6	3.908	74.8	3.943	77.0
6	3.653	72.9	3.701	75.9	3.747	78.8	3.790	81.7	3.831	84.4	3.869	87.1	3.906	89.8	3.941	92.4
7	3.651	85.0	3.699	88.5	3.745	92.0	3.788	95.3	3.829	98.5	3.868	101.7	3.905	104.7	3.940	107.7
8	3.649	97.2	3.698	101.2	3.744	105.1	3.787	108.9	3.828	112.6	3.867	116.2	3.903	119.7	3.939	123.1
9	3.648	109.3	3.697	113.8	3.743	118.2	3.786	122.5	3.827	126.7	3.866	130.7	3.903	134.7	3.938	138.5
10	3.648	121.4	3.696	126.5	3.742	131.4	3.785	136.1	3.826	140.7	3.865	145.2	3.902	149.6	3.937	153.9
11	3.647	133.6	3.696	139.1	3.741	144.5	3.785	149.7	3.826	154.8	3.864	159.8	3.901	164.6	3.936	169.3
12	3.646	145.7	3.695	151.8	3.741	157.6	3.784	163.3	3.825	168.9	3.864	174.3	3.901	179.6	3.936	184.7
13	3.646	157.9	3.695	164.4	3.740	170.8	3.784	176.9	3.825	183.0	3.863	188.8	3.900	194.5	3.936	200.1
14	3.645	170.0	3.694	177.1	3.740	183.9	3.783	190.6	3.824	197.0	3.863	203.3	3.900	209.5	3.935	215.5
15	3.645	182.2	3.694	189.7	3.740	197.0	3.783	204.2	3.824	211.1	3.863	217.9	3.900	224.4	3.935	230.9
20	3.644	242.9	3.693	252.9	3.739	262.7	3.782	272.2	3.823	281.5	3.862	290.5	3.899	299.3	3.934	307.8
25	3.643	303.6	3.692	316.2	3.738	328.4	3.781	340.3	3.822	351.8	3.861	363.1	3.898	374.1	3.933	384.8
30	3.643	364.3	3.691	379.4	3.737	394.1	3.781	408.3	3.822	422.2	3.861	435.7	3.898	448.9	3.933	461.8
40	3.642	485.8	3.691	505.9	3.737	525.4	3.780	544.4	3.821	562.9	3.860	580.9	3.897	598.5	3.932	615.7
60	3.641	728.7	3.690	758.8	3.736	788.2	3.779	816.7	3.821	844.4	3.859	871.4	3.896	897.8	3.932	923.5
∞	3.640	∞	3.689	∞	3.735	∞	3.778	∞	3.819	∞	3.858	∞	3.895	∞	3.931	∞

SOURCE: The approximation for d_2* is based on the approximation given by P. B. Patnaik in the paper, "The Use of Mean Range as an Estimator of Variance in Statistical Tests", Biometrika, Vol. 37, pp. 78-87, 1950. The calculation for the degrees of freedom is based on an extension to the approximation given by P. B. Patnaik, which was presented by H. A. David in the paper, "Further Applications of Range to the Analysis of Variance", Biometrika, Vol. 38, pp. 393-407, 1951, to improve the accuracy for $k > 5$.

Section 3.5 – Data Collection

3.5.1 Section Objectives

In this section, we provide effective techniques for acquiring the right amount of the right kind of data to guide research efforts toward improvement. Readers should come away with an understanding of the fundamentals of the statistical Design of Experiments (DOE), including its historic origin, its purpose and its advantages over competitive methods of inquiry and improvement. The section is not intended to be comprehensive of all DOE theory and methods. Rather, a summary is offered, with coverage of some of the most frequently used designs and with references to greater detail to be found elsewhere.

3.5.2 DOE History and Ronald Fisher's Contributions

During the early 20th century, England found itself unable to feed its population. Insufficient agricultural yield was augmented by imports, and while agricultural experiment stations existed, their research was helter-skelter by modern standards. Enter onto the scene a young mathematician, a Cambridge University honor graduate. With degree in hand, Ronald Fisher had left the university to work for an investment company, then entered farming and then school teaching. None of these occupations suited him, and his mentors and tormentors were not impressed with his performance. Some say he was not tactful. Others say he did not suffer fools lightly. He returned to academe.

Further in his research into modeling genetic behavior (for which he was knighted), he did not see eye-to-eye with the key statistical leadership of his day, Karl Pearson. He dropped his early work on eugenics because he saw it was leading nowhere, but that activity further ostracized him from the statistical mainstream despite the fact that its members were greatly impressed by his mathematical skill and insight. (Salsburg, 2001)

In 1919, Fisher was offered a position at the Rothamstead Agricultural Experiment Station where staff had accumulated a mountain of data and little knowledge of how to extract its meaning. While he spent years pouring over Rothamstead's volumes, it is doubtful that he learned much of real value from them. The problem, in brief, was one-factor-at-a-time experimentation. Information-less data was often the consequence of attempting a lone, sans control experiment one year with a follow-up characterized by a slight variation the following year. Of course, the yearly difference was influenced by many of Mother Nature's favorite attributes, including shifting temperatures, varying rainfall, and different soil conditions.

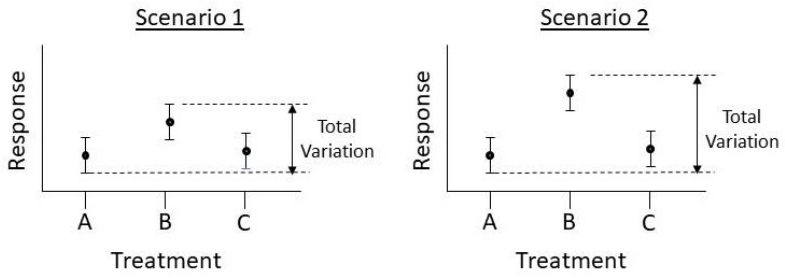
What Fisher brought to the table was sensible, planned, multifactor experimentation techniques that minimize the effects of natural and human-induced variation through randomization, balance and replication. He sat back while those clinging to their old ways raged. Experimental design had been used before, but randomization was foreign to a deterministic world.

Then he explained in mathematical terms how this new way of experimentation works and what it brings. He introduced what he called the analysis of variance. Explanations were later published (Fisher, 1925) with more underpinning mathematical detail to follow.

In this sense, Fisher may be considered to have been the world's first statistical engineer!

We might wonder how information about variances might impart knowledge of differences among means. Consider alternative scenarios as in Figure 3.15. Here we see three treatment means depicted graphically and surrounded by statistical intervals (Hahn and Meeker, 1991) In Scenario 1, the intervals overlap hinting that the variation among the means may be due to chance, alone. The opposite is true in Scenario 2; the intervals surrounding the means fail to overlap. Notice, also that the total variation, as depicted in each scenario as the distance between the lower bound of the lowest mean and the upper bound of the highest mean, differs between scenarios. It is much larger in Scenario 2 because of the differences among means. This is because the total variation in all the data is a function of the variation associated with an individual treatment combined with the variation among the treatment means. Fisher's genius is in recognizing this phenomenon and generalizing it to multiple factors, their levels and their interactions (described in the next section). Hmm … why didn't we think of that?

Figure 3.15 How the Analysis of Variance Identifies Mean Differences



Largely due to Fisher's genius and advice, and due to the diligence of those who followed it, England now feeds itself. But it did not end there, and it certainly did not end with agricultural experimentation, while many thought it should.

After the Second World War, George Box and J. Stuart Hunter, accompanied variously by others and sponsored in part by the Chemical Division of the American Society for Quality Control (Now the Chemical and Process Industries Division of the American Society for Quality) traveled the United States preaching the gospel of statistical methods, especially the statistical design of experiments, greatly to the influence of chemical, automotive and other technically based industries. (Box, Hunter and Hunter, 2005; Box, 2013) This sparked a revolution in the approach to all manners of experimentation beyond those initial applications and extending into pharmaceuticals, foods, electronics, health care, psychology and countless other activities. Their efforts changed the way the world executes scientific inquiry.

Fisher's genius extended beyond those already mentioned. He brought us the concept of statistics, itself. Earlier, Karl Pearson had argued that data, even sampled data, were wholly

contained of information whereas Fisher's view – the one that stuck – was that sampled data are used to obtain statistics which are estimates of parameters of a larger population. Fisher also brought us the notion of degrees of freedom, discriminant analysis, maximum likelihood estimation and many other original ideas. He is correctly called the Father of Statistics (and the father-in-law of George Box).

For the record, Fisher was not a fan of the formalized Neyman-Pearson school of statistics. Jerzy Neyman was a Polish born mathematical statistician. Neyman and Egon Pearson, Karl Pearson's son, taught together at University College, London. Together, they espoused a rigorous, highly probabilistic form of hypothesis testing which is taught today in beginning statistics courses and is highly embraced by government agencies globally and nearly worshiped in many industries, including pharmaceutical research.

While Fisher had done the math and calculated exact probabilities, e.g., F- Distribution tail areas, his approach to research downplayed the probabilities in favor of allowing seemingly rare or near-rare events to guide further research. He advocated using what he named "p-values" to categorize factors into "significant," meaning potentially influential, and others which did not emerge above the noise inherent in an experiment. Those others were to be kept in mind for future experimentation. It is likely that he did this because of the murky waters of probability, itself. Philosophically, what is probability? How can one embrace the calculation of probabilities based on a "null hypothesis" whose related experiment would never be run if those in control really believed it is true?

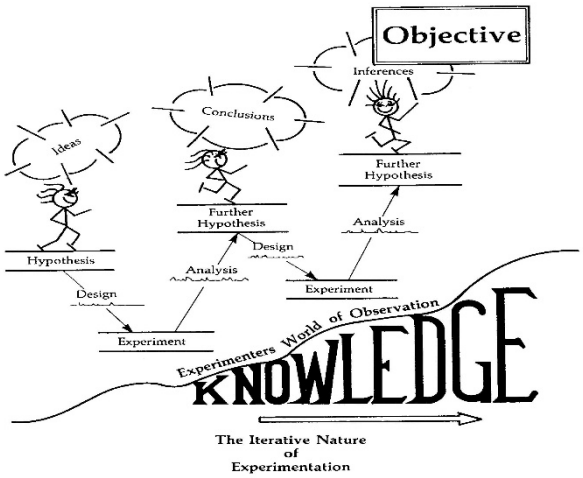
For more about Fischer and the history of statistics, see Salsburg (2001). For more on the shaky ground of probability and hypothesis testing see Weisberg (2014). For more about the perils of reliance on p-values, see Wasserstein and Lazar (2016).

3.5.3 Purpose and Strategy of DOE

Students who survived beginning statistics courses might come away believing that DOE is simply an extension of Student's t-test, allowing fair comparisons among multiple treatments. It is all about controlling the probability of a Type I error (rejecting the hypothesis of no difference when it is actually true), easily forgotten or put aside in favor of more important things. After all, those survey courses, well taught as they were, covered so many topics and tools that they caused confusion about what tool to use when and why. The resulting resolution is avoidance. It is to seek expert assistance, and then only when absolutely necessary.

This is a sad commentary in that it misses opportunities for understanding and using DOE as a powerful, strategy for improvement. Certainly, DOE can be used for comparing multiple treatments. Certainly, it can be used to ferret out multiple sources of variation. But its greatest impact comes when it is used as a strategy to cut through the cacophony of candidate variables shrouded in noise, correlation and confusion.

Figure 3.15 The Iterative Nature of Experimentation



Scientific inquiry is an iterative process beginning with synthesis, – meaning ideas emerging from subject matter knowledge combined with pie-in-the-sky thinking – turning to analysis and then returning to synthesis, all in repetition, and all the while, knowledge grows. This wash, rinse and repeat loop can be a bitter pill for those seeking quick, one-step solutions. W. Edwards Deming (1986) is quoted as saying, "It does not happen all at once. There is no instant pudding." To be fair, he was talking about organizational change, but the same is true of a sound process of experimentation, which is fundamental to process change and improvement.

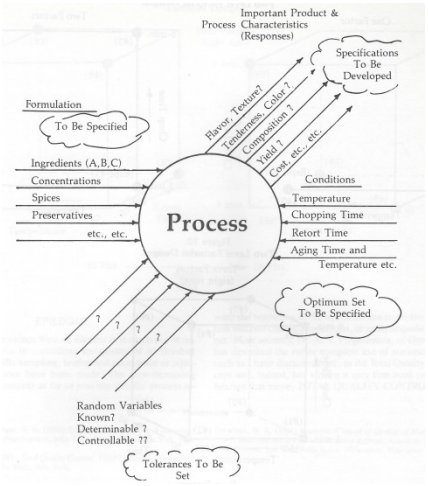
Truth be told, and like it or not, scientific inquiry moves along this path described in the opening pages of Statistics for Experimenters (Box, Hunter and Hunter,2005) and other DOE references, stemming from the scientific method of inquiry.

A favorite illustrative diagram (Figure 3.15) was modified from the lecture notes of Prof. Horace P. Andrews (Snee, Hare and Trout, 1985) whose lecture style and artistry could capture the imagination of the sleepiest of students. He depicted the iterative cycle of experimentation as a pleasurable journey, contrary to what some might think, from initial conjecture (hypothesis) to objective through the building of knowledge. A most important point: it is the creation of knowledge, not a single experiment nor a design point from a collection of experiments, which leads to the objective.

In a phrase, the purpose of DOE is to build knowledge.

But how? A strategy that works well in most situations is one of applying screening designs, followed by characterizing designs and then followed by optimizing designs. Screening designs assess broad linear factors which are thought possibly to influence responses.

Figure 3.16 Key Aspects of Processes



Consider Figure 3.16. It first appeared (Andrews, 1964) to explain the role of statistics in setting food specifications, but it is easily generalized to all process improvement situations. The central point of focus is the process. It has various input sources, here categorized as formulation, conditions and random variables, and it has output responses that, in this case, aid in the setting of specifications. The important thing to note is that the diagram captures key aspects of the process for all to embrace at a glance. It serves an interdisciplinary team as a planning aid.

Now, the team's first reaction might be "Holy cow" or some other food-related comment about the sheer number of factors. "We have to eliminate some of them. There are too many to run in an experiment. We don't have the resources." Then, the temptation will be to eliminate some of the factors. "We all know that spice isn't important," someone will say. However, "What gets us into trouble is not what we don't know. It's what we know for sure that just ain't so." (Mark Twain) Do we have data to suggest that spice does not make a difference or that it does not combine with other ingredients or input factors to make a difference? If not, there may be danger in eliminating it.

The alternative to data-less elimination of factors is to use a class of screening designs, described later, to identify the factors that will point us most efficiently in the direction of success. The resulting decision is databased and nearly opinion-free.

A great strategy for following up screening designs is the use of characterizing designs. These designs aid in the isolation and identification of interactions or joint effects among factors. The data generated by them are analyzed using relatively simple linear models to estimate the effects of factors and their interactions.

What are interactions? Consider, for example, the ever-popular hot fudge sundae. It is not so much the amount of fudge that makes it so wonderful; nor is it the amount of ice cream. It is how the fudge and the ice cream work together. The bitterness of the fudge contrasts with the sweetness of the vanilla; the hot of the fudge clashes with the cold of the ice cream. Together,

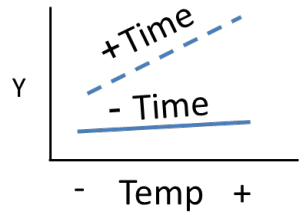
these two tantalizing battles operate in concert to dazzle the taste buds and send fireworks of delight to the brain.

Interactions of this sort are commonplace. Chemists refer to their effects as either antagonism or synergism depending on whether they combine negatively or positively. Other disciplines may use other words, but the impact is the same. It is how two or more factors work together to influence a response apart from the sum of the individual factors, themselves. Failure to recognize their existence is all too often a major cause of project failure. That is how important planning for their existence and measurement is. Those who doubt the importance of interactions may not be aware of what happens when molecular hydrogen and oxygen are combined and allowed to react.

From the sequence of screening designs and characterizing designs comes the identification of the relatively few factors and their interactions that guide the path to knowledge building necessary to success. Finding the sweet spot is the task of optimizing designs. Later sections of this chapter show that most screening designs and characterizing designs hold factors at two levels each. That is convenient because it helps to minimize the number of experimental treatment combinations. There it is assumed that the world is locally linear or at least gently sloping.

But as we approach the mountain peak, the terrain is more curved, so we need more exacting tools for assessment. More than two levels of each factor are required as are more model terms. Effects of linear, quadratic and interaction terms are usually estimated, and sometimes, depending upon response complexity, even higher order terms are necessary.

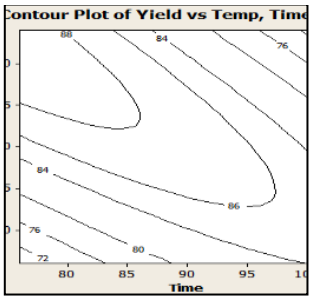
Figure 3.17 A Simple 2-Factor Interaction Plot



Regardless of design complexity, data modeling is usually essential to interpretation and understanding of resulting data. Models, however, rarely stand alone. To enhance their message, graphical techniques are essential. In the case of simple linear models, such as those used to analyze data from screening and characterizing designs, straight line graphs aid understanding and facilitate communications. Figure 3.17 shows a hypothetical example. There, a response (Y) is shown as a joint function of both processing time and temperature. Better than words can, this graph shows that if processing time is brief, increasing temperature has little effect on the response, but if processing time is longer, increasing temperature increases the response.

In optimizing situations, graphs are even more vital. Figure 3.18 shows a response surface, with "Time" on the horizontal axis and "Temperature" on the vertical axis. Labeled lines, or contours,

Figure 3.18 A Response Surface Plot



show predicted locations of constant response. These lines are analogous to isobars on a weather map and contour lines appearing on a topographical map. From Figure 3.18 it is clear that as Time decreases and Temperature increases, Yield increases. As both Time and Temperature increase together or decrease together, Yield decreases.

Carried out correctly, DOE can be leveraged to guide research, thereby avoiding blind alleys and dead ends. Head-on-head or one-factor-at-a-time experimentation is often expensive of resources and morale because it often leads nowhere, but the best designs lead the users to subsequent experimentation through the establishment of relations among factors.

Take, for example, a situation in which a beginning researcher – call her Mary – is asked to investigate methods for process improvement. After some study, she suspects that increasing temperature might yield positive results. Examination of process data including estimates of variation together with revisiting her hardly worn statistics text led her to believe that 8 observations, randomized of course, at each of two temperatures, will provide the power necessary to find differences, should they exist, that might be important to the company. She presents Table 3.13a to the boss. (The numbers are not real, but she knows the boss likes numbers.)

The boss thinks about this plan for a while and, being the boss, suggests a way to get more information out of these same 16 experimental treatment combinations: how about if we run half of them at the present operating time and the other half at a slightly longer time? Mary disappears and returns later with the modified design shown in Table 3.13b. Yes, that is better, more information in the same number of runs. But you know, Mary, I have been thinking. What if we were to add agitation rate? We could still run only 16 experimental treatment combinations, but we would get even more information. Mary revisits her spreadsheets and comes back with Table 3.13c.

Table 3.13a A Simple Experiment

- Temp	+ Temp
5.1	4.2
4.7	4.5
2.3	4.2
3.5	5.4
4.8	4.8
5.7	4.1
4.2	4.9
4.1	5.3

Table 3.13b A Simple Experiment – Stage 2

	- Temp	+ Temp
-Time	5.1	4.2
	4.7	4.5
	2.3	4.2
	3.5	5.4
+Time	4.8	4.8
	5.7	4.1
	4.2	4.9
	4.1	5.3

Table 3.13c A Simple Experiment – Stage 3

		- Temp	+ Temp
-Time	-RPM	5.1	4.2
		4.7	4.5
	+RPM	2.3	4.2
		3.5	5.4
+Time	-RPM	4.8	4.8
		5.7	4.1
	+RPM	4.2	4.9
		4.1	5.3

You can guess what happens next. The boss wants to economize further, so Mary returns to her spreadsheet, incorporates the boss's new idea and returns with Table 3.13d incorporating two different oil levels.

Table 3.13d A Simple Experiment – Stage 4

			- Temp	+ Temp
-Time	-RPM	-Oil	5.1	4.2
		+Oil	4.7	4.5
	+RPM	-Oil	2.3	4.2
		+Oil	3.5	5.4
+Time	-RPM	-Oil	4.8	4.8
		+Oil	5.7	4.1
	+RPM	-Oil	4.2	4.9
		+Oil	4.1	5.3

Now, from a DOE perspective, Mary has gone from a design with a single factor at 2 levels, to a design with 2 factors each at 2 levels each, to a design with 3 factors at 2 levels each and on to a design with 4 factors at 2 levels each. Later, in Section 3.5.4.5 on fractional factorial designs we will show how Mary can add more factors while holding the number of experimental treatment combinations constant.

But first, let us consider the progression here. If Mary had only run the experiment of Table 3.13a, she might have come up with no difference, in which case she would have come to a dead end, no information. If she had only run the experiment of Table 3.13b, she would increase her chances of discovering key effects, but not as greatly as if she had run the experiments of Tables 3.13c, or even more so, 3.13d.

Not only that – and this is important – the question changes as we move through these tables from "Does this factor make a difference?" to "Which factor or factors, acting singly or in combination, will lead to process improvement?" This represents a strategic shift in planning experiments in that it leverages the data, not opinions nor guesses, to guide experimentation and the resulting knowledge building.

This strategy, coupled with DOE technology, avoids blind alleys and dead ends, and instead leverages data to build knowledge, the basis for improvement.

3.5.4 Frequently used experimental designs

What follows is an introduction and brief rationale for some popular designs. While this section is not intended to serve as an endorsement of the use of any specific design for a given situation, it is presented here to facilitate understanding of design use and how each might fit into a larger research strategy.

3.5.4.1 One-way classification

As mentioned in Section 3.5.3, a take-away from an elementary statistics class might be that DOE is useful for comparing treatment differences. Indeed, it is. While DOE is capable of so much more, the one-way classification might be an informative, introductory plan and understanding its use might open broader vistas.

Table 3.14 A One-Way Classification Layout

Observation	Treatment 1	Observation	Treatment 2	Observation	Treatment 3
1	4.7	6	4.7	12	0.1
2	7.6	7	2.7	13	8.4
3	2.8	8	2.1	14	1.7
4	0.5	9	1.0	15	4.4
5	1.0	10	6.0	16	9.3
		11	1.3	17	1.6
				18	7.7

As an example, suppose an experiment is set up to examine three distinct treatments. They might be three different suppliers of an important ingredient, with the response being the percentage of an active component. Or they might be three distinct treatments for an illness, with the response being time to recovery. There may be more than three treatments, and the number of observations under each treatment is not necessarily the same.

In setting up, the experimenters must be sure to minimize potential bias by randomizing the allocation of samples or people to treatments and by assuring uniform environmental conditions for each treatment. This is essential because the data analysis is based on the calculation of probabilities, and those calculations assume randomization.

Table 3.14 lists treatments and data corresponding to a specific one-way classification. Analysis of data generated by this and most other designs is carried out using an analysis of variance (ANOVA) model. See Montgomery et.al. (2012) 4, specifically Section 4.2 for more information about the ANOVA. More than likely, data analysts will use canned software to carry out the calculations. Care should be taken to assure that the software is doing what was expected. One check is to write out the anticipated sources of variation and corresponding degrees of freedom. These are the number of levels of a factor, minus 1. If a factor has k levels, its degrees of freedom are k – 1.

The ANOVA partitions the total variation into its assignable causes. In the case of the example illustrated in Table 3.14, the ANOVA would be used to partition the total variation into that due to treatments and that due to observations within treatments. Table 3.15 shows that because

Table 3.15 ANOVA Sources and Degrees of Freedom	
Source of Variation	Degrees of Freedom
Total	17
Treatments	2
Observations within Treatments	15

there are 18 data points in the table, there are 17 degrees of freedom in total. Treatment degrees of freedom are 2 because there are 3 treatments. Finally, the degrees of freedom for observations within treatments may be obtained two ways. One way is to count the degrees of freedom among observations within each treatment (4 + 5 + 6 = 15). The other way is to calculate them by subtraction: if there are 17 degrees of freedom, all told, and if 2 of them are associated with Treatment, the remaining 15 must relate to Observations within Treatments.

If your software produces the anticipated degrees of freedom, there is a good chance it has the corresponding sums of squares, mean squares and F-ratios right, as well. In the case of this example, the F-ratio is the ratio of the mean square for treatments to the mean square for observations within treatments, and it has 2 degrees of freedom in the numerator and 15 in the denominator. Fisher, after whom the F-ratio is named, worked out the theoretical distribution of such ratios. That was not an easy thing in his day: calculations were carried out painstakingly with the aid of a crude (by our standards) calculator. He produced tables with tail area probability points corresponding to numerator and denominator degrees of freedom. Your software goes a step further by calculating the area under the F-distribution curve above the calculated value. You read an exact value, whereas Fisher's tables bracketed one.

However, Fisher did not get caught up in exact F-distribution probabilities, and we should not either. If the tail area (or p-value) is low, we should suspect that the treatment means are different by more than chance alone would allow. If it is not low, then either our experiment was not large enough or the means do not differ. Note that we are inclined to use the word "significant" in this sense, but in this context, significant does not mean important or vital. It simply means that the difference we are seeing has risen above the noise level. The value of its impact is a business calculation.

For those data sets with very low Treatment p-values, the next question is which treatment means differ from which other treatment means. To determine this, you could carry out three separate t-tests; Treatment 1 against Treatment 2, Treatment 1 against Treatment 3, and Treatment 2against Treatment 3. While it is not recommended, for each, you could fix the desired p-value necessary to exceed at some given level, say α. (Actually, that is what Neyman and Pearson did, to Fisher's chagrin.) It turns out that if there are k treatments, the actual probability of declaring a difference when none exists is: $P = 1 - (1 - \alpha)^k$. For example, if there are 3 treatments and if α is fixed at 0.05, the experiment-wise P is 0.14. The point is that carrying out multiple t-tests compounds the experiment-wise probability of falsely declaring a difference. There is no law against doing this, and many software packages provide this option, but you should go into this territory knowing what is happening to the underlying probabilities.

Recognizing this inconsistency, John Tukey (1949) developed the Honest Significant Difference (HSD), implying somewhat less than subtly, something deceptive about the use of multiple t-tests. Most statistical software packages offer this test. The benefit is that it preserves the same experiment-wise probability of error as is used in the ANOVA.

The other, very important feature of this software is that it provides excellent graphical displays of this and other statistical intervals about the treatment means. These are essential to both interpretation and communication.

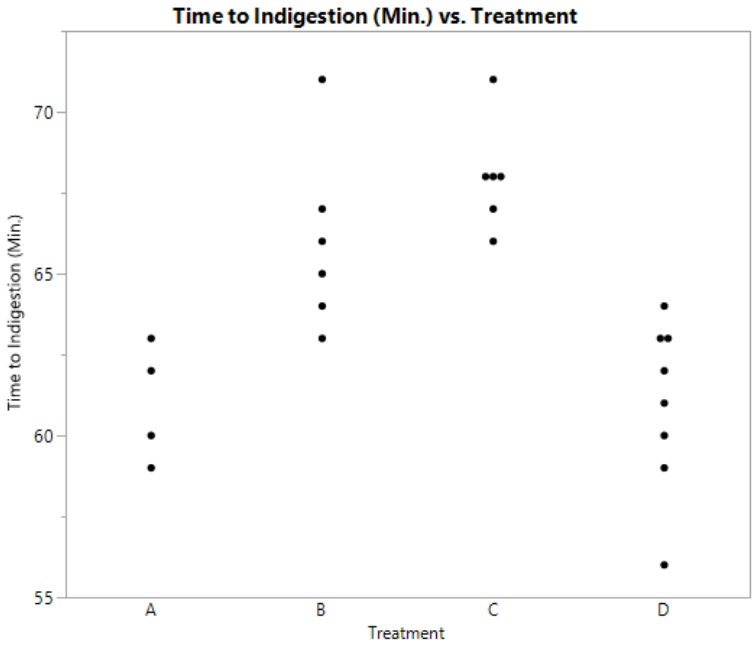
Example:

Table 3.16 lists data relating to the time in minutes to the onset of indigestion as a result of consuming each of 4 different food recipes. Twenty-four unfortunate wives endured their husbands' barbeque delights.

Table 3.16 Time (minutes) to onset of indigestion.

Recipe			
A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59

Figure 3.19 Plot of Raw Data of Table 3.16



The first and perhaps the most important thing you can do to analyze data is plot them. A first law of data analysis is, "Always, always, always, without exception, plot the data – and look at the plot." On a table or in a large column of data, if someone records 27 when they meant 72, you may not see it, but if you plot the data, you will. That and many other features stand out when you plot the data. From Figure 3.19, you may wonder about the lone high value corresponding to Treatment B, or you may wonder why some subjects within Treatments C and D are identical. Were the wives talking to each other?

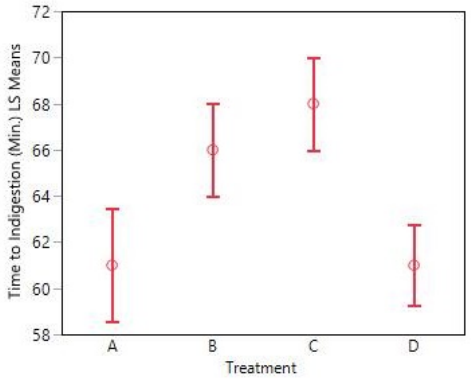
Table 3.17 ANOVA for Data in Table 3.16

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob > F
Total	23	340			
Model	3	228	76	13.57	<.0001
Error	20	112	5.6		

If you are satisfied from the plot and other information that the data are uninfluenced by sources of bias, you can proceed to the ANOVA as shown in Table 3.17. Do the degrees of freedom correspond to your understanding of the data? If not, you and your software are on different tracks. If so, it is safe to proceed to the interpretation.

Notice that the F-distribution tail area corresponding to the calculated F-ratio of 13.57 is tiny, suggesting that there are differences among recipes. The next step is to examine the graph of Tukey HSDs provided by the software as in Figure 3.20.

Figure 3.20 Tukey's HSD Intervals Surrounding Treatment Means



What is the interpretation? Recipes B and C take more time to cause indigestion, but we do not see a difference between them. This do not mean that they do not differ. However, it takes them longer to act than A and D which also cannot be said to differ.

You might wonder why some HSD intervals are longer than others. That is because there are fewer victims in the recipes corresponding to the longer intervals. Fewer observations mean greater uncertainty. You might also wonder why wives put up with this sort of thing.

A final point about the one-way classification analysis: the ANOVA is carried out under the assumption that the variation among observations within each treatment is uniform. Your software should contain a test for this assumption.

3.5.4.2 Randomized block designs

Only a modest increase in sophistication is needed to move from the one-way classification to the randomized block design. Suppose you have several different and distinct treatments, and you have subjects within treatments to consider, just as in the one-way classification.

In addition, suppose the subjects are further classified into homogeneous groups. You might imagine yields resulting from several treatments on elements of distinct batches of raw materials or liver protein levels resulting from different diets given to male siblings from several litters of laboratory rats.

Granted, the latter example is not nice to rats, but it does serve to illustrate an opportunity to remove some of the cloud of variation from the treatment comparisons. As we did with the one-way classification, we partition the total variation into its assignable causes, but with the randomized block design, we can further partition out (of the way), the block (in this case litter-to-litter) variation. Doing so sharpens the tool for detecting treatment differences.

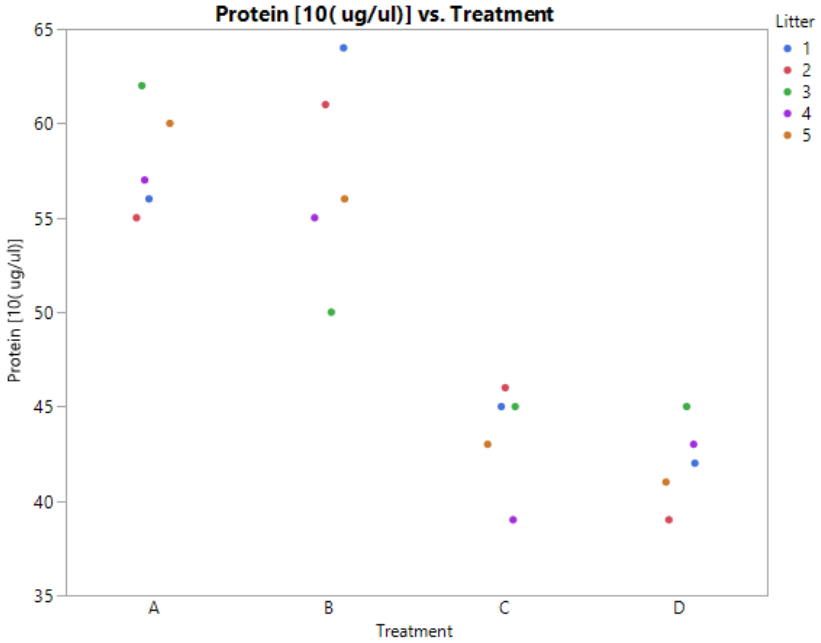
Example:

Table 3.18 lists the rat liver data for an experiment already described. Four sibling male rats within each of 5 litters were randomly assigned to the four diets. At the experiment's (and rat's) end, livers were evaluated for protein content. It is already assumed that there are differences among litters. The focus is on diet differences.

Table 3.18 Rat Liver Protein (10µg/µl) as a function of diets A, B, C and D

Litter	Diet			
	A	B	C	D
1	56	64	45	42
2	55	61	46	39
3	62	50	45	45
4	57	55	39	43
5	60	56	43	41

Figure 3.21 Protein Amount in Rat Livers Due to Diet Differences



Of course, we plot the data and look at the plot. Nothing seems too far out of line. The ANOVA summary in Table 3.19 lists Treatments, Litters and Error as sources of variation. Notice that the F-ratio is only computed for Treatments, and it has a very low probability. "Litters" in this case, is a random effect. It is not being measured so much as it is being used to rid the data of a source of variation so that Treatment differences, if they exist, may be exposed more clearly. The important F-ratio under this circumstance is that of the mean square for Treatments to the mean square for Error. Further, Error here is the failure of the differences among Treatments to be the

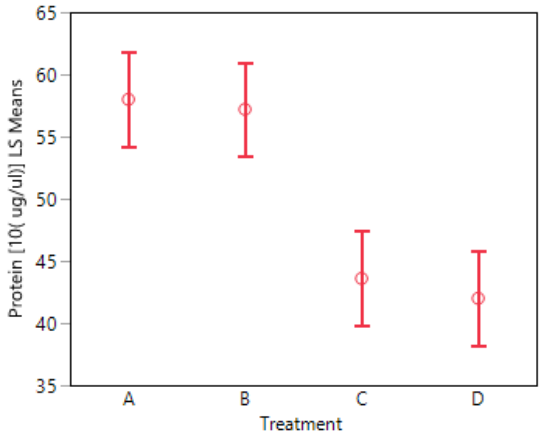
same from Litter to Litter. It measures how the Treatments relate, Litter upon Litter. The appearance of Litters in the design adds credibility to the conclusions in that the same results appear and are inferred to appear across Litters.

Table 3.19 ANOVA for Data in Table 3.18

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Total	19	1307.2			
Treatment	3	1103.2	367.7	24.2	<.0001
Litter	4	21.7	5.4		
Error	12	182.3	15.2		

Figure 3.22 is a plot following on the ANOVA to display means surrounded by HSD intervals. It shows higher protein levels for Treatments A and B, than for C and D.

Figure 3.22 Means surrounded by HSD intervals showing treatment differences



3.5.4.3 Nested Designs

Speaking of rats, the term "nested" came from experiments in which rats were literally nested within cages. Just as "regression" was generalized from an initial application examining reversion to an original form, "nested" has been generalized to refer to hierarchical arrays, e.g., rats within cages within rooms within treatments within … You get the general idea.

Applications are many and varied. Common examples may be seen in Measurement Systems Analysis (MSA, Section 4) where investigators determine the adequacy of an analytical measurement to aid in the assurance that a product conforms to specifications. This is a most important step – the M in DMAIC – in many Six Sigma projects. If the measurement process is not right, nothing downstream will be.

Nested designs are also combined with factorial designs for such applications as interlaboratory studies, but more about that later.

Often, the critical outcome of a nested design application is the generation of knowledge about the inherent variation of the process at hand. Careful data analysis includes the estimation of variance components which are the building blocks of variation. Combinations of variance components can be configured to reflect alternative future sampling schemes resulting in improved measurement systems.

Example:

Table 3.20 tabulates data from a study of the percentage of active ingredients in shipments of cleansers being received by a manufacturing facility. Two random samples (S1 and S2) from each of two random bags from each of six random lots were analyzed in duplicate (A1 and A2). Knowledge of the results of the first analysis (A1) was not available prior to the second analysis (A2).

Table 3.20 Nested Design – Active Ingredients (%) in Shipments of Cleanser

	Lot 1					Lot 2					Lot 3			
	Bag 1		Bag 2			Bag 1		Bag 2			Bag 1		Bag 2	
	S1	S2	S1	S2		S1	S2	S1	S2		S1	S2	S1	S2
A1	29	28	29	27	A1	29	27	26	24	A1	32	29	25	30
A2	29	27	29	28	A2	28	28	24	25	A2	30	30	27	31

	Lot 4					Lot 5					Lot 6			
	Bag 1		Bag 2			Bag 1		Bag 2			Bag 1		Bag 2	
	S1	S2	S1	S2		S1	S2	S1	S2		S1	S2	S1	S2
A1	29	30	28	30	A1	30	27	25	26	A1	29	31	29	29
A2	29	31	28	28	A2	29	28	28	26	A2	31	32	30	31

A plot of these data, Figure 3.23, shows some points of concern. Is the variation among duplicate analyses uniform? Why are Lot 2, Bag 2 active percentages low? Could this have happened by chance? Some light is shed on these questions and more, by a simple control chart as shown in Figure 3.24. These charts are available in most modern statistical software packages. The use of these and other forms of data plots is encouraged.

It turns out that analysis variation may be large, but it cannot be said to differ among the samples in this set of data. Bag variation within a lot is a different story.

Figure 3.23 Raw Data From Table 3.20 – Active Ingredients in Cleansers

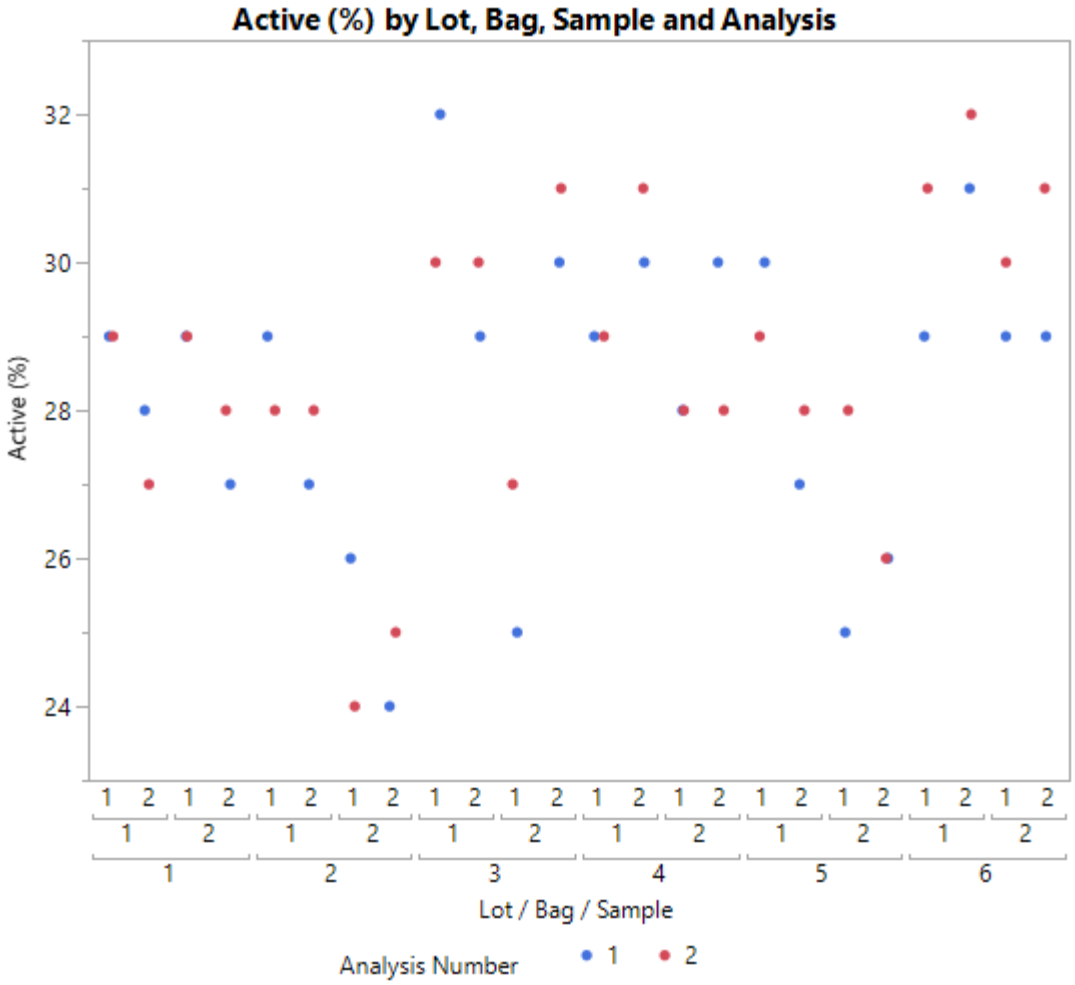


Figure 3.24 Standard Deviations of Duplicate Analyses

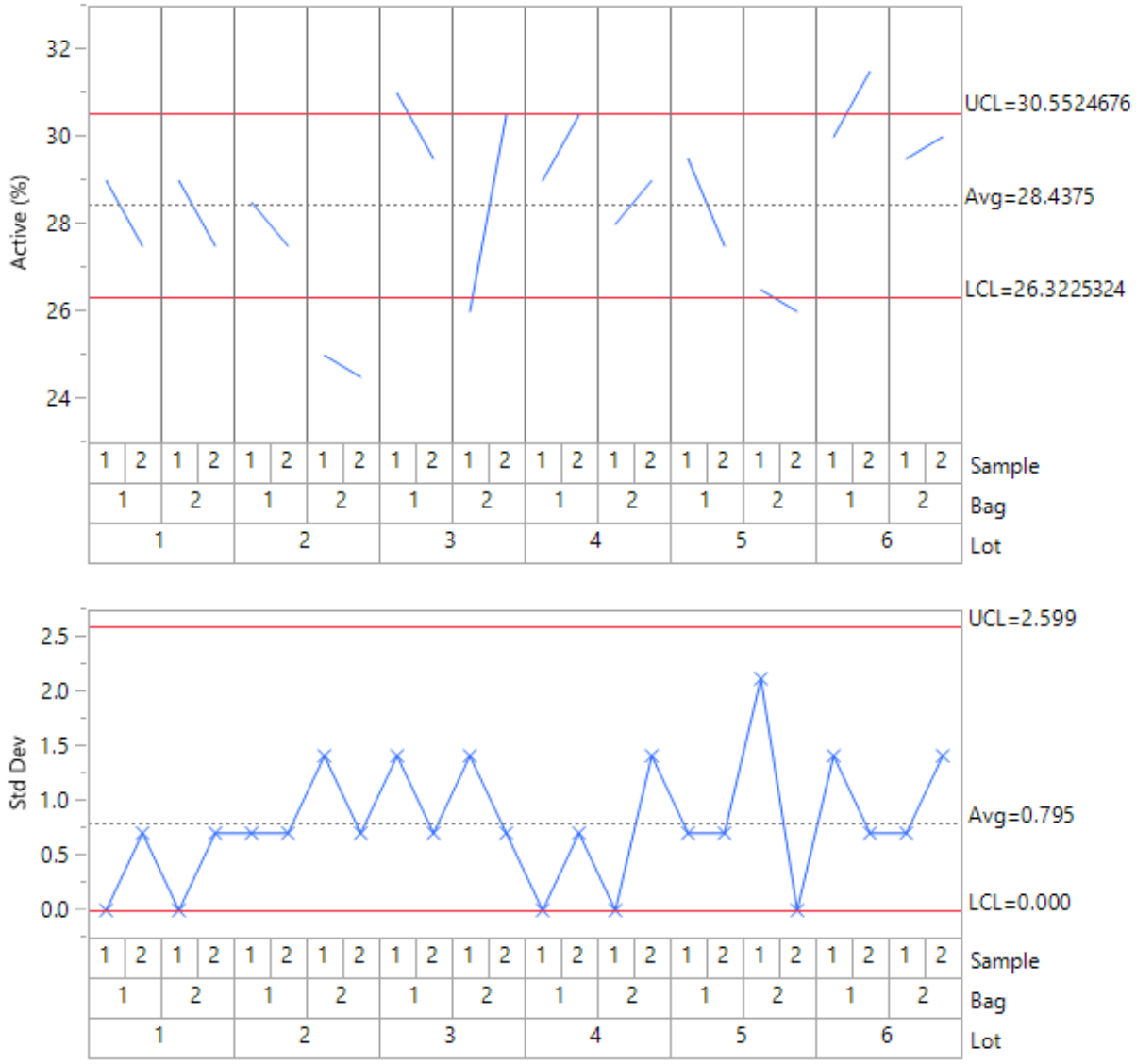


Table 3.21 shows the ANOVA with the variance components listed in the extreme right column.

Table 3. 21 Active Ingredients ANOVA with Variance Components

Source	DF	SS	MS	Var. Comp.
Total	47	183.81		
Lot	5	78.69	15.74	1.04
Bag	6	44.38	7.40	1.05
Sample	12	38.25	3.19	1.13
Analyses	24	22.50	0.94	0.94

No single component of variance stands out as being a single major source of variation, although if you were the plant manager, you might want to sit down to tea with the vendor to discuss Bag 2 of Lot 2. But if you hold to the assumption that all sources of variation in this study are completely random, you can estimate the variation likely to be seen from a single analysis of a random lot, bag and sample. It is:

$$s = \sqrt{1.04 + 1.05 + 1.13 + 0.94} = 2.04.$$

This suggests a broad uncertainty for the prediction of the vendor production. A 95% confidence interval is approximately plus or minus 4% around the result. Not a pretty sight.

Suppose, instead that the inference is being made to a specific lot, like today's. And let us suppose also that, just to be sure, we sample 2 bags by 2 samples each and carry out 2 analyses on each sample. In this case, the standard deviation becomes:

$$s = \sqrt{\frac{1.05}{2} + \frac{1.13}{4} + \frac{0.94}{8}} = 0.96$$

The 95% confidence interval about the observed mean is plus or minus 1.9%.

Much more can be said about the use of variance components, and much more can be gleaned from the use of nested designs. For more information see Box, Hunter and Hunter (2005) and Montgomery (2013).

3.5.4.4 Mixed Crossed and Nested Designs

Combining the message of previous sections where multi-way experiments are discussed, with the nested designs of Section 3.5.3 introduces one of the next levels of complexity. This is the broad class of designs with both mixed and nested factors. They are experienced in Lean Six Sigma studies as part of the Gauge R&R Study suite, and they are very useful in interlaboratory studies, often referred to as Round Robins.

An example scenario has multiple laboratories with different brands or types of instrument measuring the same response and with different operators on the same or different brands of equipment. From a central source, the laboratories receive multiple samples, some or all in blind replicate; all craftily coded and randomized so analysts cannot carry over information from predecessor to successor sample results.

Applications of these designs abound, and there are no limits, except those practical and logistical, to the number of crossed and nested factors. During the analysis of the resulting data, the random and fixed testing rules of the previous section hold.

Example:

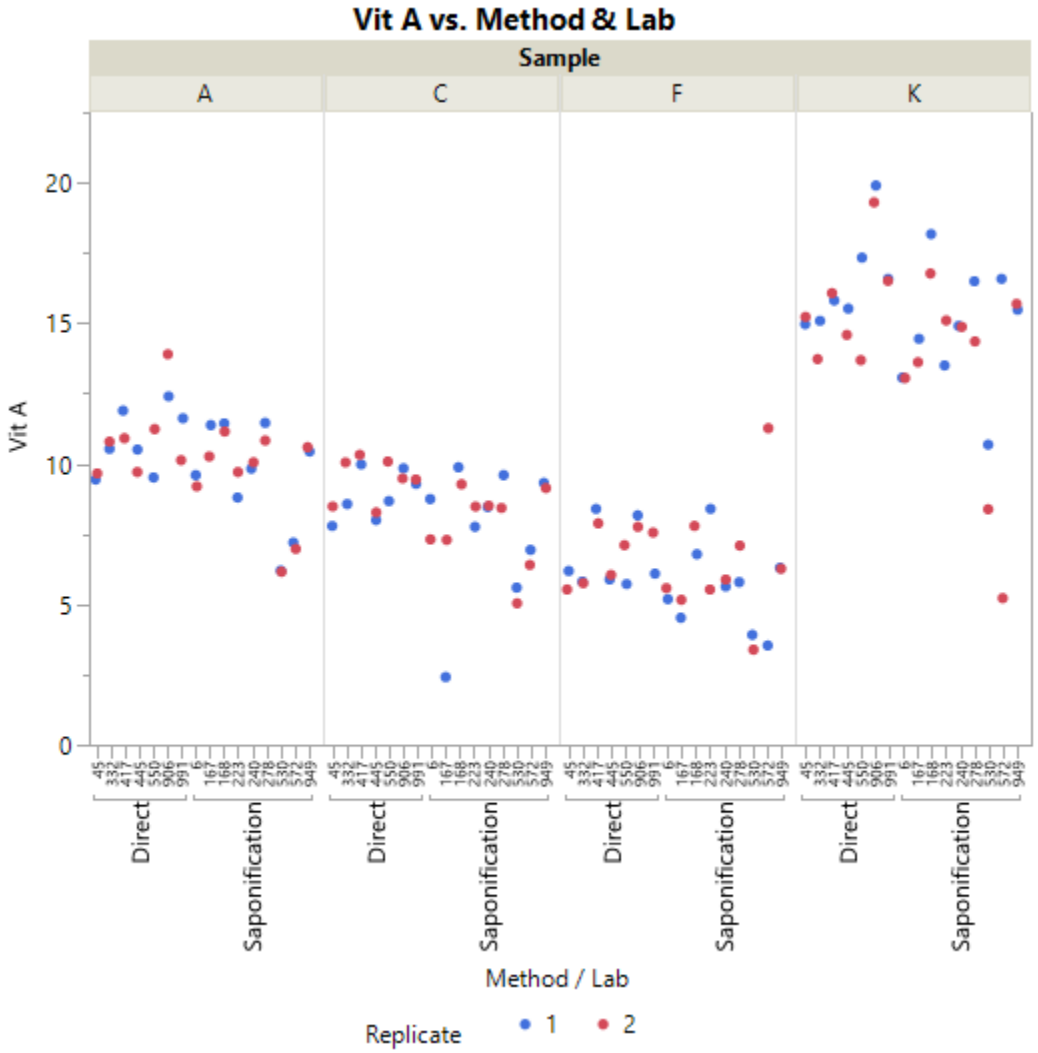
Multiple government laboratories may test food samples to aid in verification of claimed levels of nutrients, Vitamin A among them. Some of the laboratories use a direct extraction method while others use saponification. In this sense, the laboratories are nested within the methods. The measure of Vitamin A is reported in International Units per Pound (IU/Lb.). Table 3.22 lists data generated from a Round Robin study. IU values have been divided by 1000.

Table 3.22 Round Robin Study of Lab and Method Differences in the Assessment of Vitamin A

		Sample A		Sample C		Sample F		Sample K	
Method	Lab	Rep. 1	Rep. 2	Rep. 1	Rep. 2	Rep. 1	Rep. 2	Rep. 1	Rep. 2
Direct	45	9.44	9.66	7.79	8.49	6.19	5.53	14.97	15.23
	332	10.53	10.79	8.57	10.05	5.81	5.76	15.09	13.73
	417	11.89	10.91	9.98	10.32	8.40	7.89	15.81	16.07
	445	10.51	9.71	8.00	8.27	5.89	6.05	15.52	14.59
	550	9.52	11.24	8.68	10.08	5.73	7.11	17.33	13.69
	906	12.40	13.90	9.84	9.48	8.17	7.76	19.90	19.30
	991	11.62	10.13	9.29	9.44	6.09	7.56	16.58	16.51
Saponification	6	9.60	9.20	8.75	7.31	5.19	5.58	13.07	13.05
	167	11.38	10.26	2.42	7.30	4.53	5.16	14.45	13.62
	168	11.44	11.15	9.88	9.27	6.79	7.79	18.17	16.77
	223	8.80	9.71	7.76	8.48	8.40	5.53	13.50	15.10
	240	9.83	10.05	8.46	8.51	5.64	5.88	14.91	14.87
	278	11.46	10.83	9.60	8.43	5.79	7.09	16.50	14.36
	530	6.21	6.16	5.60	5.04	3.92	3.39	10.68	8.39
	572	7.20	6.98	6.94	6.41	3.54	11.27	16.58	5.22
	949	10.44	10.59	9.33	9.14	6.30	6.27	15.48	15.69

Of course, the first step in the analysis of these data is a plot of them, Figure 3.25.

Figure 3.25 Raw Data from a Round Robin Study



While carefully examining this graph, you might wonder about Lab 572 in the Saponification set. The distance between the replicate observations seems extreme by comparison to those of the other labs. A useful tool for learning about the uniform variation, or lack of it, is the standard deviation chart. See Figure 3.26, a and b.

Figure 3.26a Standard Deviation Chart for Vitamin A, Method=Direct

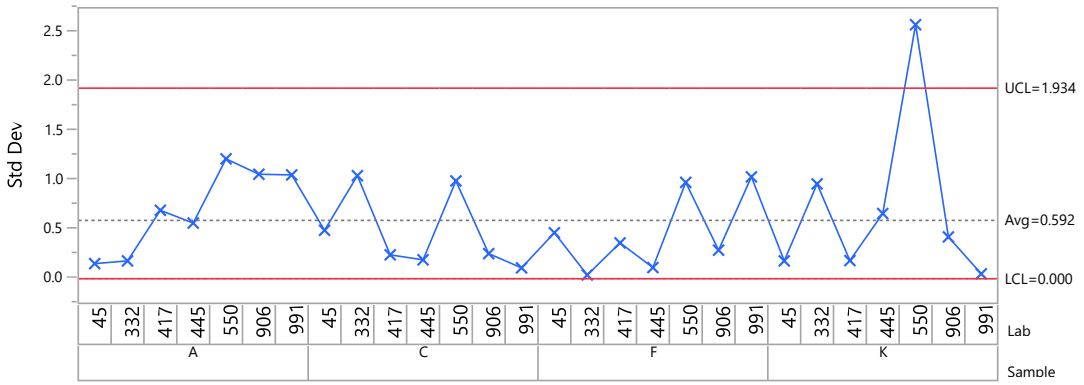
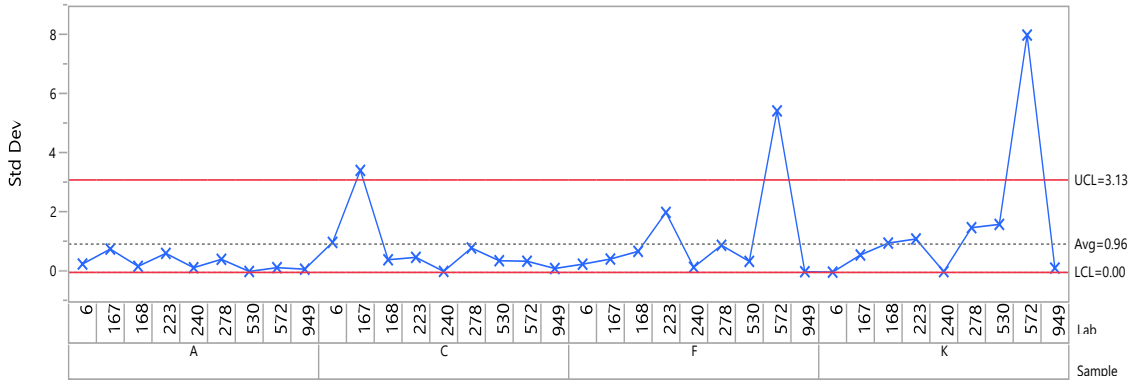


Figure 3.26b Standard Deviation Chart for Vitamin A, Method=Saponification



It would appear that Lab 550 in the Direct Extraction Method group had disagreeing duplicates on sample K, and clearly, Lab 572 in the Saponification Method group had difficulty reproducing its results for Samples F and K. Another Lab may be cause for concern: Lab 167 in the Saponification Method set. How should those discrepancies be handled?

In most cases, statisticians have absolutely no business eliminating data. (See Section 3 of this chapter.) This is the purview of the subject matter expert. In the specific case of a Round Robin study like this one, the responsible investigators acting on the evidence of Figures 3.6 a and b would contact the lab managers involved to point out the extreme differences between replicates and, working together, would seek causes. Discovery of the source might involve wading through records for accuracy or retraining technicians or both. In the meantime, the analysis would continue, with the difficult data set aside, at least temporarily.

With the data that remain, we can proceed with the analysis of variance, remembering to count degrees of freedom to assure that your software is doing what you think it is – two methods, so 1 degree of freedom; six labs within the Extraction Method, so 5 degrees of freedom there, and seven labs within the Saponification Method, so 6 degrees of freedom there, all for a total of 11 degrees of freedom for Labs within Methods, etc.

Before proceeding to the ANOVA, it is important to check that the variation among duplicates does not differ between methods. The error mean squares corresponding to the direct and the saponification methods are 0.32 and 0.57, with 24 and 28 degrees of freedom, respectively. Their F-ratio is 0.56 and is not significant at any probability level large enough to cause concern.

The full ANOVA is shown in Table 3.23.

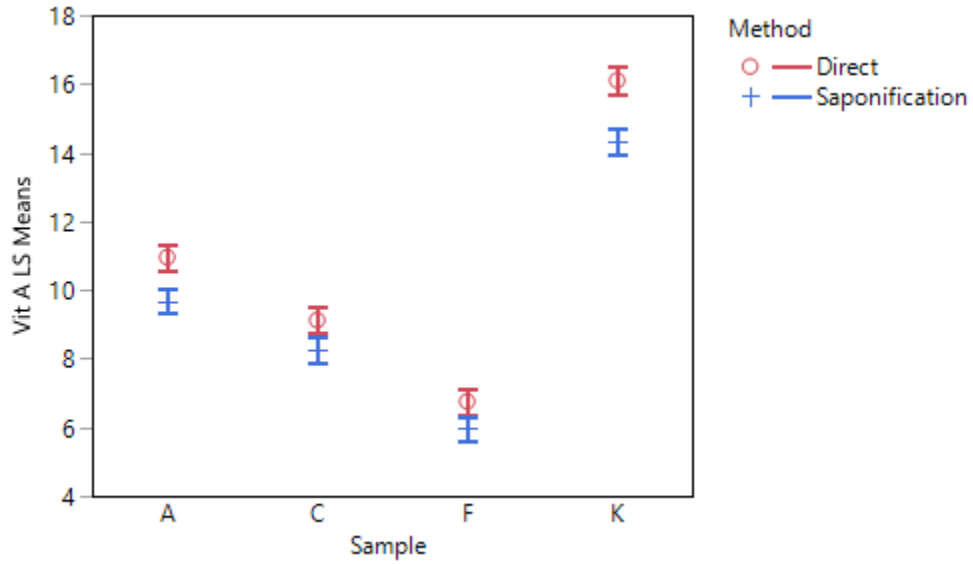
Table 3.23 ANOVA for the Data of Table 3.22

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Total	103	1371.48			
Method	1	36.11	36.11	79.03	<.0001
Lab[Method]	11	186.54	16.96	37.12	<.0001
Sample	3	1089.54	363.18	794.89	<.0001
Sample*Method	3	4.02	1.34	2.94	0.04
Lab*Sample[Method]	33	35.02	1.06	2.32	0.00
Error	52	23.76	0.46		

If the samples had been a random selection, in the ANOVA table, the test for Methods would involve the ratio of the Methods mean square to the Sample*Method mean square. In this case all factors are fixed. Their F-ratios are calculated with the error mean square in the denominator. The error mean square is the variance of the duplicates. These ratios suggest that all factors are statistically significant. What do we make of that? Does it seem likely? A possible interpretation is that the laboratory technicians are cleverer than the Round Robin designers thought. This may not have been their first rodeo, the exception being the green horn in Lab 572. Did the others see this test coming and adjust results accordingly?

Notice that the largest mean squares are those representing Methods and Methods within Labs. They might be most deserving of attention. Figure 3.27 shows method means by sample. On average, the Direct Injection method provides results that are approximately 1200 IU/Lb. higher on average than the Saponification method. This must be cause for concern.

Figure 3.27 Method Means by Sample Surrounded by Tukey's HSD Intervals



In addition, both methods show Lab-by-Sample interactions. Figure 3.28a shows the interaction for the Direct Injection method, and Figure 3.28b shows it for the Saponification method. The interpretations are a bit involved, especially if we believe that the variation among duplicate analyses is really as low as calculated. Still, if you look at the figures, you will notice that lab-to-lab variation increases with the sample mean in both methods and each method shows one lab that is usually in disagreement with the others.

Figure 3.28a – Direct Extraction Method

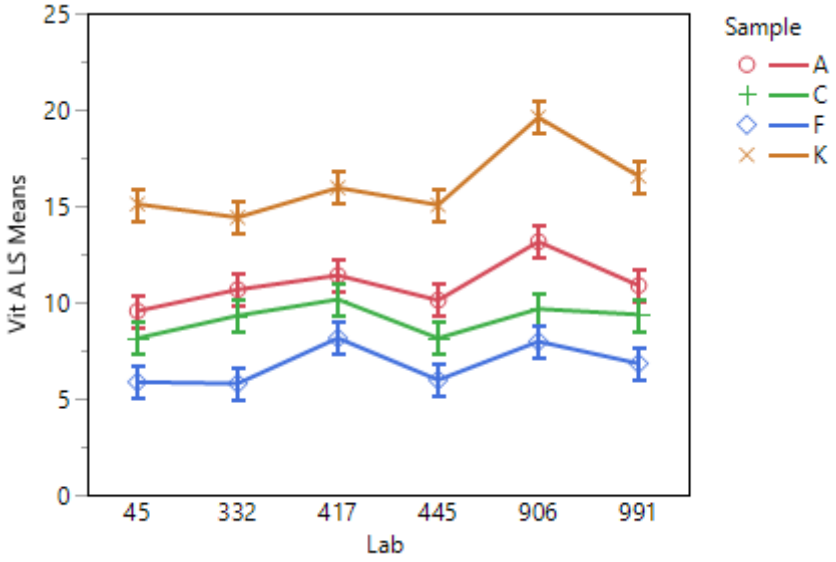
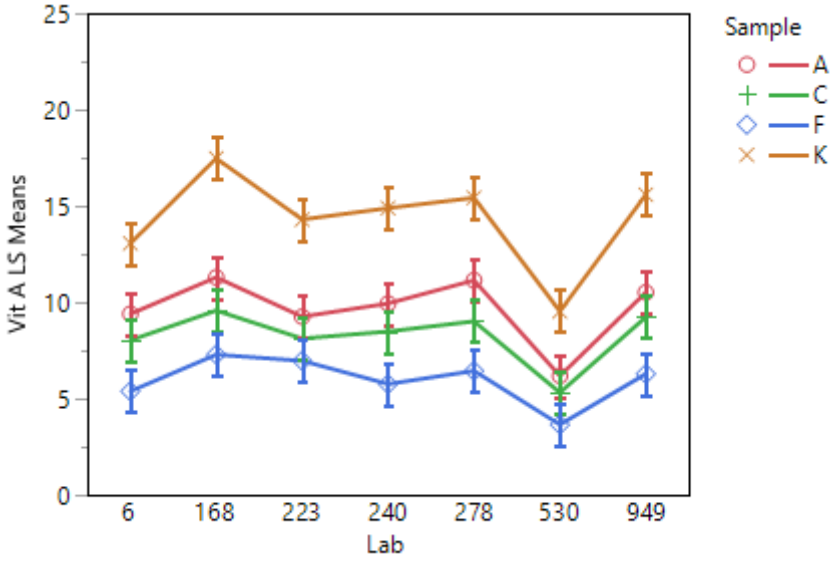


Figure 3.28b – Saponification Method



This one design result has not answered all questions. It was never intended to, but it does light the path for much more work to be done to obtain laboratory and method agreement. The use of this mixed, crossed and nested design, like any other beginning design is only a start of serious investigation.

3.5.4.5 Factorial designs and their fractions

When we left Mary at the close of Section 3.5.3, she was laboring through her boss's "economizing" of her two-level experiments by adding more factors through the reduction of the number of replicate observations for each experimental treatment combination. While this may have seemed frustrating because, at first glance it appeared to rob the experiment of needed power to detect differences, it pointed to a highly successful strategy of experimentation. Recall that it changed the question from "Do any of these factors make a difference?" to "Which factors, singly or in combination, are most important for process improvement?"

Presently, there are 4 factors, each at 2 levels for a total of 16 experimental treatment combinations. The design is fully saturated, meaning that all the experimental points are used up, and there is no room (replication) to estimate experimental variation directly. Would it be possible to add another factor without increasing the total number of experimental treatment combinations?

That question was answered by D.J. Finney, a student of Fisher (Finney, 1947). He showed how to fractionate 2- and 3-level designs, and he showed how to calculate what information was lost in doing so.

Following his lead, suppose a fifth factor, Cooling Rate, were added to Mary's experiment. Then, all factors considered, we have:

A. Time
B. Temperature
C. Agitation Rate
D. Oil
E. Cooling Rate

If each were held at 2 levels there would be 32 experimental treatment combinations. Can we reduce that? The answer is yes, but there may be some loss of information.

To see how fractionation is accomplished, we can examine a simpler case. Suppose there were only three factors, A, B and C, each at two levels. Then, there are 8 experimental treatment combinations. Code the low levels of each factor using "-1" and the high level using "+1." The resulting design is called a 2^3 (two cubed) factorial design, meaning there are 3 factors, each at 2 levels. Its design "matrix" is shown in Table 3.24. See Figure 3.29 for the cube.

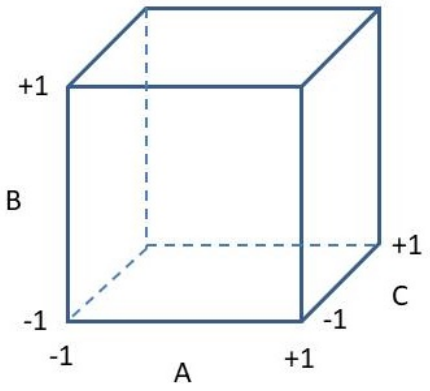
Table 3.24 Design matrix for a 2^3 factorial design

Treatment	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	1	1	1	-1
2	1	-1	-1	-1	-1	1	1
3	-1	1	-1	-1	1	-1	1
4	1	1	-1	1	-1	-1	-1
5	-1	-1	1	1	-1	-1	1
6	1	-1	1	-1	1	-1	-1
7	-1	1	1	-1	-1	1	-1
8	1	1	1	1	1	1	1

For each of the 8 experimental treatment combinations, the -1 and 1 settings specify the combinations of factors A, B and C. Their paired products AB, AC and BC are simply the products of the factors going into the pairs. So, for Treatment 1, the AB interaction is (-1)(-1) = 1; for Treatment 2, it is (1)(-1) = -1, etc. (Interactions are described in Section 6.3.)

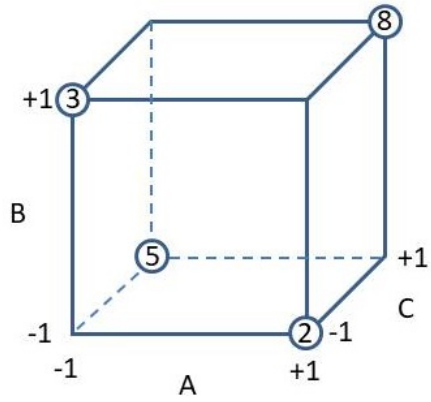
In nature, main effects are more likely to occur than 2-factor interactions, and 2-factor interactions are more likely to occur than 3-factor interactions. For a 3-factor interaction to exist, it must be that the nature of the 2-factor interaction involving A and B at the low level of C is different from its nature at the high level of C. Of course, that can happen, but it happens rarely. If, in the course of designing an experiment, you must sacrifice the estimation of an effect, you would do well to give up the quantification of a 3-factor or higher level interaction. After all, its existence is a low probability event.

Figure 3.29 A 2^3 Factorial Design



To fractionate this design, you might take only the shaded treatments of Table 3.24. These are the treatment combinations whose 3-factor interaction is calculated to be 1. (See Figure 3.30 for the half replicate.) If you were to do this, then the three factors, A, B and C, would each be balanced – they would each have the same number of -1s and 1s. Moreover, they would still be independent of each other. Knowledge of the effect of one would impart no knowledge of the effect of another. That is the good news. The bad news is that the effects of A, B and C would be confounded (or "hopelessly confused," not damned in the Biblical sense) with the interactions; A with BC, B with AC, and C with AB. This is not what most would consider a satisfactory design.

Figure 3.30 A half replicate of a 2^3 design. Numbered vertices correspond to the shaded rows of Table 3.24.



Still, this example shows how to fractionate a design. Incidentally, you could also choose the unshaded rows of Table 3.24, which is the other half-replicate of the design.

Designs of this type form a general class of what are called 2^{k-p} factorial designs, where k represents the number of factors and p stands for the degree of the fraction; 1 for a half replicate,

2 for a quarter replication, etc. There is such a thing as a quarter replicate. For more on this class of designs, see, Box and Hunter (1961).

A nagging question is how might you know what factors are confounded with what other factors without writing out the entire design matrix as is shown in Table 3.24? Most modern software packages do the math and tell you about the confounding. Still, it is useful to know so you are sure of what you are getting when you use the software. Table 3.24, the design matrix, is made up of vectors in columns. Squaring the elements of any of these vectors yields a column of 1s. We call that the identify vector, **I**. (Vectors always appear in bold type. Nobody knows why.) If we choose to fractionate on the 3-factor interaction, we are essentially setting up a defining contrast **I** = **ABC** and if you want to know what **A** is confounded with you multiply both sides of that equation by **A**. **AI** = **AABC**, which is the same as **A** = **BC** because anything times the identity remains the original anything and because **AA** = **I**.

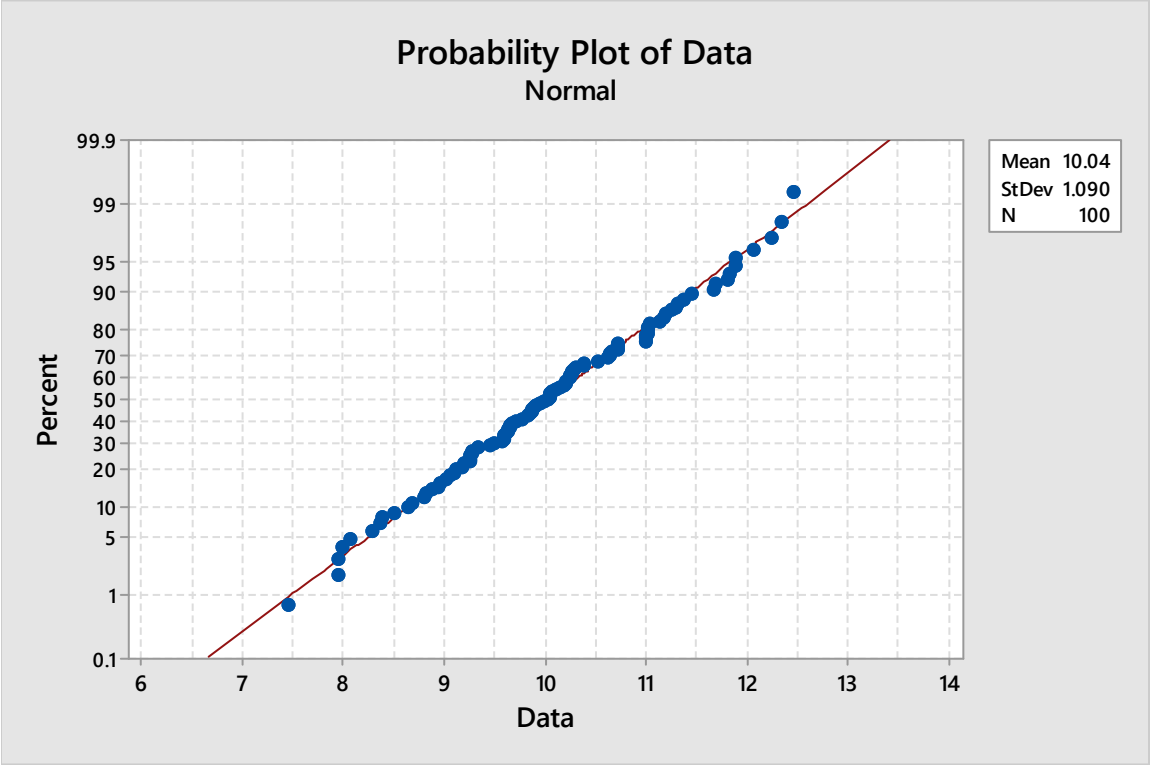
Remember Mary? How does any of this help her? Mary has factors A through E. If she wants to run the half replicate of the 2^5 factorial, a 2^{5-1} design, she might choose **I** = **ABCDE**. This indicates the full confounding pattern. The factor A has an alias. It is not Joey Baga Donuts, it is BCDE. B's alias is ACDE, etc. The defining contrast tells you that for Mary's fraction, all main effects are confounded with 4-factor interactions. That is comforting because the probability that a 4-factor interaction exists is usually low. Two-factor interactions are confounded with 3-factor interactions. The safer bet is on the 2-factor interaction.

Mary could delight the boss with her new found technology. She can add one factor without increasing the number of experimental treatment combinations. The risks in doing so are minimal.

There was only one problem. There is no built-in replication, so there is no error term against which to test the factor mean squares. It turns out that you can look at the distribution of the effects. The effect of a factor is the average response at the high level of the factor minus the average response at the low level of the factor. Effects are averages, so if there is nothing to disturb them, they should be normally distributed.

A good visual test is the normal probability plot as in Figure 3.31. It shows 100 observations taken from a distribution whose mean is 10 and whose standard deviation is 1. The vertical axis is transformed from linearity purposely, so the data form a straight line if they come from a normal distribution. A departure from normality should stand out visually.

Figure 3.31 A normal probability plot



Example:

Jack, a food scientist, learned from Consumer Affairs that consumers are complaining about chicken noodle soup. It is strange because half the complaints are that the soup is too salty, and the other half are that it is not salty enough. Jack thinks the problem may be due to the dosing; that is, the "paste plug" which is a viscous mixture of protein and spices, lacks density uniformity. This causes paste plug weights to vary highly, meaning that some pouches of this dry mix soup are heavier and therefore more salty than others. Quality records confirm that the package weights are out of both upper and lower specification limits by a substantial margin. Jack would like to have known that earlier.

Jill, the plant manager is flummoxed. She thinks that anything she does to reign in the plug weight distribution will slow the process. She gets rewarded for high productivity, but she is a good company citizen, and she understands the value of keeping the consumers happy.

At their summit, they discuss factors that may influence the plug weight distribution and come up with the following list:

A. The number of mixer ports through which the oil component is added during mixing. 1 or 3 ports. One port is easier and saves manufacturing time.
B. The mixer temperature, roughly controlled by using ambient temperature or running cold water through the mixer jacket. Cold water costs more.

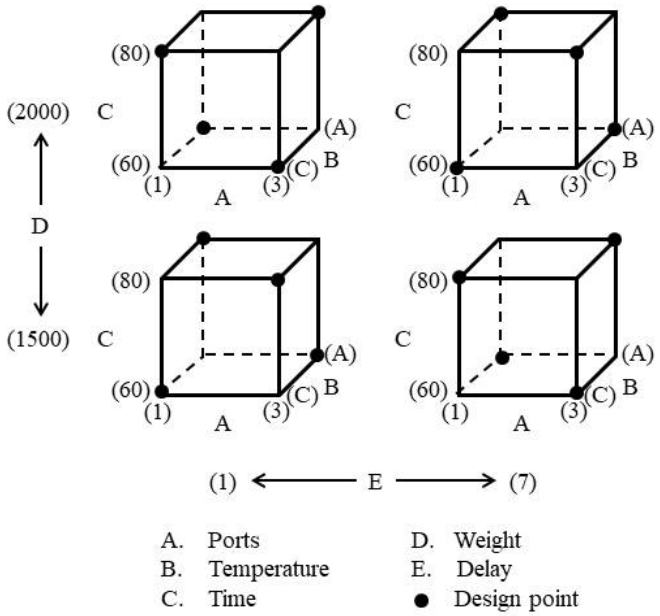
C. The mixing time in seconds: Shorter mixing time means greater productivity. 60 or 80 seconds.
D. The batch size. The mixer manufacturer recommends a maximum of 1500 pounds, but Manufacturing pushed for 2000 pounds to increase productivity.
E. The delay in days between mixing and packaging. Research said the product should have a week to "set up," but Manufacturing wanted to pack ASAP after mixing. Delay was studied at 1 and 7 days.

Jill was forced by the manufacturing schedule to limit time for experimentation to 16 runs, and that was a stretch. Jack came up with a half replicate of a 2^5 factorial design based on information he learned in a company short course. Together, they organized operations staffed jointly by Research and Manufacturing, and they produced the data in Table 3.25 following the schematic diagram of Figure 3.32.

Table 3.25 Variation in paste plug weights from 2^{5-1} design

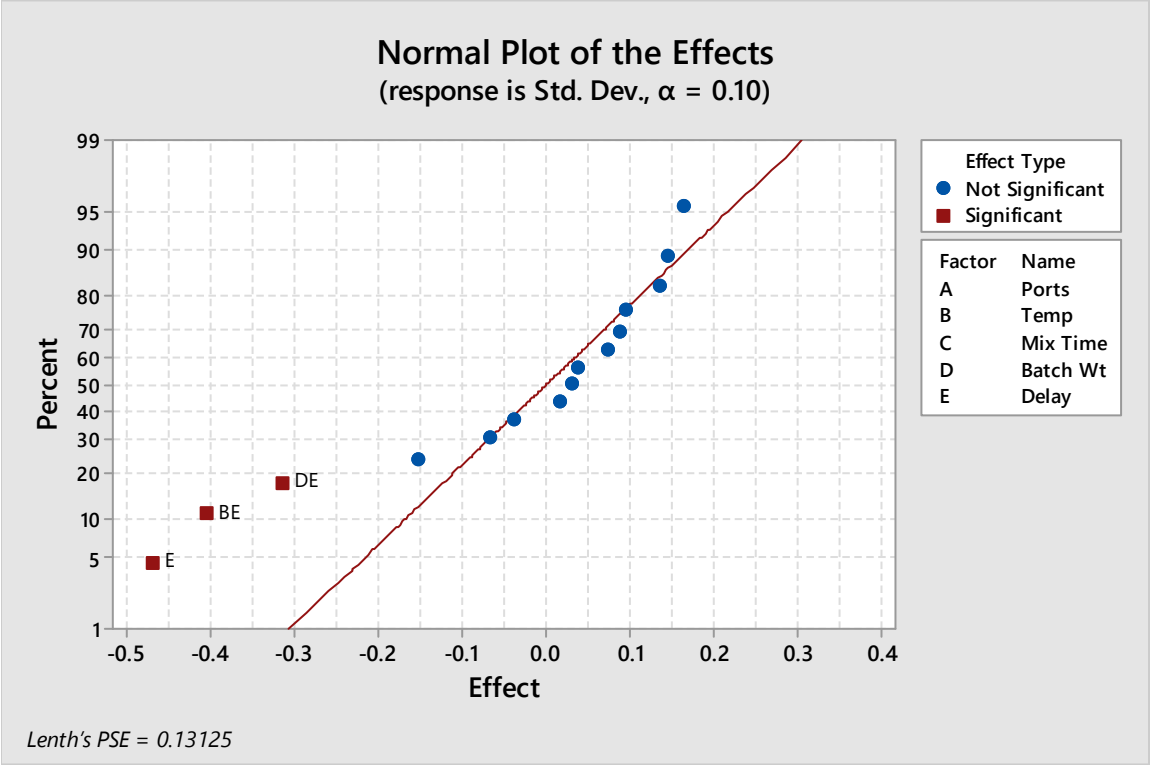
Ports	Temperature	Mix Time (Min.)	Batch Wt. (Lbs.)	Delay (Days)	Std. Dev.(g)
1	Cold	60	2000	7	0.78
3	Cold	80	2000	7	1.10
3	Ambient	60	1500	1	1.70
3	Cold	80	1500	1	1.28
1	Ambient	60	1500	7	0.97
1	Cold	80	1500	7	1.47
1	Ambient	60	2000	1	1.85
3	Ambient	80	2000	1	2.10
1	Ambient	80	2000	7	0.76
3	Ambient	60	2000	7	0.62
1	Cold	80	2000	1	1.09
1	Cold	60	1500	1	1.13
3	Cold	60	1500	7	1.25
3	Ambient	80	1500	7	0.98
3	Cold	60	2000	1	1.36
1	Ambient	80	1500	1	1.18

Figure 3.32 Schematic diagram of the paste plug weight variation design



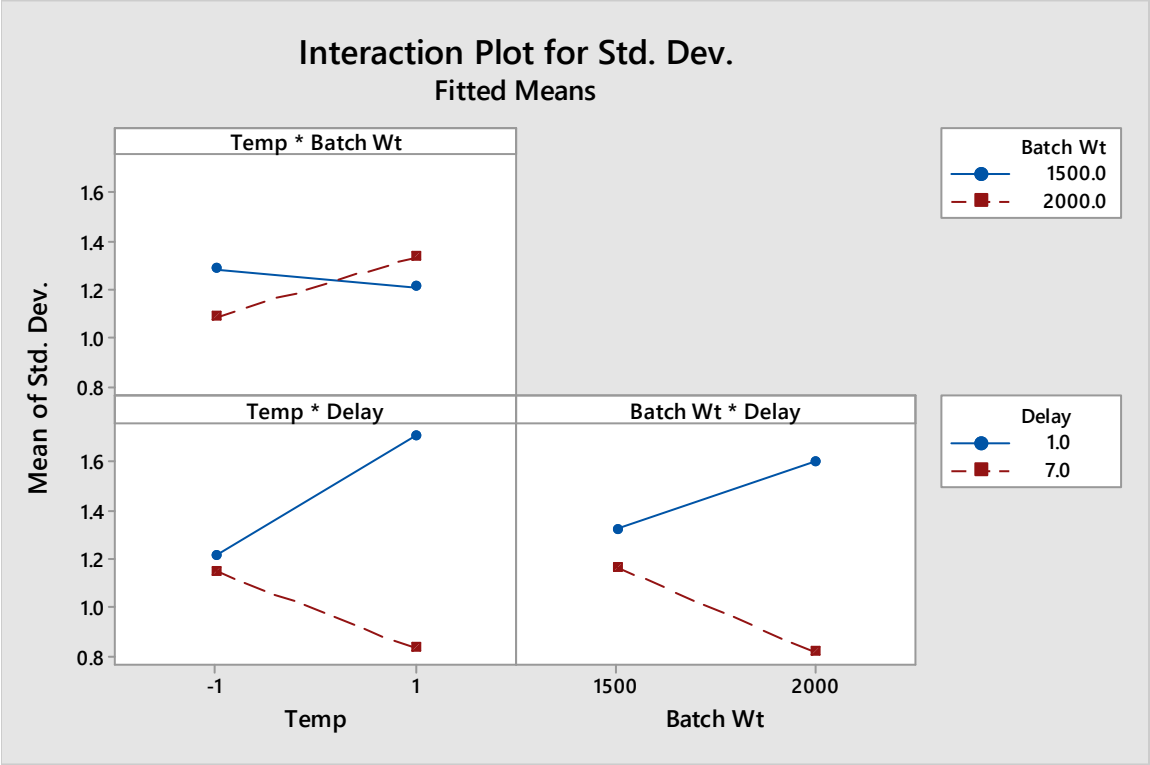
Of course, it took some time to consolidate all the data, but Jack carried out the analysis and came up with the probability plot of the effects in Figure 3.33.

Figure 3.33 Effect plot of paste plug variation data



Two interactions emerged as being of value. As an aside, notice that a lenient error probability of $\alpha = 0.10$ is used. As this is an initial design being used to explore the factors and their possible interactions for improvement opportunity, it is best to allow for effects to emerge. Verification studies may be conducted later. Given that, there is likely value in exploring the Temperature-by-Delay and the Batch Weight-by-Delay interactions as shown in Figure 3.34.

Figure 3.34 Interactions in the paste plug weight variation data



A major conclusion is that the delay works jointly with the temperature ("-1" means Cold and "1" means Ambient) and with the batch weight to indicate compromises that might maintain productivity and improve customer satisfaction. To minimize plug weight variation, delay packaging for a week to let the product set up and then use ambient temperature water in the mixer jacket. If the product must be used immediately after mixing, use cold water in the mixer jacket.

The second conclusion involves the interaction of delay and batch weight. If Manufacturing can delay 7 days between mixing and packaging, they can take advantage of the economies of larger batch sizes. If not, smaller batch sizes should be used.

As discussed previously, a single design rarely stands alone to find the path to improvement. Questions remain: Would paste plug weights be within specification if the delay between mixing and packaging were 3 days? What would happen if Manufacturing used intermediate weight, say 1750 Lb. batches, in the mixer? Would it help to control the mixer jacket water temperature to levels between simply cold and ambient? (Hare, 1988)

This example ends here, but it should be understood that it is rare when the results of a design stand alone. That is not usually intended. Instead, the design strategy is meant to continue, as in Figure 3.15 to build knowledge.

Before leaving this section on factorial designs and their fractions the reader should be aware of some other facts:

- Fractional factorial designs have been classified into "Resolutions." (Box, Hunter and Hunter, 2005).
    - Designs of Resolution III are those whose main effects are free of confounding with each other but are confound with 2-factor interactions.
    - Designs of Resolution IV have main effects free of confounding with each other or any 2-factor interactions. The main effects, however, are confounded with 3-factor interactions, and some 2-factor interactions are confounded with other 2-factor interactions.
    - Designs of Resolution V have no main effects or 2-factor interactions confounded with each other. Main effects are confounded with 4-factor interactions, and 2-factor interactions are confounded with 3-factor interactions.
- Depending on the design strategy, it is sometimes wise to allow some confounding, especially in situations where broad screening is the goal. This recognizes that further experimentation will pursue identification and quantification of interactions. Fractional factorial designs can be used for screening and for characterization, i.e., identification of meaningful interactions, depending on their use.
- The notion of fractionation can be used to divide a factorial design into blocks with, for example, each half replicate of the design assigned to one of two blocks. In such cases, blocks may be confounded with other design effects.
- There may be value in some cases to include center points in 2-level factorial designs to assess the potential lack-of-fit of the model there. If the lack of fit is significant, some curvature of the response is indicated. This suggests that a more substantial design should be used.

3.5.4.6 Optimizing Designs

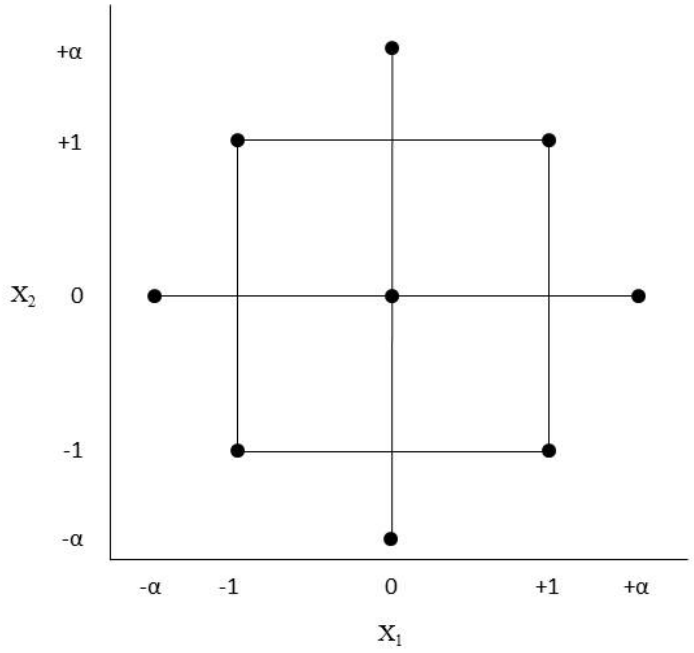
Applications of screening and characterizing designs assume the world is locally linear. That is usually a safe assumption during the early stages of experimentation. However, once key factors and their possible interactions have been discovered, the experimental world gets a little more complex. The new aim is to find the "sweet spot" or optimum response in a curved response space. Instead of models with only first order terms, or perhaps second order terms to measure interactions, we must graduate to models with additional second order, quadratic terms. A typical model might include a constant term, a term for each of the main effects, a term for each possible 2-factor interaction and a squared term for each main effect. In more complex situations, even higher order terms may be needed to explain the data.

Which designs might be best for these situations? To begin, it would seem that at least three levels of each factor must be needed. With two factors, there would be 9 experimental treatment combinations. That might seem reasonable, but with three factors, there would be 27 experimental treatment combinations. That high number could tax resources, and it would be even more taxing to follow this logic as the number of factors increases. Is there a better way?

Of course. Figure 3.35 is a sketch of a 2-factor central composite design. It shows 9 unique experimental treatment combinations, but they are not arrayed as one might in a 3-by-3 factorial

design. Instead, there are 5 levels of each of the 2 factors, better serving the experimenter's desire to assess response curvature. It is built on a 2^2 factorial design which is augmented by "star" points whose distance from the design center is alpha (α). The setting of α depends on a design characteristic called rotatability (Box and Hunter, 1957). In the context of the central composite design, a good design is considered to be one in which the variance of prediction is uniform at points equidistant from the center. For rotatability to be assured, the value of α is the fourth root of the number of factorial points in the design. So, for the present design, $\alpha = \sqrt{2}$. If there were 3 factors (Figure 3.36a), the number of factorial points would be 8, and α would be 1.682.

Figure 3.35 A 2-Factor Central Composite Design



Certainly, designers are free to choose other values of α. If there is concern about the repeatability of experimental treatment combinations, for example, experimenters may choose α = 1 (Figure 3.36b). True, this choice spoils part of the curvature assessing advantage of the original central composite design (CCD), but it provides a safety check by alignment of the star point findings with other points on the same factor level. The resulting array is called a face centered cube design.

Another alternative is to use a spherical CCD (Myers and Montgomery, 2002) with $\alpha = \sqrt{k}$ where k is the number of design factors. Its design points are all on the surface of a sphere with radius α. This is the best choice if the region of interest (the design space) is a sphere.

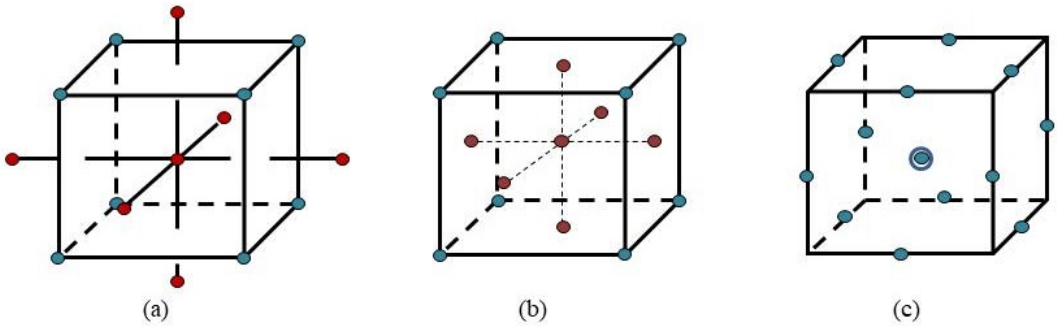
In situations where extremes are to be avoided, an alternative to the CCD is the Box-Behnken (1960) design (Figure 3.36c). Its experimental treatment combinations are located at the centers of the cube (or cuboidal) edges.

Example:

Spray drying towers are used to create crystals from tea liquors to make instant tea. Crystal size, friability, dissolution and density are all important performance factors and are believed to be functions of dryer temperature, liquor throughput and dryer air velocity. If all three of these factors were held at their high levels, there would be dryer shrapnel in the manufacturing facility and tea in the street. Box-Behnken designs were used to temper experimentation, to find best settings and to avoid disasters.

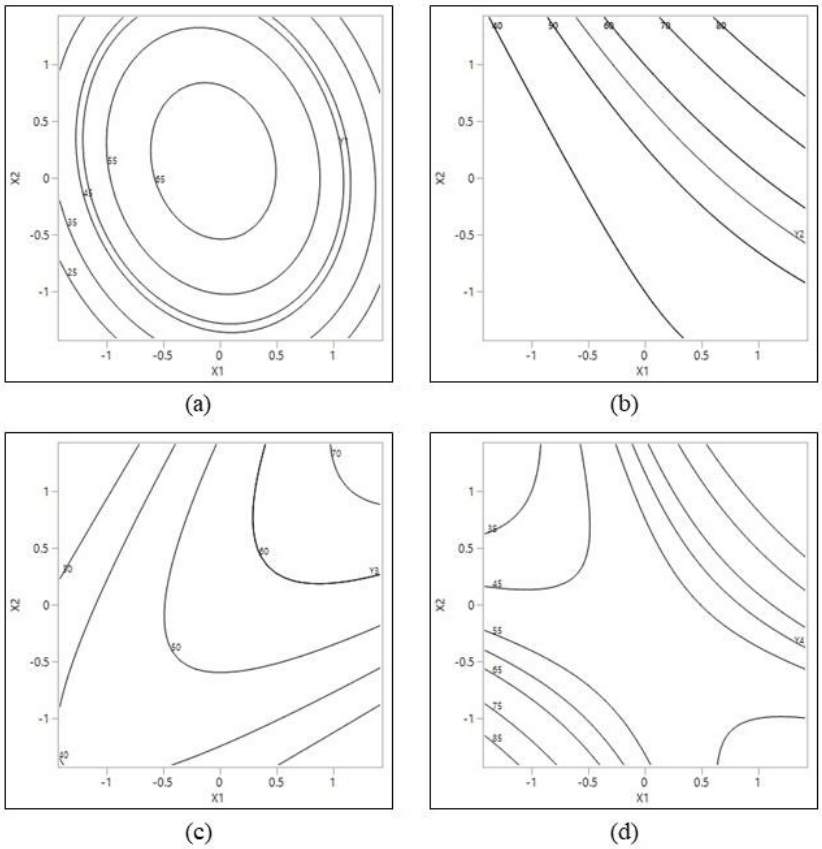
It is highly recommended that these central composite designs be carried out with 3 to 5 replicated center points, randomized along with all other points, to assure a reliable estimate of the experimental variation.

Figure 3.36 Some 3-Factor Response Surface Designs – (a) Rotatable CCD, (b) Face-Centered Cube, and (c) Box-Behnken Design



As discussed earlier, response surfaces (Box and Wilson, 1951) are essential for analyzing, interpreting and communicating findings from designed experiments, and data from optimizing designs are no exception. Second order response surface models generate four different patterns depending on the model coefficients. These are illustrated in Figure 3.37.

Figure 3.37 Response Surface Patterns, (a) Relative Maximum or Minimum, (b) Stationary Ridge, (c) Rising Ridge, (d) Saddle



Part (a) of this figure illustrates a relative maximum or minimum. Departures from the center of the inner contours indicate movement away from a stationary point. Part (b) shows a stationary ridge; the direction toward higher valued responses involves increases in either X1 or X2 or both. Part (c), the rising ridge shows that X1 and X2 should be increased together to obtain higher valued responses. And part (d), the saddle shows increased valued responses when X1 and X2 are either both high or both low. If one is high and the other is low, the response becomes lower.

In most industrial applications, multiple responses are measured. In anticipation of this, vendors of statistical software have provided features that overlay response surfaces on the same grid, enabling users to find admissible areas of in-specification operation, simultaneous optimization or trade-offs among responses for various factor settings.

Only the simplest situations involve two factors. With more factors, response surface methods can and should be used. To analyze, explore and communicate, use two of the most important factors from the model to form a grid, and then use settings of the third, fourth and beyond factors to display the response(s). Several statistical software packages supply this feature.

Example:

Fluoride retention is an important aspect of mouth rinse. It is thought to be a function of some or all of the various ingredients. A face-centered cube design in 5 factors was planned to assess their linear, squared and interactive effects on several responses including the area under the fluoride retention curve (AUC) as a measure of effectiveness.

A full 5-factor CCD would require 50 experimental treatment combinations, including a recommended 8 replicated center points. Carrying out the full design would be a resource and logistical strain. A half replicate of the full design was proposed. It involves all the star points as the full design but only the half replicate of the imbedded 2^5 factorial design. With the recommended center point replicates, there would be 32 experimental treatment combinations. But there was another stumbling block. Not all 32 formulations could be produced on the same day, and it was believed that various environmental, personnel and lurking variables could creep into the sample generation process. The use of 2 blocks with fewer experimental treatment combinations each was recommended.

The required 36 experimental treatment combinations would require two long laboratory days. Was there a way to pare this down? Against some statistical protestations, it was decided to sacrifice the center point replication. The resultant design experimental treatment combinations are listed in Table 3.26. Actually, the statistician knew that residual error could be used as an estimate of random variation against which to test factor effects in the ANOVA.

Table 3.26 Experimental Treatment Combinations for Fluoride Retention CCD

Block	Run	A:CPC	B:PG	C:Humectants	D:PEG40	E:Flavor	Y:AUC
Day 1	1	0.05	10	30	1	0.14	902.34
Day 1	2	0.05	10	16.8	0.24	0.14	78.42
Day 1	3	0.03	5	23.4	0.62	0.06	887.14
Day 1	4	0.03	5	23.4	1	0.1	863.46
Day 1	5	0	10	30	0.24	0.14	904.4
Day 1	6	0.03	0	23.4	0.62	0.1	902.08
Day 1	7	0.05	10	30	0.24	0.06	904.4
Day 1	8	0	10	16.8	0.24	0.06	902.83
Day 1	9	0.03	5	16.8	0.62	0.1	859.81
Day 1	10	0	10	16.8	1	0.14	746.38
Day 1	11	0.03	5	23.4	0.62	0.1	902.43
Day 1	12	0	10	30	1	0.06	355.91
Day 1	13	0	0	30	0.24	0.06	691.74
Day 1	14	0.05	0	30	0.24	0.14	901.48
Day 1	15	0.05	5	23.4	0.62	0.1	851.4
Day 2	16	0	0	16.8	1	0.06	802.15
Day 2	17	0.03	10	23.4	0.62	0.1	845.6
Day 2	18	0.05	0	16.8	0.24	0.06	463.73
Day 2	19	0.05	0	30	1	0.06	844.9
Day 2	20	0.03	5	30	0.62	0.1	822.81
Day 2	21	0.05	0	16.8	1	0.14	808.07
Day 2	22	0.05	10	16.8	1	0.06	41.19
Day 2	23	0	5	23.4	0.62	0.1	769.34
Day 2	24	0.03	5	23.4	0.62	0.1	839.48
Day 2	25	0.03	5	23.4	0.62	0.14	767.44
Day 2	26	0.03	5	23.4	0.62	0.1	844.7
Day 2	27	0	0	30	1	0.14	578.35
Day 2	28	0.03	5	23.4	0.24	0.1	845.6
Day 2	29	0	0	16.8	0.24	0.14	188.12

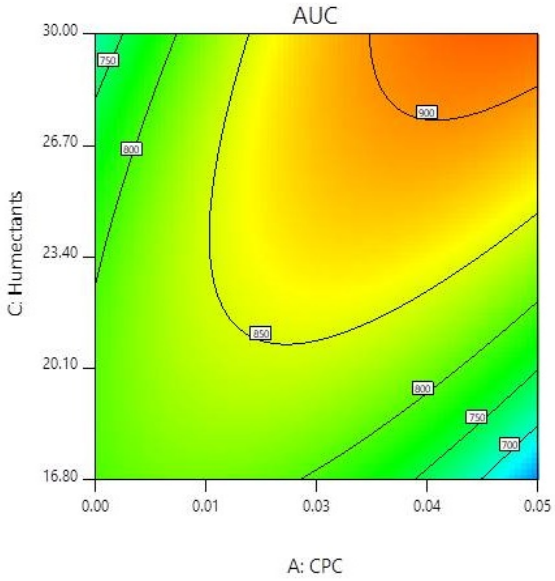
The corresponding ANOVA table, 3.27 is not a pretty sight. Nearly everything is statistically significant, – p-values are calculated with the residual mean square in the denominator – and there is significant lack of fit, albeit based on only a single degree of freedom.

Table 3.27 ANOVA Table for Fluoride Retention CCD

Source	Sum of Squares	df	Mean Square	F-value	p-value
Block	0.0633	1	0.0633		
Model	2.79	20	0.1393	8.59	0.0036
A-CPC	0.0796	1	0.0796	4.91	0.0623
B-PG	0.136	1	0.136	8.39	0.0231
C-Humectants	0.3594	1	0.3594	22.16	0.0022
D-PEG40	0.001	1	0.001	0.0628	0.8093
E-Flavor	0.0004	1	0.0004	0.0228	0.8843
AB	0.374	1	0.374	23.06	0.002
AC	0.3838	1	0.3838	23.67	0.0018
AD	0.0004	1	0.0004	0.025	0.8789
AE	0.0305	1	0.0305	1.88	0.2123
BC	0.1236	1	0.1236	7.62	0.0281
BD	0.1577	1	0.1577	9.72	0.0169
BE	0.0585	1	0.0585	3.61	0.0992
CD	0.0563	1	0.0563	3.47	0.1047
CE	0.0231	1	0.0231	1.42	0.2719
DE	0.4746	1	0.4746	29.26	0.001
A^2	0.0186	1	0.0186	1.15	0.32
B^2	0.0041	1	0.0041	0.2529	0.6305
C^2	0.008	1	0.008	0.4956	0.5042
D^2	0.0062	1	0.0062	0.3841	0.5551
E^2	0.0105	1	0.0105	0.6504	0.4465
Residual	0.1135	7	0.0162		
Lack of Fit	0.1135	6	0.0189	5220.41	0.0106
Pure Error	3.62E-06	1	3.62E-06		
Cor Total	2.96	28			

Recall that the response is the area under the fluoride retention curve (AUC). There is nothing that says that the underlying distribution of these areas should follow a normal distribution. Fortunately, the software used in this example provides many diagnostics including several types of response transformations (Box and Cox, 1964). These are especially useful when data span orders of magnitude. In this case, the recommendation is to analyze the AUC data taken to the 2.5 power. Admittedly, taking these data to such a large power lacks intuitive appeal, but it does not hurt to look. When this is accomplished, fewer factors stand out as being significant. Large among them is the interaction between CPC and Humectants as shown graphically in Figure 3.38. As can be seen, increased levels of both, from a starting point of (0, 16.8) will increase the AUC.

Figure 3.38 A Response Surface Illustrating the CPC-by-Humectants Interaction Evident from the Data of Table 3.27.



3.5.4.7 Mixture Designs

Statistical literature about mixtures first appeared late in the DOE game (Claringbold, 1955). Among the first industrial applications involved coffee formulations at General Foods by Mavis Carroll, a pioneering statistician there, who hired Henry Scheffé as a consultant to help figure out how to modify conventional designs to treat situations where the response is a function of the proportions of the components. Scheffé went on to develop and publish special designs for mixture experiments called simplex lattice and simplex centroid designs (Scheffé, 1958, 1963). This seems to have left people confused, not understanding what had been written, until some expository papers had been written (Snee, 1971; Cornell, 1973) to set off a wave of interesting and productive research.

What sets mixture experimentation apart from the conventional designs discussed so far is the main operating constraint that the sum of the components is constant. To those not familiar with mixture experimentation, this can be a bit baffling. A technician's supervisor asked the statistician if the technician was explaining factors for experimentation correctly. The statistician replied that the original list of components did not sum to 100%, but the problem was resolved, and design development was progressing nicely. The supervisor commented that this explained why they were having so much trouble getting the product into the bottle.

What characterizes mixture experimentation is that the response is a function of the proportions of the components, not their amounts. Consider Macbeth's witches' brew (Shakespeare, 1606):

"Fillet of a fenny snake
In the cauldron boil and bake

Eye of the newt and toe of frog
Wool of bat and tongue of dog
Adder's fork and blind-worm's sting
Lizard's leg and owlet's wing …"

It is not so much the amount of each of these delectable savories that make the brew right; it is their proportions relative to everything else in the cauldron. Get this wrong and there goes your hex.

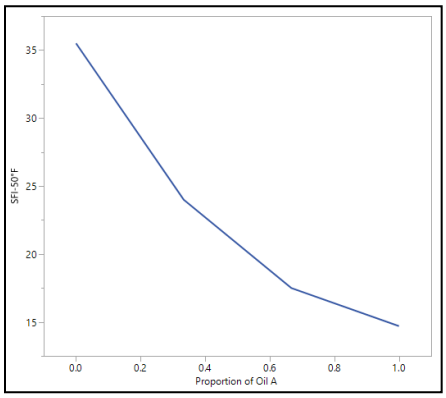
A simple, two-component example might help explain (Hare, 1974). Lipid chemists were blending vegetable oil to obtain synergies as measured by a solid fat index at 50°F. Fats such as margarine and shortening are designed to have melting characteristics at various temperatures. They should melt near body temperatures, otherwise they cannot be digested, and they should be solid at refrigerator temperatures so they can be handled as butter is. Chemists' practice is to run bottled blends of oils through baths at a range of temperatures, refrigerator to body, and examine the solid contents of each. Further, their hope was that the use of palm oil, which was less expensive than domestic oils, would produce blending synergies – higher solids at a lower cost.

The decision was to make up and run 4 mixtures, Table 3.28, of the two oils through the series of baths to learn of any synergy might take place. SFI is the Solid Fat Index.

Table 3.28 SFI Resulting from Vegetable Oil Blends

Run	Proportion of Oil A	Proportion of Oil B	SFI-50°F
1	1	0	14.7
2	0	1	35.5
3	2/3	1/3	17.5
4	1/3	2/3	24.0

Figure 3.39 SFI-50°F Resulting from Oil Blends

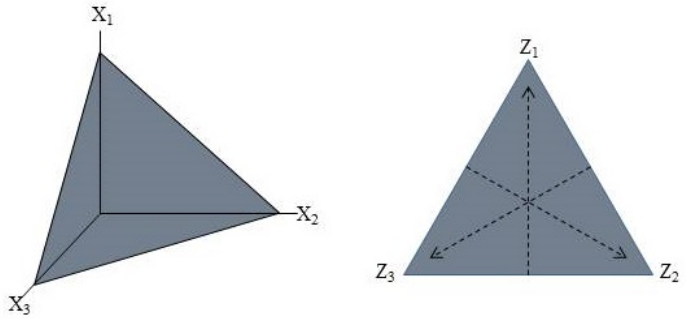


A simple plot of the data ends the discussion. The curve of Figure 3.39 showing SFI-50°F against increasing proportions of Oil A, which is the same as decreasing proportions of Oil B, is concave, indicating antagonism, not the hoped-for synergism. Is this real? Is this somehow statistically significant? Further details of mixture experimentation should help.

Unlike the experimental environment that serves as background for the designs suggested so far, in which factors are to free vary independently of each other, the mixture experimental environment is constrained so that the sum of the factors is a constant. You can see this easily in the oil blending experiment whereas one of the oil's proportion increased, the other's decreased. Their sum is always 1.0 or 100%.

When there are three components, the mixture space is limited to an area called a simplex. Figure 3.40 is illustrative. The shaded area shows the space for experimentation when the sum of the components is constrained to a constant, say 1.0. In side (b) of the figure, the experimental area is removed from the independent variables space and laid flat to form the simplex. New variable names are awarded, X to Z, so the estrangement does not make them feel orphaned. And new axes, represented by the dashed arrows run from 0 at the center of the edge opposite the axis name to 1.0 at the vertex.

Figure 3.40 (a) A Three-Component Mixture Space in Independent Variable Coordinates and (b) as a Simplex.



The constraint that the sum of the components is always a constant renders conventional models (See Montgomery, et.al, (2012)) different. A simple linear model in independent factors becomes

$$E(y) = \sum_{i=1}^{q} \beta_i Z_i$$

a quadratic model reduces to

$$E(y) = \sum_{i=1}^{q} \beta_i Z_i + \sum\sum_{i<j}^{q} \beta_{ij} Z_{ij}$$

and a special cubic model is

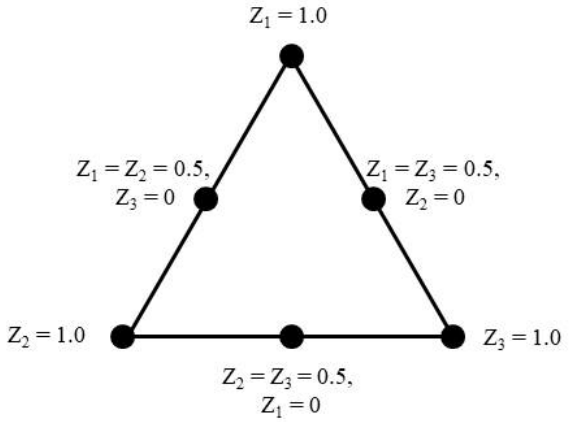
$$E(y) = \sum_{i=1}^{q} \beta_i Z_i + \sum\sum_{i<j}^{q} \beta_{ij} Z_{ij} + \sum\sum\sum_{i<j<k}^{q} \beta_{ijk} x_i x_j x_k$$

in mixture terms. Also, the βs in the mixture model are not the same as the βs in the independent variable mode. Instead, the mixture model βs are a combination of their linear model counterparts. To model mixture data in most software packages, you must let them know that you are working with mixture variables by forcing the constant term to zero or otherwise alerting the modeling routines.

Extensive work has been carried out on mixture models (Becker, 1968) and on test statistics for mixture models (Marquardt and Snee, 1974). Many of their developments have found their way into software packages.

In his pioneering work, Scheffé, mentioned earlier in this section, developed mixture lattice designs designated by {q,m} where q is the number of components and m is the degree of the model intended to be fitted to the data. The intent is that there should be m + 1 levels of each factor in m equal steps. The simplest of these designs is depicted in Figure 3.41.

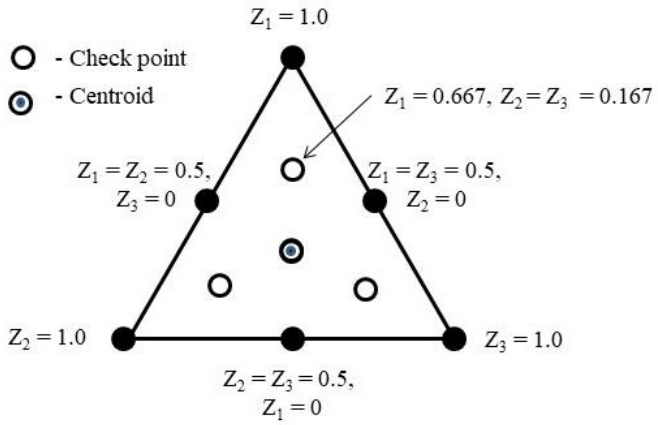
Figure 3.41 A Scheffé Simplex Lattice {3,2} Design



While they have nice statistical properties making them appropriate for the intended model, they lack intuitive appeal. The m = 2 designs have no interior points, and those are in areas where scientists have the greatest curiosity. Designs with m > 2 have only sparse interior points, and they involve many more distinct formulations.

A good compromise is the simplex-centroid design (Cornell, 2011). Rich with interior points, it positions "check points" midway between the centroid and the vertices. These can be used to check the model's ability to fit the data well. It is often recommended but too often ignored that the center point be replicated.

Figure 3.42 A Simplex-Centroid Design in Three Components



As the number of components increase, these designs can become quite large. For 3 components there are 10 design points; for 4, there are 19, and for 5 there are 36, if check points are included. Reduced complex centroid designs consist of only the pure components, all possible (½, ½) blends and a centroid. For them, when there are 3, 4, 5 or 6 components, the numbers of distinct blends required are 7, 15, 31 and 63, respectively (Snee and Hoerl, 2016).

In a follow-up experiment, our intrepid oil chemists studied a third oil source, again to track solids at 50°F in pursuit of blending synergies. Their experimental design and resulting data are shown in Table 3.29.

Table 3.29 SFI Results from a Simplex-Centroid Design of Vegetable Oil Components

Run	Proportion of Oil A	Proportion of Oil B	Proportion of Oil C	SFI-50°F
1	1	0	0	4.6
2	0	1	0	35.5
3	0	0	1	55.5
4	1/2	1/2	0	14.5
5	1/2	0	1/2	25.7
6	0	1/2	1/2	46.1
7	1/3	1/3	1/3	27.4
8	2/3	1/6	1/6	14.5
9	1/6	2/3	1/6	32.0
10	1/6	1/6	2/3	42.5

A model found to fit these data well is:

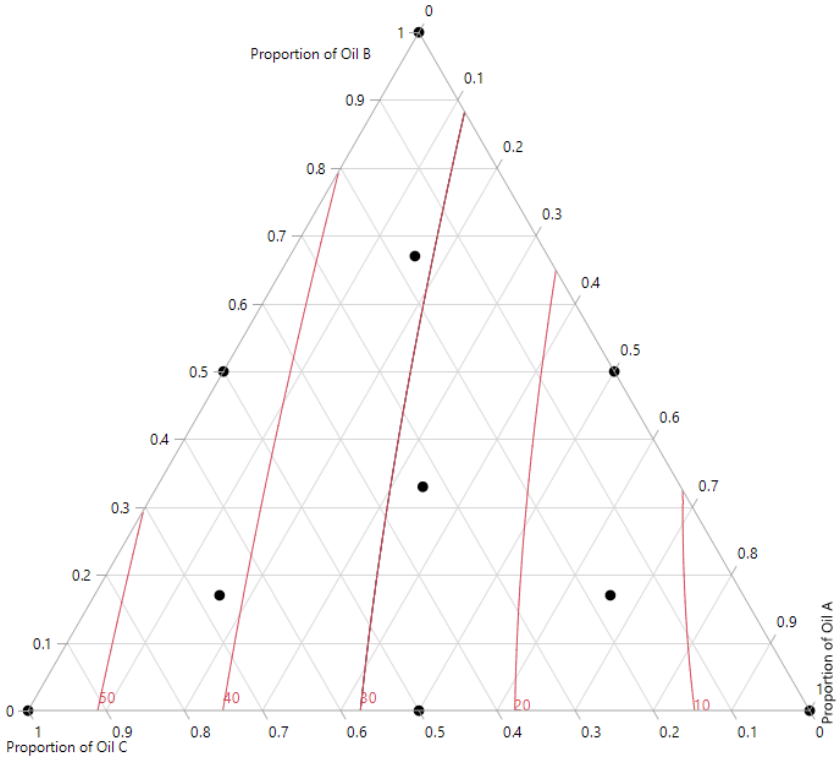
$$Y = 4.5Z_1 + 35.9Z_2 + 55.9Z_3 - 21.5Z_1Z_2 - 16.6Z_1Z_3,$$

where the Y represents the SFI-50°F, and the Zs represent the proportions of the oils. By itself, the model provides estimates of the "true" values at the vertices or pure components, and it tells us about possible non-linear blending or response curvature. (We avoid the use of the word

interaction in mixtures because of the correlations among components.) But the model is not satisfactorily informative.

It helps to display the model in response surface form. This is done by superimposing response surface contours on the simplex as in Figure 3.43.

Figure 3.43 Response surface drawn from the data in Table 3.29. Red lines are contours of constant response as marked. Black dots mark design points.



Our oil chemists may have been disappointed at the lack of synergy as indicated by the contours being nearly straight lines, but an advantage of the study's resulting response surface is that it shows various ways a specific SFI-50°F might be attained. For example, if the chemists were seeking an index of 30, it could be obtained by Oils A, B and C at 17%, 69% and 13%, respectively, or it could be obtained by them at 29%, 33% and 37%, respectively or many other formulations along the "30" line on the response surface. By imposing costs, the chemists can learn how to obtain the least expensive formula with the desired SFI profile.

The notion of imposing cost constraints on the SFI profile came about as an unintended revelation of the blending experiments. This is typical of the accelerated discovery process that takes place as a result of well-planned experimentation.

You cannot make concrete with 100% sand, and if you try to make pudding with 100% milk, it will not pud. The point is that many if not most formulation studies have constraints on compositions.

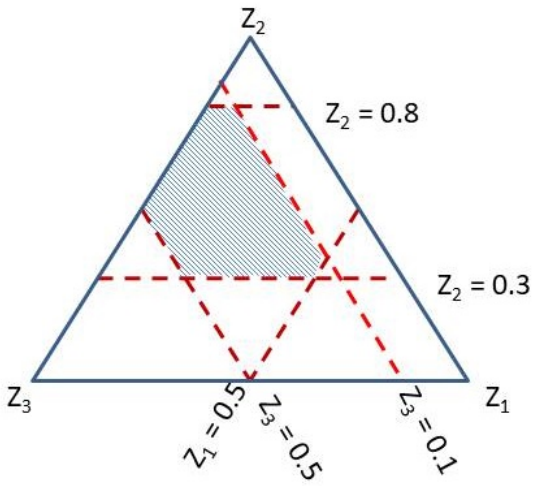
Surfactant contents of body washes all have lower and upper practical bounds. Suppose we examine only the surfactant content of a body wash, leaving other components such as water, thickeners and aromas out of consideration. We could consider the sub-ingredient mix of surfactants to be its own composition, holding the proportions of all other ingredients constant. There may be known constraints on the surfactant mix, itself. Prior research and cost constraints may impose bounds as:

A. $0.0 \leq Z_1 \leq 0.5$
B. $0.3 \leq Z_2 \leq 0.8$
C. $0.1 \leq Z_3 \leq 0.5$,

where the Zs compose 100% of the surfactant blend.

These constraints can be drawn on the simplex as in Figure 3.44. Realistic experimentation may only take place in the shaded region.

Figure 3.44 A Three-Component Constrained Mixture Space



It turns out that the most efficient experimental designs in constrained mixture regions usually contain some or all of the vertices of the region, and the number of vertices increases dramatically with the numbers of components and constraints. McLean and Anderson (1966) developed an algorithm for finding these vertices, and others have developed algorithms for finding the best subsets of vertices, centers of long edges and constraint plane centroids for given numbers of blends and model forms (Piepel, 1988, Snee, 1975, 1985).

It should be noted that much of the work done to develop mixture design technology parallels that done to develop conventional (or independent variable) exploration. The same principles hold. In situations involving large numbers of components, for example, the strategy presented in Section 6.3 applies; that is use screening followed by optimizing designs. For more on mixture screening designs, see Snee and Marquardt (1976).

Development of mixture design technology persists and includes blocking, incorporation of process variables, e.g., exploring the effects of cake ingredients and of baking variables in combined designs, and estimation of component effects. Software development keeps up with the technology, and the best packages include helpful guidance and hints.

3.5.4.8 Split Plot Designs

Complete randomization of experimental treatment combinations is fundamental to the probability theory underpinning the analysis of the resulting data. If there is no randomization, all bets are off. In the world of experimentation however, some factor levels are harder to change than others. Naturally, from the perspectives of those engaged in the physical experimentation, the hard to change factors should be changed less often than those whose levels are easy to change. In many situations, full randomization is not possible.

A solution which can save much labor and expense is the application of split plot designs (Fisher, 1925). It does not forsake randomization so much as it capitalizes on it by executing the design in tiers. The name *split plot* derives from agricultural experimentation involving several whole plots of land which must receive entire treatments of one kind but can be subdivided to receive treatments of other kinds. The terminology (plot) stuck while many design applications are outside agricultural science.

A split plot design can be thought of as a design within a design. There is a whole plot design and a subplot design. In an agricultural experiment, there may be fields which can be plowed one way or another but might have fertilizers, herbicides and pesticides applied randomly in factorial order within the fields. The plowing method constitutes the whole plot and has one kind of inherent variation, while the remaining treatments form the subplots and have another kind of inherent variation.

Table 3.30 lists sensory responses measuring baked cracker degree of golden brownness – low numbers are too brown and higher numbers are too gold – resulting from a split plot design involving oven temperature, target final moisture, and order of shortening addition – before or after water in the mix.

Table 3.30 Cracker Color as a function of Oven Temperature, Initial Moisture and order of Shortening Addition.

Whole Plot	Oven Temperature	Moisture	Shortening	Golden Brownness
1	375	3	After	6.5
	375	3	Before	5.1
	375	4	Before	5.3
	375	4	After	4.6
2	400	4	Before	6.9
	400	3	After	6.4
	400	4	After	6
	400	3	Before	7.1
3	350	3	After	1.2
	350	4	After	2
	350	3	Before	4
	350	4	Before	5.7
4	400	3	After	6.4
	400	4	Before	6
	400	3	Before	5.2
	400	4	After	7.8
5	350	4	Before	6
	350	3	Before	5.6
	350	4	After	2.8
	350	3	After	3.9
6	375	3	After	2.2
	375	3	Before	3.7
	375	4	Before	4.7
	375	4	After	6.7

In each whole plot, the moisture and shortening addition variables form a 2^2 factorial design, and each of the three temperatures is replicated to form six whole plots. Table 3.31 shows the analysis of variance summary. It is important to count degrees of freedom to assure that the ANOVA generated corresponds to what is intended. Notice that the numerator degrees of freedom for oven temperature is 2, as expected. It should be tested against the whole plot replication degrees of freedom, one for each unique whole plot. That comes to 3, so the stated F-value and corresponding p-values are appropriate. The counting of degrees of freedom are appropriate for the remaining sources of variation as well. Together, the oven temperature and its error term (Replicates within Whole Plots) use up 5 degrees of freedom. Because there are 23 degrees of freedom in all the data, and because the remaining terms account for an additional 9 degrees of freedom, there are 9 left over to measure the error. That is why there are 9 denominator degrees of freedom listed for all terms below oven temperature in the table. The analysis makes sense. If any factor is important to get just the right level of golden brownness it is the oven temperature, perhaps as it interacts with the order of shortening addition.

Table 3.31 ANOVA Summery of Data in Table 3.30

Source	DF Numerator	DF Denominator	F-Value	P-Value
Oven Temperature	2	3	6.81	0.077
Moisture	1	9	1.23	0.295
Shortening	1	9	1.84	0.208
Oven Temperature*Moisture	2	9	0.11	0.901
Oven Temperature*Shortening	2	9	3.84	0.062
Moisture*Shortening	1	9	0.01	0.928
Oven Temperature*Moisture*Shortening	2	9	0.28	0.764

Working with split plot designs and analyzing the resulting data can be a challenge. For one thing, experimenters often split the plot without informing the designer. That is the road to perdition. If it leads to correct answers, it is a coincidence.

Many modern statistical software packages shy away from the conventional analysis of split plot data in favor of ordinary least squares (OLS) solutions. While OLS is more appropriate, especially if there are missing data, the user must be careful while working to understand the output.

The good news in this still developing field is that there is much good advice (Jones and Nachtsheim, 2009) about designing split plot experiments to include factorial designs in both whole and split plots as well as fractional factorial designs and mixture designs in sub-plots and about their analysis and interpretation.

3.5.4.9 Incomplete Block Designs

Suppose you work in cosmetics and you want test three underarm deodorants for effectiveness after 24 hours. Yes, they have sniff tests for such things. Let us not go there. Problem is people only have two arms. What to do?

While you could have many people in each of three separate groups carry out the evaluations, the variation from person to person could mask the differences among deodorants. If only there was a way to take advantage of the randomized block idea of partitioning out the differences among people. Take heart, there is!

Suppose you divide subjects into three groups or blocks. Block 1 receives deodorants A and B; Block 2 receives A and C; and Block 3 receives B and C. There is a healthy balance there. Each block is missing one treatment and is therefore incomplete, hence the name.

Table 3.32 A Simple Balanced Incomplete Block Design

Block	Deodorant A	Deodorant B	Deodorant C
1	X	X	
2	X		X
3		X	X

Of course, things are never quite that simple. An expert may point out the human left side secretions might differ from right side. A possible solution is to divide the members of each block into two groups, with one group receiving its products in left-right order and the other in right-left order.

Table 3.32 shows an example of a "2 of 3" BIB design. Of course, there are many other BIB designs. The most practical have been catalogued by Cochran and Cox (1950) who show plans, the number of treatments per block and the number of treatment replications. They also list Partially Balanced Incomplete Block (PBIB) designs – those that hold some slight correlations among treatment effects.

BIB designs are available in some statistical software packages. If you have the resource to run as many blocks as you like, many combinations of treatment and block sizes are available.

Often, BIB designs are used in consumer studies where there may be many more treatments than a consumer is capable of or willing to experience in one sitting. Studies have shown that consumer fatigue increases as the number of samples experienced exceeds 20 or thereabout. The use of the BIB design results in the loss of some statistical efficiency but probably increases the testing validity.

For example, an original product and/or process design may be a 5-factor response surface study involving 35 experimental treatment combinations including replicated center points. Experimenters may choose 35 blocks with 17 treatments per block. Of course, other strategies such as blocking by fractional replication, as discussed earlier, are design alternatives.

Data analyses are carried out by regression packages that partition out the treatment and block effects and leave the residual error which may be a mix of random error and other, assumed smaller effects.

3.5.4.10 Definitive Screening Designs

Dramatic advances in computer searches and simulations augmented the great insight of Jones and Nachtsheim (2011), aiding in the development of what are called "Definitive Screening Designs (DSD)." While the name seems a bit boastful, it is justified, at least in part, by the design characteristics:

1. The number of required experimental treatment combinations is only one more than twice the number of factors
2. Main effects are free of confounding with two-factor interactions
3. Two-factor interactions are not completely confounded with each other
4. Quadratic effects are estimable when models comprised of linear and quadratic terms are used
5. Quadratic effects are not completely confounded with two-factor interactions
6. With at least six factors, a complete quadratic model can be fit to three or fewer factors

to paraphrase the reference.

In practice, the initially recommended number of experimental treatment combinations is augmented by 4 additional runs. The runs in Table 3.33 are deliberately left unrandomized to show that DSD runs are in mirrored pairs, runs 1 through 12, inclusive. Runs 13 – 16 are augmented mirrored pairs, and run 17 is an overall center point.

Table 3.33 A Six Factor DSD

Run	X_1	X_2	X_3	X_4	X_5	X_6
1	0	1	1	1	1	0
2	0	-1	-1	-1	-1	0
3	1	0	1	1	-1	1
4	-1	0	-1	-1	1	-1
5	1	-1	0	1	1	1
6	-1	1	0	-1	-1	-1
7	1	-1	-1	0	1	1
8	-1	1	1	0	-1	-1
9	1	1	-1	-1	0	1
10	-1	-1	1	1	0	-1
11	1	-1	1	-1	-1	1
12	-1	1	-1	1	1	-1
13	1	1	-1	1	-1	1
14	-1	-1	1	-1	1	-1
15	1	1	1	-1	1	1
16	-1	-1	-1	1	-1	-1
17	0	0	0	0	0	0

These designs stand in contrast to the fractional factorial designs which are augmented with center points, in that DSDs may estimate distinct parameters of curvature, whereas the fractional factorial designs will alert the user to the presence of curvature but not permit estimation of its parameters.

The authors have pressed forward to develop DSDs for situations where categorical factors are involved in the design and/or where blocking is necessary. In each case, only a few extra runs may be required, but the confounding pattern becomes slightly more complex.

Point 6, among the design characteristics is a very strong one. It is consistent with many statisticians' experience that in experiments with many factors only three to six will stand out as being truly important. Given this, it might be tempting to embrace a DSD as the one-time experiment that answer all questions. Certainly, this is not the developers' intent. DSDs are still screening designs and the principles of building knowledge via the sequential application of screening, characterizing and optimizing designs as described earlier still hold.

Research into the utility and effectiveness of DSDs continues (Jones, 2016) as developers evaluate and compare alternatives.

3.5.4.11 More Designs

There may be as many designs as there are applications for them. Experimentally, each situation is unique in some sense. In this context, the designs described above are broad categories to be tailored to the situation at hand. Following are some design categories.

3.5.4.11.1 Optimal Designs

In some pharmaceutical applications, experimental treatment combination can cost over a million dollars. Squeezing every last drop of information from the data is imperative. Computer generated optimal designs (Goos and Jones, 2011) can help. These are designs in various spaces, constrained or otherwise, that assume the model is correct and minimize the number of experimental combinations needed to fit it.

Temptations are strong to save money by applying these designs at every chance. Caution is in order because we never know the model.

3.5.4.11.2 Latin Square Designs

Latin square designs and their extensions, Graeco-Latin square designs and hyper-Graeco-Latin square designs, have their roots in the same theory that spawned Sudoku puzzles. In these squares, only one unique symbol may appear in each row and column. Rows, columns and symbols all represent design factors, each at the same number of levels.

Latin square designs measure only main effects. As a result, their direct use can lead to incorrect conclusions if interactions are present (Hunter, 1989). This is not to suggest that they should never be used. Rather, strong subject matter should be present beforehand.

These designs might be more useful as they are imbedded in other, more extensive designs such as consumer studies and clinical trials in which the order of sample presentation may influence attitudes and efficacies. Very useful among these designs is a subset called Williams Latin Squares (Williams, 1949) which balance the order of presentation so that within a square no treatment follows another treatment more than once. Table 3.34 lists some examples.

Table 3.34 Three Williams Latin Square Designs

Three Treatments		
A	B	C
B	C	A
C	A	B

Four Treatments			
A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

Five Treatments				
A	B	E	C	D
B	C	A	D	E
C	D	B	E	A
D	E	C	A	B
E	A	D	B	C

3.5.4.11.3 Robust Parameter Designs

Some years ago, Saturday Night Live had a skit about a floor wax that doubled as an ice cream topping. Now, there is an example of a rugged product. But if you think about it, you will come to discover that some of the most successful products on the market are among those that are most rugged.

Tea bags, for example, make a passable product no matter how used or abused – lemon, milk (not together, please), hot or iced, made with hard or soft water. Some people even get two cups of tea out of the same bag. And who has not dropped a cell phone, immediately thinking the worst, only to be relieved when it is discovered to be fully functional? The former happened by chance, but the latter by design.

Designs for aiding in the creation of rugged products are called Robust Parameter Designs (RPD). The ideas date to the early twentieth century but were formally proposed by Taguchi and Wu (1980) and Taguchi (1987). The notion caught hold industry wide. Envision a situation where scientists and engineers have an experimental design covering many product characteristics together with an array of environmental characteristics within whose combinations the product is expected to perform. Taguchi's idea was to form an "outer array" of product characteristics and an "inner array" of environmental characteristics. His original notion was to cross the two arrays to form a resulting design.

While the fundamental idea was pioneering and even inspiring to many, the suggested methods of data analysis and interpretation left a great deal to be desired. Much has been written since about more efficient designs and more informative methods of analysis. (Box, Hunter and Hunter, 2005) and (Montgomery, 2013) Fundamentals include the combined use of response surface methods and split plot designs discussed above.

Section 3.5 References

Andrews, H.P. "The Role of Statistics in Setting Food Specifications," Proceedings of the Sixteenth Research Conference of the Research Council of the American Meat Institute, (1964): pp. 43-56.

Becker, N. G. "Models for the Response of a Mixture," Journal of the Royal Statistical Society, B. Vol. 30, (1968): pp. 349-358

Box, G. E. P. *An Accidental Statistician*, John Wiley and Sons, New York, 2013.

Box, G. E. P. and D. W. Behnken. "Some New Three Level Designs for the Study of Qualitative Variables," Technometrics, Vol. 2, (1960): pp. 455-476.

Box, G. E. P. and K.G. Wilson. "On the Experimental Attainment of Optimal Conditions," Journal of the Royal Statistical Society, B. Vol. 13, (1951): pp. 1-45.

Box, G. E. P. and D.R. Cox. "An Analysis of Transformations," Journal of the Royal Statistical Society, B. Vol. 26, (1964): pp. 211-243.

Box, G. E. P. and J.S. Hunter. "Multifactor Experimental Designs for Exploring Response Surfaces," Annals of Mathematical Statistics, Vol. 28, (1957): pp. 195-242.

Box, G. E. P. and J.S. Hunter. "The 2^{k-p} Fractional Factorial Designs" Technometrics, Vol. 3, (1961).

Box, G. E. P., Hunter, J. S. and W. G. Hunter. *Statistics for Experimenters*, John Wiley and Sons, New York, 2005.

Claringbold, P.J. "The Use of the Simplex Design in the Study of Joint Action of Related Hormones," Biometrics, Vol. 11, (1955): pp. 174-185.

Cochran, W. G. and G.M. Cox. *Experimental Designs*, 2^{nd} Ed. John Wiley and Sons, New York, 1957.

Cornell, J.A. "Experiments with Mixtures: A Review," Technometrics, Vol. 15, (1973): 437-455.

Cornell, J.A. *A Primer on Experiments with Mixtures*, John Wiley and Sons, New York, 2011.

Deming, W. E. *Out of the Crisis*, Massachusetts Institute of Technology, Center for Advanced Educational Services, Cambridge, MA, 1982.

Finney, D.J. "The Fractional Replication of Factorial Arrangements," Annals of Eugenics, 12, (1947): pp. 291-301.

Fisher, R.A. *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburg and London, 1925.

Goos, P. and B. Jones. *Optimal Design of Experiments*, John Wiley and Sons, West Sussex, UK, 2011.

Hahn, G. J. and W. Q. Meeker. *Statistical Intervals*, John Wiley and Sons, New York, 1991.

Hare, L. B. "Mixture Designs Applied to Food Formulation," Food Technology, Vol. 28, No. 3, (1974): pp. 50-56, 62.

Hare, L. B. "In the soup: A Case Study to Identify Contributions to Filling Variability," Journal of Quality Technology, Vol. 20, No. 1, 1988.

Hunter, J.S. "Let's All Beware the Latin Square," Quality Engineering, Vol. 1., (1989): pp.453-465.

Jones, B. "21st Century Screening Experiments – What, Why and How," Quality Engineering, Vol. 28, No. 1, (2016): pp. 98-106.

Jones, B. and C.J. Nachtsheim. "A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects," Journal of Quality Technology, Vol. 43, No.1, (2011): pp. 1-15.

Marquardt, D. W. and R.D. Snee. "Test Statistics for Mixture Models," Technometrics, Vol. 16, (1974): pp. 533-537.

McLean, R.A. and V .L. Anderson. "Extreme Vertices Design of Mixture Experiments," Technometrics, Vol. 8, (1966): pp. 447-454.

Montgomery, D.C. *Design and Analysis of Experiments*, 8th ed., John Wiley and Sons, New York, 2013.

Montgomery, D. C., E. A. Peck and C. G. Vining *Introduction to Linear Regression Analysis*, 5th ed., John Wiley and Sons, Hoboken, New Jersey., 2012.

Myers, R. H. and D C. Montgomery. *Response Surface Methodology: Product and Process Optimization Using Designed Experiments*, 2nd ed., John Wiley and Sons, New York, 2002.

Piepel, G. F. "Programs for Generating Extreme Vertices and Centroids of Linear Constrained Experimental Regions," Journal of Quality Technology, Vol. 20, (1988): pp. 125-139.

Scheffé, H. "Experiments with Mixtures," Journal of the Royal Statistical Society, B. Vol. 20, (1958): pp. 344-360.

Scheffé, H. "The Simplex-Centroid Design for Experiments with Mixtures," Journal of the Royal Statistical Society, B. Vol. 25, (1963): pp. 235-263.

Shakespeare, William, "Macbeth", 1606.

Snee, R. D. "Design and Analysis of Mixture Experiments," Journal of Quality Technology, Vol. 3, (1971): 159-169.

Snee, R. D., L. B. Hare and J.R. Trout, Eds. *Experiments in Industry*, American Society for Quality, Milwaukee, WI, (1985).

Snee, R. D., and R.W. Hoerl. *Strategies for Formulations Development*, SAS Institute, Inc., Cary, NC, 2016.

Snee, R. D. "Experimental Designs for Quadratic Models in Constrained Mixture Spaces," Technometrics, Vol. 17, (1975): pp. 149-159.

Snee, R. D. "Computer-aided Design of Experiments: Some Practical Experiences," Journal of Quality Technology, Vol. 17, (1985): pp. 222-236.

Snee, R. D. and D. W. Marquardt. "Screening Concepts and Designs for Experiments with Mixtures," Technometrics, Vol. 18, (1976): pp. 19-29.

Salsburg, D. *The Lady Tasting Tea*, W. H. Freeman and Company, New York, 2001.

Taguchi, G. and Y. Wu. *Introduction to Off-Line Quality Control*, Central Japan Quality Control Association, Nagoya, Japan, 1980.

Taguchi, G. *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost*, UNIPUB, White Plains, NY, 1987.

Tukey, J. "Comparing Individual Means in the Analysis of Variance," Biometrics 5 (2), (1949): pp 99-144.

Wasserstein, R. L., and N.A. Lazar. "The ASA's Statement on p-Values: Context, Process and Purpose," The American Statistician, Vol. 70, No. 2, (2016): pp 129-133.

Weisberg, H.I. *Willful Ignorance*, John Wiley and Sons, New York, 2014.

Williams, E. J. "Experimental Designs Balanced for the Estimation of Residual Effects of Treatments," Australian Journal of Scientific Research, Vol. 2, No. 3, (1949): pp. 149-168.

Section 3.6 – Chapter Summary

Data collection is no simple task. If sound decisions are to be made, it should not be taken lightly.

This discourse begins with discussions of data terminology and use followed by basic statistical principles, accompanied by examples of important concepts leading to sampling for inference in the shadows of variation. Discussions of the nature of variation, the distinction between sample and population, measurement systems, common and special cause variation, and other basic concepts set the stage for the progression of topics to follow.

Of course, data collection can be a rocky road, scattered by potholes and pitfalls. For value-rich data to be gathered, their basic nature must be determined. Are they qualitative or quantitative, subjective or objective, and especially in the case of "big data" are they fraught with errors and redundancies?

To pave the road, considerations of data pedigree must be made. We must know their sources, where and with whom have they been? What is their chain of custody, and has any tampering taken place? These are issues too often overlooked. Assumptions are made that the data are "the data", and if we have enough data, the rough places will be made smooth. Errors in this logic are pointed out, and steps in the assessment of data pedigree are presented.

An additional concern also overlooked is the issue of the measurement system itself. Is it accurate and precise? How might we know? That is, what techniques should be used to assess accuracy and precision of measurement processes, and what must be done to assure measurement variation does not interfere with the quality of decisions to be made?

Planning for data collection requires teamwork and considerable thought to assure that the right amount of the right kind of data are accumulated sensibly for excellent decision making. A brief history of experimental design (DOE) technology is presented. It is followed with informal discourse regarding types of designs, from simple to complex; some with worked examples.

All elements of this chapter are to further the aims of statistical engineering, including identifying and capitalizing on opportunities for improvement and solving difficult problems.

Chapter 4 - Data Exploration

Table of Contents

Preface

Statistical Engineering is a data-based methodology and, as such involves data analysis. The first step in data analysis is data exploration which helps the analyst understand the data, the data pedigree discussed in Chapter 3, characterize the variation in the data and what some potential predictor (causal) variables might be. In this chapter three important aspects of data exploration are addressed: theory of data exploration, data cleaning and using Exploratory Data Analysis (EDA). A high-level discussion of EDA is presented to provide the foundation for data exploration. A "Global Positioning System" for an effective EDA journey is presented. Data are rarely "clean" and can have a variety of problems and limitations. Five types of data cleaning problems and methods for conducting data cleaning are discussed. The chapter concludes by showing how EDA can be used to understand data prior to doing formal statistical analyses. EDA helps the analyst to be alert to unexpected patterns, relationships and extraordinary cases. Some tools useful in using data analysis are described and illustrated.

Section 4.1 - Philosophy of Exploratory Data Analysis

4.1.1 Objectives

Exploratory Data Analysis is defined and addressed at a high level providing the foundations for the methodology. The use of the resulting philosophy of data analysis is discussed and compared to other approaches. A GPS for an effective EDA journey is presented.

4.1.2 Outline

After defining EDA, the EDA journey is discussed along with competing models, understanding outliers, viewing EDA as a philosophy of science, EDA and other methods, and the norms of EDA.

4.1.3 What Is Exploratory Data Analysis?

Exploratory Data Analysis (EDA), pioneered by John W. Tukey (1915-2000) is widely recommended as an initial step in any data analysis. EDA is a philosophy of critical thinking about data along with practical tools and approaches for data analyses. It calls for being alert to unexpected patterns, relationships and extraordinary cases. And these can arise or be revealed almost any time in the course of an analysis, so it is wise to be mindful of EDA philosophy throughout an analysis.

Tukey wrote extensively about the philosophy of EDA, for example, in his 1962 declaration "The Future of Data Analysis." Yet his recommendations continue to stand in contrast to common practice in statistical analyses and data science. This is in part because they focus, not on hypothesis tests, optimization, and drawing conclusions from data, but on understanding what the data are trying to tell us. Specifically, he suggests that data analyses should rely on indication rather than confirmation. EDA seeks ways to allow the data to speak for itself and to indicate paths for investigation without being restricted to previously selected hypotheses.

4.1.4 The EDA Journey

Data exploration is a journey. Exploring often requires that we leave the beaten path. But it does not require that we travel without a GPS. The challenge is not just to know where we are but to then see where we are going. Our path moves between data-driven attempts to refine working models and model-driven attempts to perfect the data. Perfecting the data may include choosing re-expressions that simplify its structure, nominating

Figure 4.1

A *GPS* for Exploratory Data Analysis

- Data exploration is a journey; plan accordingly
- Critical EDA check points: display, re-expression, residuals, resistance, and iteration
- Make extensive use of graphics along the way
- Expect graphical displays to reveal the unexpected
- Always be on the lookout for non-random patterns
- Focus on simplicity and parsimonious models
- Models are always imperfect; work to enhance models
- EDA is inherently subjective; use subject matter knowledge whenever possible

possible outliers for special consideration, and the selection of variables to include in a model.

Lacking a map, we need to take frequent sightings. We make many displays of the data, and we continue to do so all along the way. EDA is rich in graphical methods and readily adopts new ones as they are made available by technology. Early methods for pencil-and-paper work such as stem-and-leaf displays and boxplots have been joined by computer-driven methods such as rotating plots, plot matrices, plot brushing, and the ability to identify individual points interactively on the computer screen and search for information about them on the internet. As part of the journey, we refine the questions being asked of the data. We often approach a data set with some goals in mind, but these should not blind us to noticing unanticipated patterns, relationships, or extraordinary cases. For example, a display might reveal unexpected subgroups in the data. That could call for introducing a new variable that identifies the subgroups, locating a previously unincluded variable that accounts for the subgroups, or simply analyzing each of the subgroups separately. As a result, we could find ourselves following unplanned paths.

But we are not obliged to follow steep paths when gentler ones are available. We re-express data to make it more nearly symmetric and thus easier to summarize, to make simple (e.g., linear or additive) models fit more appropriately, and to identify outliers. Data analyses should prefer simpler models over more complex ones and should prefer models for which we must estimate fewer parameters whenever possible. The value of parsimony has been noted by many philosophers of science, often citing "Occam's Razor." EDA provides a practical way to realize this preference by favoring re-expression over the fitting of more complex (e.g., quadratic or exponential) models and by perfecting the data during the analysis process.

4.1.5 EDA and Models

Progress in an exploratory analysis builds on imperfect models. Although they are imperfect, they are a constructive part of the data analysis process, not an impediment. Others have opined on this subject. Box famously declared that "All models are wrong, but some are useful." But we can trace the idea back to the founder of scientific thinking, Francis Bacon (1561–1626), who wrote (in Latin; this translation is by Urbach):

> But since **truth will emerge more readily from error than from confusion**, I consider it useful…for the understanding to be given leave to exert itself…[1]
> (*Novum Organum II 20*)

EDA follows this advice, working with models that make errors, for example, examining those errors in the residuals and trying alternative forms.

[1] Tukey and Wilk quote the boldface part of this quotation in "Data Analysis and Statistics." It has been suggested that Bacon may have been quoting an earlier Scottish proverb.

4.1.6 Identifying and Understanding Outliers

A related focus of EDA is the identification and understanding of outliers and exceptions. Exploratory methods are resistant to the influence of outliers whenever possible. Of course, outliers are defined relative to some working model; hence, the importance of having a model, however imperfect, to work with and improve. But where some analysts fear or dislike outliers, the data explorer seeks them out for deeper understanding. Here I must again appeal to Bacon:

> …errors of Nature, sports and monsters...correct the understanding in regard to ordinary things and reveal general forms. … For **whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways.** (*Novum Organum II 29*)

I know of no more concise statement of the interplay of models and data cleaning than this 300-year-old advice.

4.1.7 EDA – A Philosophy of Science

The philosophy of Exploratory Data Analysis is fundamentally philosophy of science. The challenge in the practice of EDA is that, as described here, EDA is inherently subjective. The paths we follow are not pre-determined, and the decisions cannot be automated. Because of this subjectivity, different analysts may arrive at different models. Indeed, the exploratory data analyst may entertain alternative models for the same data, allowing them to compete in a "survival of the best fit."

The exploratory data analyst must apply judgment at many steps along the path. In his 1962 address, "The Future of Data Analysis," Tukey noted that those who think of statistics as optimization tend to think that "data analysis should not *appear to* be a matter of judgment." He italicized "appear to" because, of course, judgment is required in the selection of methods and their criteria. He continues noting that, by contrast, "In data analysis we must look to a very heavy emphasis on judgment."

4.1.8 Interfacing EDA with Other Methods

The requirement of judgment can put EDA at odds with machine intelligence, the Lasso, and other automated methods. Those who wish to use such methods should, at a minimum, explore their data using EDA approaches before turning it over to automated algorithms; the resulting analyses are almost certain to be substantially improved. As Tukey advises, "Scientists know that they will sometimes be wrong; they try not to err too often, but they accept some insecurity as the price of wider scope. Data analysts must do the same." (ibid.)

The tools EDA brings to bear on data are summarized by Velleman and Hoaglin as Display, Re-expression, Residuals, Resistance, and Iteration.

Displays excel at revealing the unexpected. Computer-based displays now can move, morph and rotate. They can be interactive so that the analyst can identify a case simply by pointing to it. All of these capabilities support the basic motivation to think critically about data, never assuming that it is coherent or consistent with the model you intended to use.

Re-expression should be a common part of data exploration because we should doubt whether the data are expressed in the right units for analysis. EDA recognizes that much—perhaps even most—data come to us in a form not optimal for analysis. Re-expression can often make distributions more nearly symmetric, stabilize variance across groups, straighten a relationship, or make the effect of a categorical factor additive. Mosteller and Tukey suggest we regard Re-expression as "first aid" – something to do at the beginning of any analysis. But it should also be considered as a response to features found during the analysis—for example, features revealed in residuals.

Residuals come from summarizing the patterns found using a temporary model and subtracting that summary from the data to reveal departures and additional patterns. They are a major step in the process of fitting models and then using them to correct the data and try again.

Resistance refers to using methods that are not affected by outliers and anomalies. Many traditional statistical models tend to be "self-justifying." That is, an outlier will pull the model close, so that the corresponding residual will not expose it. Resistant methods can more effectively reveal model violations. Many EDA methods are based on order statistics and are thus resistant to the occasional outlier.

Iteration is the heart of the process of data exploration. A model or test is never the final word, but just the best we have done so far, and open to improvement at any time.

In summary, the philosophy of exploratory data analysis brings the underlying philosophy of science home to statistics. It informs how we interact with data to find a path forward and understand what can be learned from the data. It should be present at each step of an analysis. Look at the residuals, inquire about outliers, consider re-expression, and entertain possible alternative models. Each of these actions is appropriate at any point in a data analysis, and all are likely to improve the final analysis and understanding.

4.1.9 Norms of EDA (Best Practices)

As noted earlier, EDA is a journey. The norms for conducting this journey are summarized in the sidebar titled "A GPS for Exploratory Data Analysis". You are encouraged to use this GPS wisely.

Section 4.1 References

Bacon, Francis. *Novum Organum*, John Gibson (Ed), Peter Urbach (Ed), (1620) 1994.

Hoaglin, D.C., Mosteller, F., and Tukey, J. W. *Understanding Robust and Exploratory Data Analysis,* Wiley, 1983.

Mosteller, F., and Tukey, J. W. *Data Analysis and Regression: A Second Course in Statistics,* Pearson, 1977.

Tukey, J. W., *Exploratory Data Analysis,* Pearson, 1977.

Tukey, J. W., and Wilk, M. B., "Data analysis and statistics: An expository overview," Proceedings of the November 7-10, 1966, fall joint computer conference, (1966): pp 695-709.

Velleman, P. F. and Hoaglin, David C., "Exploratory Data Analysis," APA Handbook of Research Methods in Psychology: Vol. 3. Data Analysis and Research Publication, H. Cooper (Editor-in-Chief), Chapter 3, (2012): pp 51-70.

Section 4.2 - Data Cleaning: Outlier Detection and the Role of Automated Algorithms

4.2.1 Objectives

In this section we define data cleaning, highlight five types of data cleaning problems and discuss methods for conducting data cleaning. The goal is to ensure that the available data contain as few problems as possible and are ready for analysis**.**

4.2.2 Outline

We begin by defining data cleaning. Next five types of "dirty data" are identified and discussed. Formal and informal methods of data cleaning are presented. The section concludes with a discussion of some best practices for data cleaning.

4.2.3 What is Data Cleaning?

We use "data cleaning" to refer to detecting and, when possible, correcting actual errors in the data you receive. This is the common statistical meaning of this phrase. However, this phrase has also been used in related fields, such as machine learning, to refer to other aspects of data handling, including such data preparation steps as data reshaping.

In this article, we focus on detecting potential errors. Whether a potential error is an actual error, and, if so, whether the actual error can be corrected is a matter not addressed here.

Data cleaning is important. In real data sets it is not uncommon to find errors in the original data. This is particularly true for observational data, i.e., for data collected by simply observing the process or phenomena being studied. How much might these errors affect your conclusions and recommendations? This is, by its nature, unknown at the start of an investigation. However, prudence requires that data cleaning be done as early and as well as possible in the initial stages of data analysis.

Data cleaning is discussed from the perspective of a Statistical Engineer (SE), so we consider data-cleaning strategies for the kind of data that SEs are most likely to examine. For example, we will not discuss ideas of cleaning data from relational databases, where the cleaning may include checks for misspellings, incomplete records, wrong or missing addresses, etc. Instead, our primary focus will be on numeric data and relatively simple attribute data.

We will also not discuss the important idea of how to improve your data sources so that data cleaning can be reduced in the future. But even if we did, many SEs may be working with so many different clients that this may not be practical.

We end this subsection with two general notes on data cleaning. First, a very unusual value that is detected may be correct, while a value that does not stand out may be incorrect. In particular, any unusual values that are found to be correct, or even simply not found to be incorrect, would still normally be included in the data set. It is in your analysis where decisions will be made on how to handle them. For example, if your analysis includes sensitivity checks and you find that your conclusions are not affected by the unusual values, then it is simpler to retain them in your analysis.

Second, by its nature, data cleaning is open-ended. So, no matter what set of pre-determined rules you may follow, you are not guaranteed to have clean data. Our approach here is to look at some common questions whose answers may make data cleaner. In all cases, we assume that the data have been prepped so that each row, or record, contains a rational unit of data (see Chapter 3).

4.2.4 Five Types of Dirty

Data that is not clean is often called "dirty data." There are different taxonomies of dirty data, but here we consider such data to be in one of the following categories, all of which we will examine in more detail:

1. Logical errors
2. Inconsistent, but not outlying, values
3. Data duplication
4. Missing values
5. Outliers

The last category involves a more technical topic, so we devote a separate section to it.

4.2.4.1 Logical Errors

These errors are also called violations of integrity constraints. These errors may occur within-field or field-to-field, and can best be seen with examples:

- Are the data consistent within each field for the explicit or implicit restrictions in that field? Here are three examples:
    a. A field for gender may have four possible options (here, coded for simplicity): 1 = male, 2 = female, 3 = identify as other, 4 = choose not to answer. Any values not in {1, 2, 3, 4} indicate non-clean data. Other examples of restricted data values would be "integers between 1 and 10" and "non-negative numbers."
    b. Consider a date field, which we suppose is read in as text values. If the format is required to be MM/DD/YY, then checks should be made that this format exists in each record; that MM is in 01 to 12; that if MM=02, then DD is in 01 to 28 for years not divisible by 4 or years divisible by 100 but not 400; etc.
    c. Assuming that date fields have been transformed from text to numeric values and include DOB (date of birth), is DOB earlier than the current date?

Depending on the source of the data, these checks may have already been made—but errors can still arise in transmission, manual adjustments, etc. Similarly, checks of dates may not be needed if you plan to convert dates as text to a standard numeric format, such as in some software—as long as you know that the software will perform the checks correctly, and you know how it handles text that do not correspond to valid dates.

As implied above, data values that are not clean *should not automatically be changed or set aside*. Instead, all such values should be checked, ideally by a person who has either created or sourced the data set. This also means that the less you know about the pedigree of your data, the more difficult it will be to check such values. Assessing the data pedigree is discussed in Chapter 3, Section 3.

- Are the data consistent from field to field? This can occur when a record of data contains several fields that have restrictions connecting them. Here are three examples:
    a. Some data sets might include dates and corresponding weekdays. Are the weekdays consistent with the dates?
    b. Again assume that date fields have been transformed from text to numeric values. If dates for High School and College exist in a record, does the first date precede the second by at least, say, two years?
    c. If odometer readings are given for a vehicle on more than one date, do the odometer readings ever decrease as the dates increase? In a process that involves boiling off water in a vat, and vat weights are recorded every 10 minutes, do vat weights ever increase?

  Note that these checks may detect data that are unusual, but it does not mean that the data values are incorrect. For example, occasionally a student graduates from college very soon after high school. Recorded vat weights may increase because of a poor measurement system or because water was added to the vat between readings. As noted above, whenever possible such values should be checked back to their source.

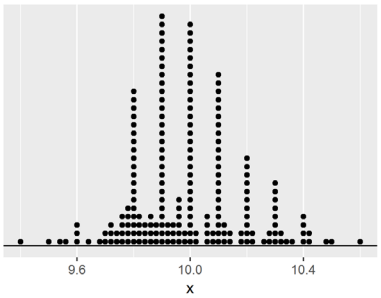
4.2.4.2 Inconsistent, but not Outlying, Values

Do non-outlying values in a field seem inconsistent with each other? Again, this idea is best seen with examples. In all cases we consider data on a process whose output is intended to be the same. For this reason, we rely on tools that are designed to reveal patterns in the one-sample case: dot plots, histograms, line plots and boxplots. Whenever possible and reasonable, one should first make plots of the data versus other features for which non-random patterns might present themselves. A key such feature is time, where control charts may be used, so that the time-collapsed versions we show are not misleading. For simplicity of presentation, we do not do this.

For smaller data sets in which there are only one or a small number of groups, a dot plot may reveal the most information. Here are two examples:

a. *Inconsistent rounding*. Consider a set of 200 measurement from a process for which the dot plot shows the following pattern:

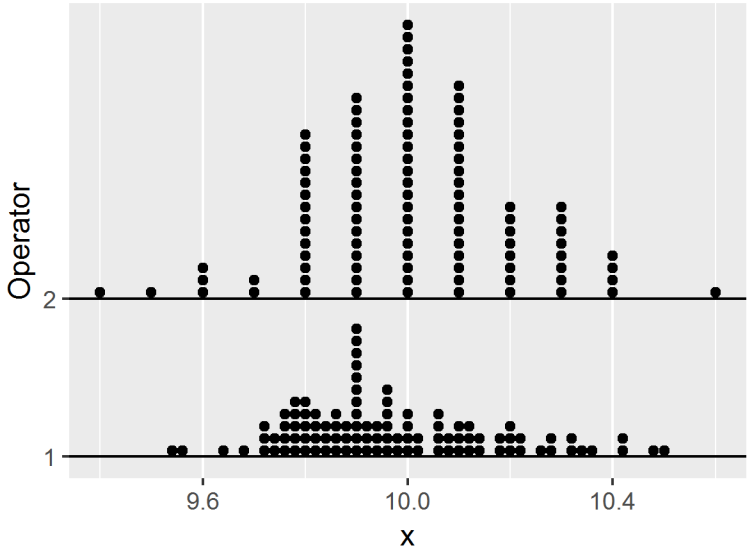
Figure 4.2 Dot plot of 200 measurements from a process.



There seems to be an unusually large number of points every 0.1 units, but there are also many other points.

The SE knows that the data were collected by two operators, so another graph was made in which the data were split by operator:

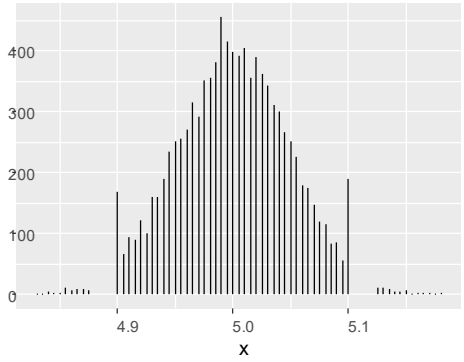
Figure 4.3 Dot plot of 200 measurements from a process, by operator.



This shows that the second operator's values were all rounded to the nearest 0.1 units. Follow-up work revealed that both operators recorded their results to the same number of decimal places and exported their data to CSV (Comma Separated Values) files, but the second operator allowed the software to round results in the exported file. A corrected file was sent to the SE to make the data consistent.

b. *Inconsistency of measurements near specification limits*. It is not uncommon for measurements to appear to be unusual near the specification limits of a product—one reason is that a small numerical difference may result in a large monetary difference, so flinching may occur. Here is an example from a high-data-volume product, whose specification limits are 5.0 ± 0.1.

Figure 4.4 Line plot of inconsistency measurements near specification limits.



You can see that most of the data are visually consistent with a bell-shaped curve. However, (a) there is an unusually large number of readings at the specification limits and (b) there is a gap of no readings in the region where the product would have been slightly out of the specifications. The SE would need to find out how such values arose and decide what to do next.

4.2.4.3 Data Duplication

Are there duplicated data records? Although it is possible that such duplication occurs by natural circumstances, it is usually more likely the data records themselves were accidentally duplicated.

This is more likely to occur when data sets are sent from multiple sources or from the same source multiple times. When either of these occur, it is important to include information in the data about each source, for example by creating another field that identifies the source or time uniquely. In that way, a search for duplicate records that reveals duplicates is more likely to point to the source of the problem.

Here is a simple example. An experiment was replicated five times, and the results of each replicate were stored in a separate file. It was found that these five files included duplicate records for the last two replicates—it turns out that the supplier of these files accidentally sent two duplicates of the fourth replicate instead of the fourth and fifth replicate.

In practice, complications may arise. Consider the five-replicate example again, and suppose each file contains six fields: the levels of each of the factors A, B, C, D; the level of the blocking factor, Time and the response Y. If the last two replicates were identical in all fields, it can be relatively easy to check for corresponding identical records. But if the last two replicates were identical except for the blocking level, one would still suspect that there is a duplication of records because the Y values were identical in both files. For this reason, good checks for data duplication may require some subject-matter knowledge.

4.2.4.4 Missing Values

If a value is known to exist, but only a missing-value code is given in the data for it, we are certain that there is an actual error. We naturally want to correct it. But if it cannot be corrected, what action can we take to try to make the data cleaner?

This is an important question, but treating this in any depth is beyond the scope of this Section. Indeed, entire books have been devoted to this subject, e.g., Little and Rubin (2002). Here, we simply note that when missing values cannot be corrected in the data-cleaning stage:

- Data may be missing for many reasons that have implications for how the situation is addressed.
- There are many statistically based imputation methods that can be used to estimate missing data. Which method to use in a given instance depends to some degree on why the data are missing.

4.2.5 Outliers

Hawkins (1980) provided a good definition of an outlier as "an observation which deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism." This definition points out two essential elements of an outlier: first, it must be seen as unusual from the other data points (or most of them—the data may have more than one outlier); and second, it must be seen as unusual compared to some reference mechanism.

We consider only the case where the observations are numeric and are such that a normal distribution may be considered a reasonable reference mechanism. This is the most common scenario in which outlier detection is used, although sometimes other approaches would be more suitable. Also, we once again only consider data on a process whose output is intended to be the same (one-sample case), and for which there is no evidence of time-varying or other non-random behavior.

4.2.5.1 Formal Methods

In this Subsection, we discuss two formal methods. The first, and one of the more common formal methods of outlier detection in the one-sample normal case, was popularized by Grubbs (1950, 1969), who considered several scenarios of detecting outliers. In the simplest case, suppose one wants to detect at most one outlier. This outlier, if it exists, is thought to be unusually large. The corresponding Grubbs' test measures the distance from the largest value to the overall sample mean, standardized by the sample standard deviation. That is,

$$\frac{x_{[n]} - \bar{x}}{s},$$

where $x_{[n]}$, the n^{th} order statistic, is the largest value in the sample, and $\bar{x}$ and s are the sample mean and standard deviation, respectively.

However, in outlier detection many other cases also have merit:

1. A two-sided test of one outlier, i.e., a test to detect either an unusually small or large value.
2. One-sided tests of two large outliers, or two small outliers, or one small and one large outlier.
3. Generalizing the previous item, one-sided or two-sided tests of exactly k outlying values, where k must be specified in advance.
4. One-sided or two-sided tests of at most k outlying values, where again k must be specified in advance.

This listing of possibilities indicates a fundamental difficulty with constructing outlier tests when we wish to preserve the α-level of the test—considering outliers is fundamentally a data-driven method, but the formal tests, if used as intended, require us to state one-sided vs. two-sided, small or large, and the value k of outliers to consider detecting before we examine the data. Suppose we select k poorly. Then we are subject to the possibility of *masking* or *swamping*. Masking occurs, for example, if we are only testing for one outlier, but two or more exist. This can affect the value of, say, the Grubbs test statistic above so that neither outlier can be detected. Swamping occurs if we specify, for example, $k \geq 2$ outliers when in fact there is only one; some tests declare either all k points or no points as outliers, and in any case the sensitivity of the test is reduced.

To address these issues reasonably, Rosner (1983) suggested the generalized Extreme Studentized Deviate (ESD) many-outlier procedure. (ESD is the basis for Grubbs' test.) In this method, one can test for up to a prespecified number *r* of outliers. This method is also the formal method that receives the highest recommendation from Iglewicz and Hoaglin (1993), who review six formal outlier tests based on the normal distribution.

The choice of *r* naturally plays an important role in this test. As they note, it is better to choose an *r* that is slightly too large to prevent masking. Using a predetermined value of the error rate α, the test proceeds in a sequential fashion where the steps shown here correspond to those in Iglewicz and Hoaglin (1993):

a. Compute $R_1 = \max_j |x_j - \bar{x}|/s$, the ESD in the full sample of size n, where $\bar{x}$ and s are the sample mean and standard deviation.
b. Find and remove the observation that maximizes $|x_j - \bar{x}|$.
c. Compute R_2 in the same way as R_1 but from the reduced sample of size $n - 1$.
d. Continue in this way until all of $R_1, R_2, \ldots, R_r$ have been obtained.
e. Using the critical values λ_i from the table in Rosner (1983), or as calculated as shown below, find ℓ, the maximum i such that $R_i > \lambda_i$. The extreme observations removed in the first ℓ steps are then declared to be outliers.

Note that in the last step, if we find that $R_1 < \lambda_1$, $R_2 > \lambda_2$, and $R_i < \lambda_i$ for $i = 3, \ldots, r$, then we declare that the first two values detected are outliers, even though $R_1 < \lambda_1$. In this way, masking is avoided.

The formula for λ_i is

$$\lambda_i = \frac{(n-i)t_{n-i-1,p}}{\sqrt{(n-i-1+t^2_{n-i-1,p})(n-i+1)}},$$

where $i = 1, 2, \ldots, r$, $t_{\nu,p}$ is the $100p$ percentage point of the t distribution with ν degrees of freedom, and $p = 1 - [\alpha/(2(n-i+1))]$. Rosner (1983) shows that this approximation is reasonably accurate for the entries given in his table and is very accurate when $n > 25$.

Using the vitamin E data set from Rosner's paper, with $n = 54$ observations, suppose that we decide that at most 10%, or $r = 5$, of the observations might be outliers, and that we use $\alpha = 0.05$ for our error rate. Then we obtain the following table.

Table 4.1 An example of the generalized ESD many-outlier procedure

i	n	Most extreme value	R	λ	$R > \lambda$
1	54	6.01	3.119	3.159	N
2	53	5.42	2.943	3.151	N
3	52	5.34	3.179	3.144	Y
4	51	4.64	2.810	3.136	N
5	50	-0.25	2.816	3.128	N

From this, we decide that the first three values are outliers.

Finally, note that even in the simple one-sample case with a one-outlier test, when the test statistic is beyond the critical value of the test, we can only logically conclude that there is either an outlier in the data or the data were not sampled from a normal distribution. For more details, see Beckman and Cook (1983).

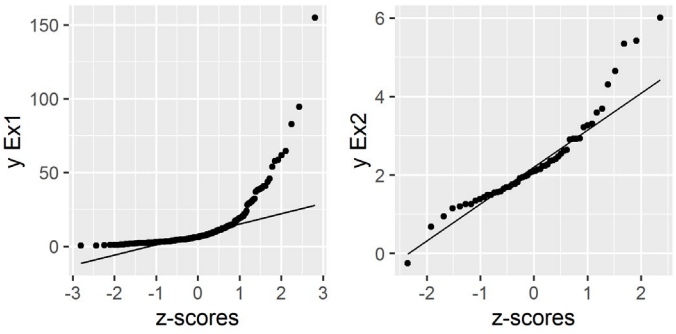
4.2.5.2 An Informal Method

Because of the inherent difficulty of formal "exact" methods—even Rosner's method still assumes that most of the data arose from a normal distribution—we also suggest that an informal, approximate, method be considered. Here are the steps of this method:

1. Examine the first so-called background assumption of the data, that the data do not exhibit time-varying or other non-random behavior.
2. If this first test is passed, graphically assess the extent to which the data (aside from outliers) appear to be consistent with a normal distribution. If this test is not passed, consider whether a transformation of the data, such as a log transformation, is both justifiable for this data and transforms the data so that this test is passed.
3. If this second test is passed, use a generally accepted robust test procedure to test for outliers.

Here are two examples where, for simplicity, we assume the first test has been passed. We start both examples with the second step, in which we construct normal-quantile plots (see Section 4.3.4.1). The Example 2 data are Rosner's vitamin E data from the previous Subsection.

Figure 4.5 Normal-quantile plots of two data sets.

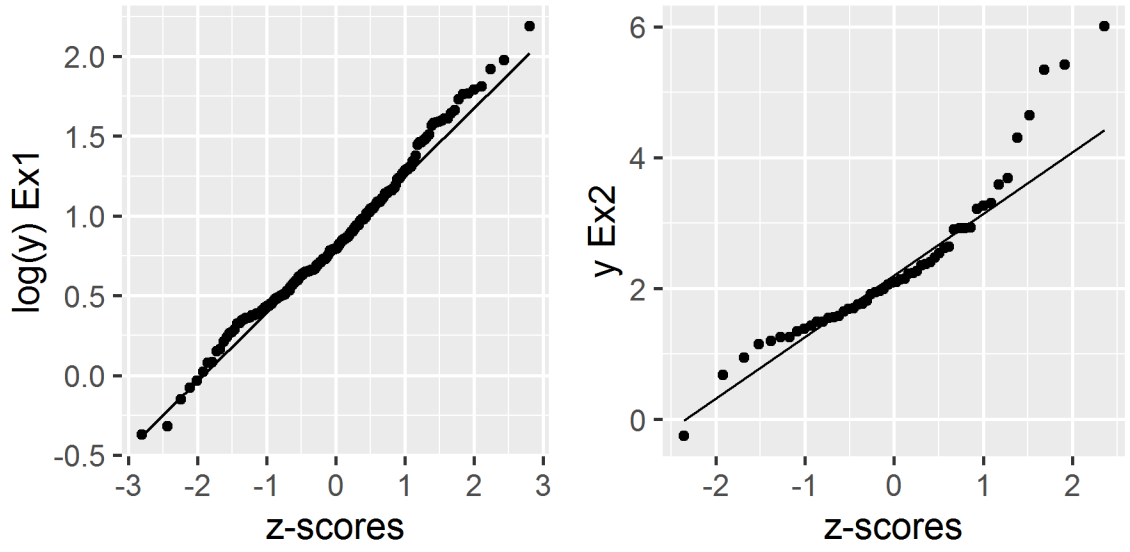


In the first plot, with sample size $n = 200$, we see what appear to be outlying values. However, in looking at the overall set of points, we also see what appears to be a smooth convex shape to the points. This suggests that a transformation to normality may be reasonable. In addition, all of the points are positive, suggesting that a log transform may be appropriate. (More technically, the data values should also be on a ratio scale for the log transformation to make physical sense, so this should be checked as well.)

In the second plot, with sample size $n = 54$, we again see what appear to be outlying values. However, in this plot it is harder to discern an overall smooth convex shape to the points. In addition, one point was less than 0 so (assuming this point itself was not incorrect), the typical transformations such as log, inverse, or square root are not appropriate. For these reasons, no transformation will be made.

The new plots are as follows:

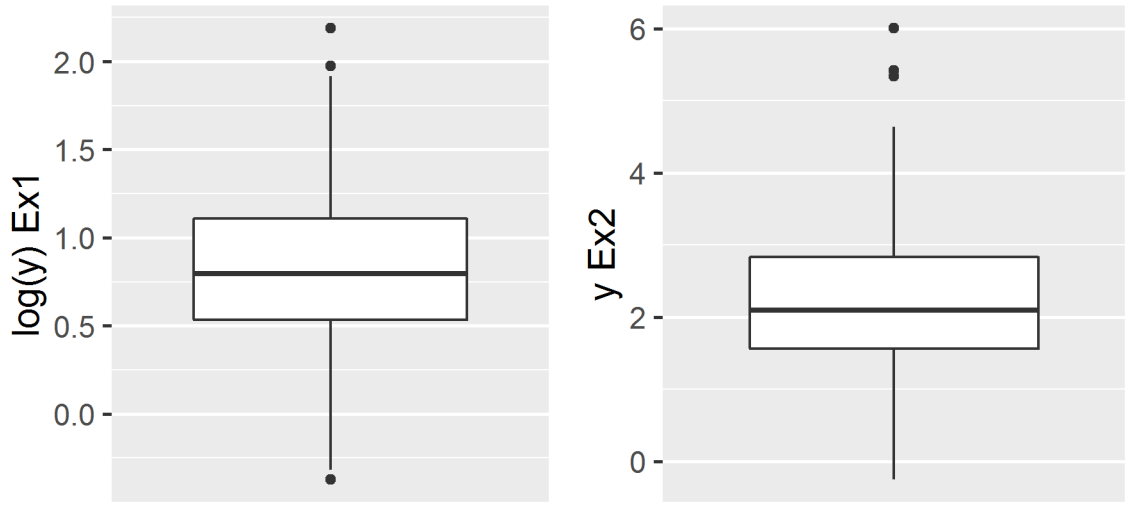
Figure 4.6 Normal-quantile plots of two data sets: first data set transformed.



The log transformation for the example-1 data appears to be effective in transforming the data to be consistent with normality. In addition, there do not appear to be any outliers in the transformed data.

We now proceed to the third step of using a generally accepted robust test procedure to test for outliers. We use the common default outlier test associated with box plots (see Section 4.3.4.1) of each data set:

Figure 4.7 Box plots of two data sets: first data set transformed.



The points on these plots indicate the outliers, so we see that three outliers are flagged for each data set.

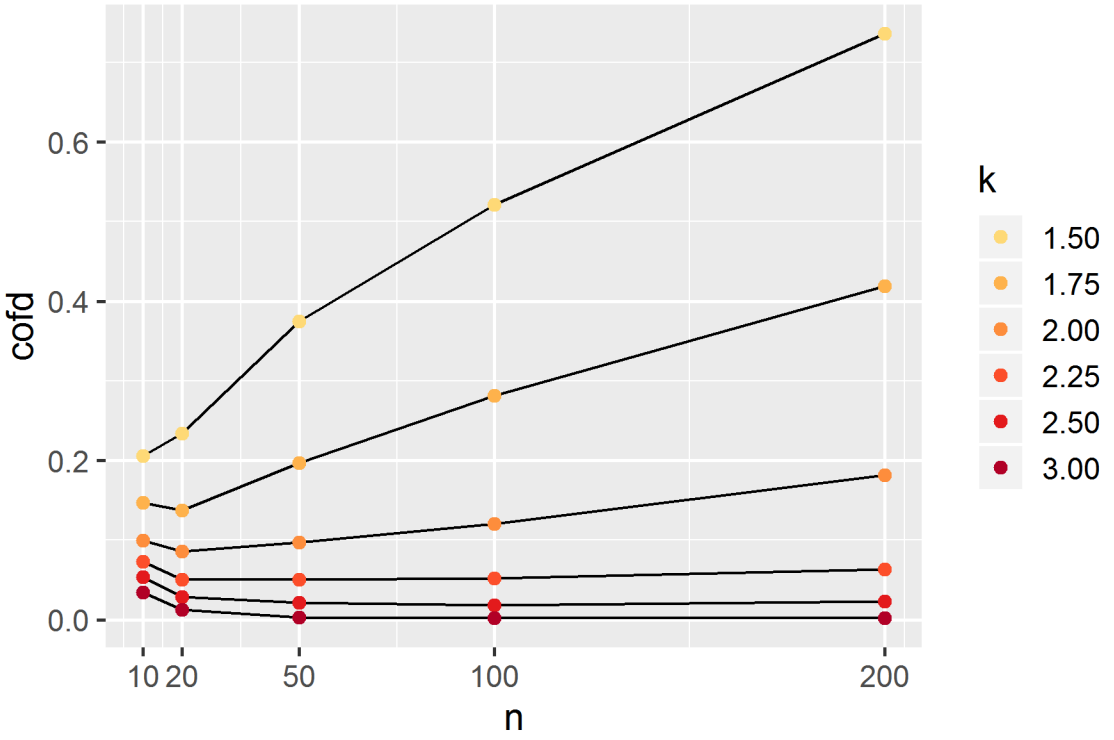
As a reminder, Tukey (1977) suggested that large outliers are flagged whenever a point exceeds $F_u + k(F_u - F_l)$, where F_u is the upper fourth (close or equal to the upper quartile), $F_u - F_l$ is the difference between the upper and lower fourth (the interquartile range), and k is a tuning constant. Small values are detected in an analogous way.

Here the default value of $k = 1.5$ was used. This value of k, aside from the variation created by estimating the fourths, corresponds for normally distributed data to a point outside of $\mu \pm 2.70\sigma$, which in turn corresponds to a probability of p=0.0070 *for any one particular point*. That is, the probability of seeing a particular point this extreme or more extreme entirely by chance is 0.0070.

This probability is quite low for one point, but in the first example, there are $n = 200$ such points. Aside from the variation in the estimates, this means that there is a $1 - (1 - p)^n$ chance that at least one point out of n exceeds this limit. For our values of p and n, this means there is about a 75% chance that at least one value will exceed the $k = 1.5$ limit erroneously when the data is from a normal distribution. We call such a chance the COFD, the Chance Of (one or more) False Detections

This suggests that $k = 1.5$ may be too liberal. The graph below shows estimated COFD values for various tuning parameters k and sample sizes n, where the COFD value of 0.75 found above is indeed reasonable, as it appears as the upper right point of the plot. (These estimates are based on simulations, where the variation in the estimates is taken into account, and the approximate standard error of the imprecision is $\sqrt{p(1-p)/n_{sim}}$, where $n_{sim} = 10{,}000$. So, for $p = 0.1$, for example, the s.e. is 0.003.)

Figure 4.8 The COFD values for various tuning parameters k and sample sizes n.



For the range k and n considered in the figure, we suggest that $k = 2.25$ should be considered to reduce the rate of false positives. Note that $k = 1.50$ is sometimes said to detect "mild outliers" and $k = 3.00$ is said to detect "extreme outliers."

Note in example 1 that the k value of 1.497 is slightly less than the $k = 1.5$ cutpoint but is still flagged as an outlier. How can this be? Well, in this note we are using Tukey's fourths, but the software used to generate these figures used quantiles. The slight difference led to a difference of one outlier. This emphasizes why the listing of the k values can help avoid the pass/fail dichotomous thinking of hypothesis testing. We also see that only the largest point in example 2, with a k value of 2.3, exceeds our suggested cut point, with milder evidence for the next two largest points. By comparison, Rosner's method detected all three of the largest points as outliers.

We hope that the preceding examples illustrate the importance of careful investigation of the data using human intervention, and the dangers that can arise by simply trying to institute a set of rules using automated methods that attempts to cover all cases.

4.2.5.3 Norms of Data Cleaning (Best Practices)

These are our recommendations for data cleaning.

1. **Reduce or eliminate errors at their source.**

   If a good proportion of your data come from repeat clients (suppliers of the data), then you have a better opportunity to improve data quality at the source. Working with your clients:
   - Learn where most data errors occur.
   - Learn about the root causes of these problems.
   - Put preventative measures in place to reduce or eliminate such problems. This includes standardized data entry and automatic checks.

2. **Create a standardized procedure for checks on data that you receive.**

   This section provides a good initial basis for such checks. As you continue to find additional problems in the particular kind of data with which you work, you will likely add checks.

3. **Consider special features of any particular data set on which you are working.**

   The notes above provide some examples. This particular practice requires you to be "close to your data," which has many benefits in addition to increasing your chances to find dirty data.

4. **Formalize communication with your clients.**
   - The client should know at the outset (written contract or plan of work, for example) that data checking and cleaning is a formal part of your work.
   - Contacts with the client on questions about the data should be formalized. For example, to whom should data-quality questions be addressed?

4.2.7 Notes

All examples but one in this chapter were based on real data sets, but the data shown was simulated for the sake of confidentiality. The example with actual data is from Rosner (1983).

See the NIST/SEMATECH handbook for other practical approaches for outlier detection.

Section 4.2 References

Beckman, R., and Cook, R. "Outlier.....s," *Technometrics*, 25, (1983): 119–149. doi:10.2307/1268541

Grubbs, F. E. "Sample criteria for testing outlying observations," *Annals of Mathematical Statistics*, 21, (1950): 27–58.

Grubbs, F. E. "Procedures for detecting outlying observations in samples," *Technometrics*, 11, (1969): 1–21.

Iglewicz, B. and Hoaglin, D. "Volume 16: how to detect and handle outliers," The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor, 1993.

Little, Roderick J. A., Rubin, Donald B. *Statistical Analysis with Missing Data* (2nd ed.), Wiley, New York, 2002.

Hawkins, D. *Identification of Outliers*. Chapman and Hall, London, 1980.

NIST/SEMATECH e-Handbook of Statistical Methods, https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm, accessed 2019/09/09.

Rosner, B. "Percentage points for a generalized ESD many-outlier procedure," *Technometrics*, 25, (1983): 165–172.

Tukey, J. W. *Exploratory Data Analysis*. Addison-Wesley, Boston, 1977.

Section 4.3 - Using Exploratory Data Analysis

4.3.1 Objectives

To show how Exploratory Data Analysis (EDA) can be used to understand data prior to doing formal statistical analyses such as creating tests of significance, creating confidence intervals, developing statistical models, etc. EDA helps the analyst to be alert to unexpected patterns, relationships, and extraordinary cases.

4.3.2 Outline

In this section we describe a few tools that are generally useful for doing EDA. We start with simple descriptive statistics that aid in characterizing the location and spread of the data. EDA is inherently graphical, so in the next section well known and widely useful graphical methods are discussed including the "Magnificent Seven" graphical tools. Next, the use of scatter plots in the visualization of relationships between variables is discussed.

Special attention is paid to the use of time plots in studying the variation in the data over time. This often leads to the identification of important causal variables. Assessment of the stability of the underlying process is also a critical outcome of the analysis.

The section concludes with a summary of the norms (best practices) one should keep in mind as EDA is used to explore data and better understand the process that generated the data.

4.3.3 Descriptive Statistics

One of the first steps in the exploration of data is to get an analytical summary of the data. Some commonly used statistics are shown in Table 4.2. These statistics help the analyst assess the location (mean/average, median) and spread (standard deviation, range, coefficient of variation, RSD) and size of the sample. Further discussion of these statistics can be found in Hoerl and Snee (2020).

4.3.4 Graphical Methods – Discovering the Unexpected

When most people hear the words *data analysis*, especially in the context of statistics, they probably think of "number crunching," mathematical formulas, and algorithms. While these are certainly part of analysis, a key element often overlooked is *visualization* through graphical analysis. The need for, and use of, graphical analysis is a theme of this section.

As you read through Chapters 1, 2 and to this point in 3, you may have noticed a number of graphical displays used to make important points. Indeed, a strategic framework for analyzing data has three critical elements: the practical, graphical and analytical. Here we focus on the graphical element of the framework, explaining why it is critical.

Table 4.2 Commonly Used Descriptive Statistics

Statistic	Interpretation	Value for Data in Table 3.3.2
Mean (average)	Central Value, Location of the distribution	10.57
Median	50^{th} percentile, The value above and below which 50 % of the distribution on the data lie.	10.50
Standard Deviation	Measures the spread of the distribution. For a normal distribution 95% of the distribution lie within the average +/- 3 standard deviations.	1.25
Minimum Value	Smallest data value	8.0
Maximum Value	Largest data value	13.0
Range = Max - Min	Difference between the maximum and minimum data values. Measures distribution spread	5.0
Coefficient of Variation (RSD)	100(Standard Deviation)/Average. Measures the variation relative to the average value. Also referred to as the Relative Standard Deviation (RSD) in some fields.	11.8%
Sample Size	Number of data points in the sample	15

Table 4.3 Sample Data to Illustrate Summary Statistics in Table 3.2

10.0	12.0	9.0	10.5	13.0	9.5	10.7	8.0	10.0	11.7	10.5	10.8	10.7	11.8	10.4

When thinking about the value of graphics (pictures of data), John Tukey perhaps said it best:

> "The greatest value of a picture is that it forces us to notice what we never expected to see."

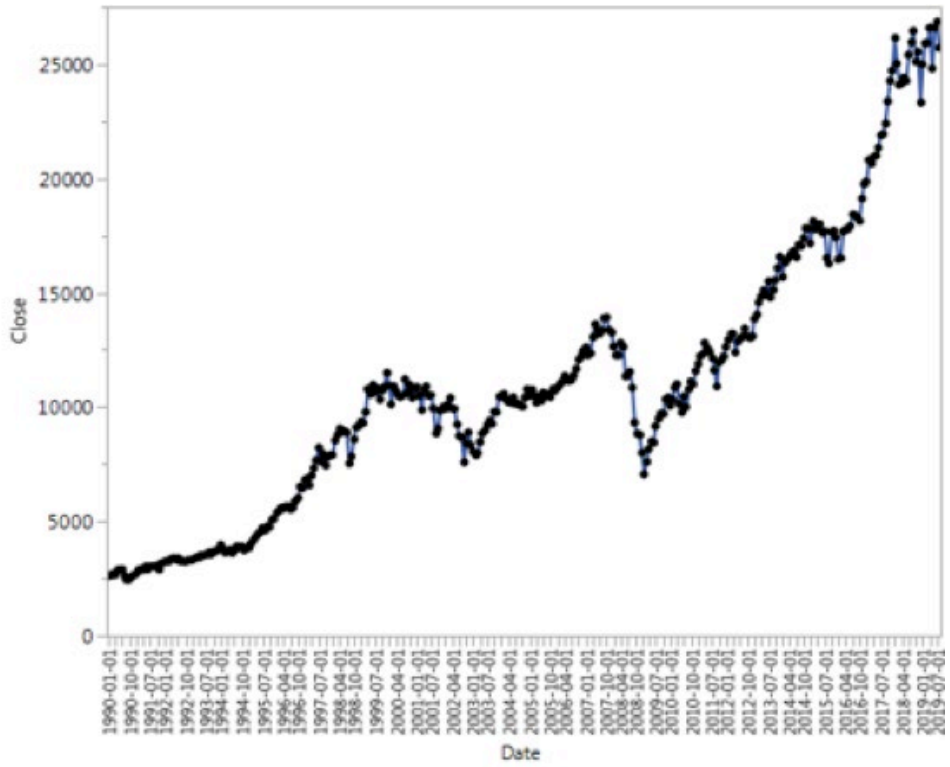
That is, if we only perform calculations, based on our existing subject matter knowledge, we may miss an obvious pattern in the data that could enhance our understanding of the process of interest. For example, someone calculating statistics on the Dow Jones Industrial Average might totally miss the sudden decline in 2008/2009 without looking at a plot of the data over time (Figure 4.9).

Graphics provide pictures that have many uses in the exploration, analysis, display and communication of data, including the following:

- **Communication of results.** We have all heard the old saying "A picture is worth a thousand words." Graphics enable the communication of volumes of information and complex relationships quickly and clearly.
- **Stimulate Insights.** As Tukey points out, graphics enable you to see the unexpected which in turn enables the development of new theories and innovative ideas.

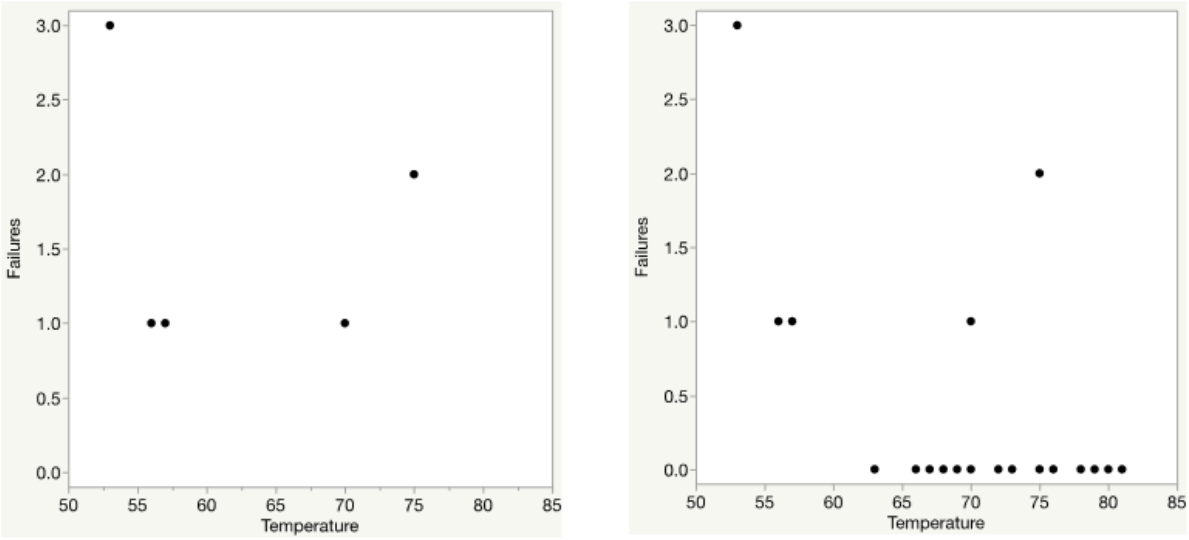
- **Identify large effects.** Graphics identify the presence of large effects that enable us to develop "ballpark" answers and compare the size of the effects from a practical perspective.
- **Provide a check on statistical analysis.** Large effects or differences seen in graphics will almost assuredly show up in the formal statistical analysis. If not, there is likely something wrong with the analysis. That is, we are using the wrong statistical model.
- **Condense and summarize data.** Patterns in data are easily seen in graphical displays, even when large volumes and complex relationships are involved. Visualization is a core element of modern Big Data Analytics (data science).
- **Provide insight into complex mechanisms.** Our eyes can see patterns in data plots, even if we are unable to develop models to quantify them.
- **Increase the probability of finding a useful solution.** Patterns seen in graphics spark our curiosity and imagination, causing us to spend more time thinking about the problem and how the process works. Critical thinking about the process usually leads to better solutions.

Figure 4.9 Dow Jones Industrial Average Monthly Closing Prices: 1990 – 2019



Ron Snee, a contributing author of this Handbook remembers that his professor, Ellis Ott, admonished students at Rutgers University to "plot the data and think." It is clear from the plot of the Challenger Space Shuttle data (Figure 4.10) that the launch temperature of 31° F was well below the temperature of any previous launch, and well below the temperature of any successful launch with respect to O-ring failure.

Figure 4.10 Challenger Data: Number of O-Ring Failures/Launch Versus Temperature – Left Panel: Five (5) launches that Had O-Ring Failures; Right Panel: All Launches With and Without O-Ring Failures



If someone had constructed Figure 4.10 and actually looked at it prior to the launch, seven lives would likely have been saved.

Real estate agents often joke that the three most important factors in determining the price of a home are: location; location; location. When it comes to graphical analysis of data, there are only two rules to remember:

1) Always plot the data
2) When in doubt, refer to Rule #1.

More than a few people have recounted a time when they were unable to resolve a problem at work, but that a solution suddenly came to them while driving or bicycling home that evening. While this is not a psychology textbook, there is a theory that most analytical work is performed on the left side of the brain. However, two-dimensional positioning, which people do while driving or bicycling in traffic, utilizes the right side of the brain. In many cases, using a different part of the brain stimulates a new way of thinking about the problem, often resulting in a novel solution. Similarly, some have found it helpful when faced with a statistical analysis that is not working, to rely totally on graphics to look for a new approach to analyzing the data. As a result, people are more fully engaging the right sides of their brains and giving the left sides a rest.

Leonardo da Vinci stated about his art that "Simplicity is the ultimate sophistication." We believe that this applies to graphical analysis as well. A good graphic should be clear and easy to

understand. We also note that the graphic should be understandable by both the presenter and potential users of the graphic, who may have very different backgrounds. The presenter must clearly understand the graphic in order to present it clearly to the user in written or oral form.

4.3.4.1 The "Magnificent Seven"

There are many different types of graphical displays that one can use in the analysis of data. Popular graphs that arguably get the most frequent use are known as the "Magnificent Seven." Included is this group are: Histogram, Dot Plot, Box Plot, Normal Probability Plot, Scatter Plot, Pareto Chart, Time/Sequence Plot and Control Chart. The uses of these plots are summarized in Table 4.3. This is a typical set of seven. Sometimes analysts will add others to the list just as the Big Ten College Football Conference has 14 teams in it now. The point is that there are a group of graphical displays that are very effective and are used frequently.

Data sets are typically analyzed using several of these graphical tools as well as other graphical displays. The different tools and variety of important variables identified suggest different models for the phenomena being studied. In the process our minds are iterating between different models for the data set.

Table 4.3 Magnificent Seven Graphical Tools

Graphical Display	Uses	Illustrative Figure
Histogram and Dot Plot	Assess the location, spread and shape of the data distribution	3.3.3
Normal Probability Plot	Compare the data distribution to the normal distribution	3.3.4
Run/Time/ Sequence Chart	Study data variation over time or sequence of collection	3.3.5
Control Chart	Run/Time/Sequence Chart with upper and lower limits to detect non-random patterns of variation in the data	3.3.6
Box Plot	Compare Groups of Data	3.3.7
Scatter Plot	Study relationship between two variables	3.3.8, 3.3.9, 3.3.10
Pareto Chart	Study relationship between frequency of occurrence of an event and the causes of the events	3.3.11

The Gasoline data in Table 4.4 will be used to illustrate the first five plots in Table 4.3. First some comments on the pedigree of the gasoline data. The auto that generated these data was used to commute to work and to run errands on weekends. It was operated by a single driver. The tank was filled up to the top each time gasoline was purchased. The time period covered was approximately four years. The miles/gallon was recorded for each fill up. For reference the summary statistics for these data are summarized in Table 4.5.

Table 4.4 Gasoline Mileage Data

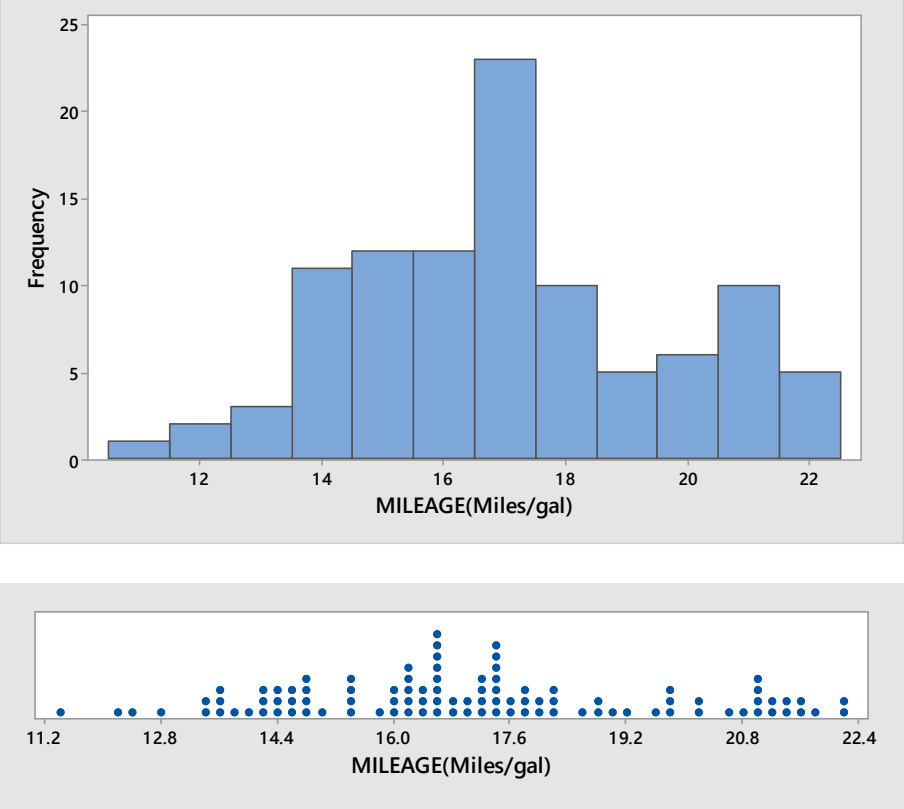
FILL UP	MILEAGE	FILL UP	MILEAGE	FILL UP	MILEAGE	FILL UP	MILEAGE
1	14.6	26	14.6	51	13.6	76	14.8
2	22.2	27	12.8	52	14.5	77	15.4
3	16.1	28	14.1	53	14.3	78	16.1
4	17.1	29	13.4	54	16.5	79	14.1
5	18.1	30	15.9	55	16.3	80	14.2
6	15.8	31	16.1	56	17.4	81	17.5
7	17.0	32	16.6	57	17.4	82	16.5
8	17.1	33	15.9	58	17.4	83	14.7
9	15.3	34	16.6	59	19.1	84	17.3
10	16.3	35	17.3	60	19.5	85	18.2
11	15.3	36	18.7	61	20.9	86	21.1
12	15.4	37	16.9	62	20.7	87	19.7
13	13.5	38	16.8	63	20.9	88	20.9
14	18.5	39	18.7	64	21.3	89	19.0
15	16.1	40	17.3	65	22.2	90	21.5
16	13.9	41	17.1	66	21.5	91	20.2
17	12.4	42	17.7	67	19.8	92	20.6
18	13.8	43	15.9	68	20.1	93	21.3
19	11.4	44	16.6	69	21.2	94	20.9
20	13.3	45	16.4	70	19.8	95	21.8
21	12.1	46	15.0	71	17.9	96	18.1
22	14.3	47	16.8	72	17.7	97	17.7
23	14.7	48	13.5	73	16.6	98	17.3
24	14.4	49	16.6	74	17.6	99	16.2
25	14.7	50	16.6	75	17.2	100	18.0

Table 4.5 Gasoline Mileage Data – Summary Statistics

Variable	Mean	StDev	CoefVar	Minimum	Median	Maximum	Range
MILEAGE(Miles/gal)	17.018	2.540	14.92	11.400	16.800	22.200	10.800

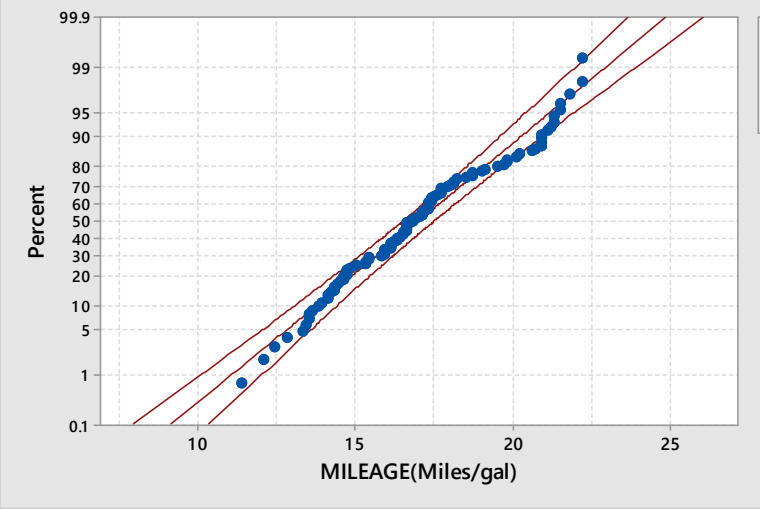
The Histogram of the gasoline data (Figure 4.11) shows that the MPG varies from approximately 11 – 22 with a center location around 16 – 17. The shape of the distribution appears to have two modes (humps) around 16 and 21. The Dot Plot supports these conclusions.

Figure 4.11 Gasoline Mileage Data - Histogram and Dot Plot of Miles/Gallon



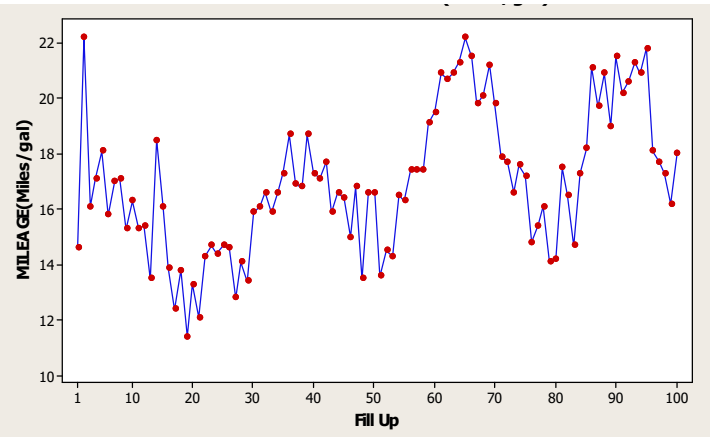
The probability plot of the mpg data (Figure 4.12) shows some departures from the normal distribution consistent with the multiple mode conjecture.

Figure 4.12 Gasoline Mileage Data - Normal Probability Plot of Miles/ Gallon with 95% Confidence Limits



A major limitation of the histogram, dot plot and probability plot is that an important variable is ignored, TIME; the time sequence in which the mileage values were determined. By plotting MPG versus time, we can identify clues as to which variables may be causing the variation in MPG (Figure 4.13).

Figure 4.13 Gasoline Mileage Data – Time Plot of Miles/Gallon



The Time Plot of the data uncovers two important non-random patterns in the data.

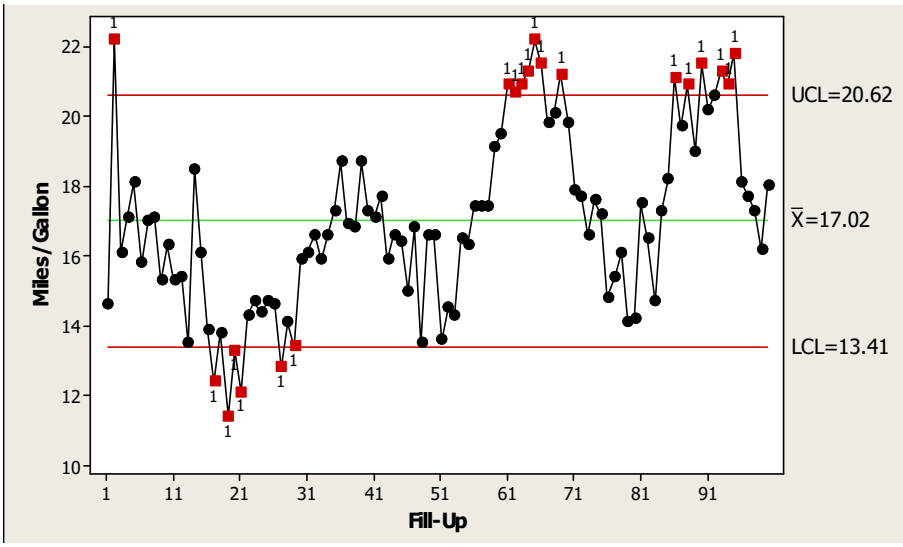
- Reoccurring cycles in the data
- Upward shift or trend beginning around fill up #50.

These patterns in the data lead us to ask whether the patterns are real or due to noise in the data. The control chart can help answer this question.

The **Control Chart** is a time/run chart with upper and lower control limits typically set at the average +/- 3 standard deviations of the data (Montgomery 2013). In Figure 4.14 we see that there are several points outside of the control limits indicating that the non-random variations in

the data are real (statistically significant) and not due to random variation (noise).

Figure 4.14 Gasoline Mileage Data - Control Chart for Miles/Gallon



The Time Plot and the Control Chart uncovered two important non-random patterns in the data.

- There are reoccurring cycles in the data
- There is an upward shift or trend beginning around fill up #50.

The obvious question to ask is "what could be the sources of this variation"? To answer this question, we return to the "pedigree of the data." We saw earlier that the data were collected over a four year period. When the fill up dates were compared to the MPG values it was clear that the high values were in the summer and the low MPG values were in the winter. Thus, the cycles were due to a seasonal effect. Table 4.6 shows the fill ups for the year-season combinations.

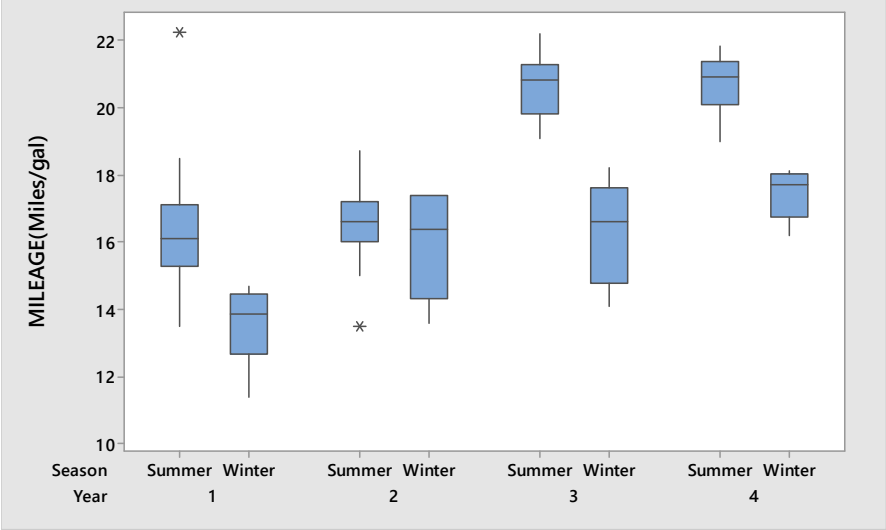
Table 4.6 Gasoline Mileage Data – Fill-Up Codes for Year – Season Combinations

Year	Summer	Winter
1	1 - 15	16 - 30
2	31 - 50	51 - 59
3	59 - 70	71 - 85
4	86 - 95	96 -100

The next question is what causes the level shift or trend? On investigation it was learned that the auto underwent a major repair. The result being an increase in gasoline mileage among other things.

The Box Plot is a good way to show the yearly and seasonal variation. In Figure 4.15 we see the upward trend and the cycles with the winter MPG being lower than the summer MPG.

Figure 4.15 Gasoline Mileage Data - Box Plot for Year and Season Combinations



Scatter Plots are used to display and study relationships between two variables, X and Y. In a scatter plot the X-Y pairs are plotted in cartesian coordinates (Y vs X). Some typical patterns, shown in Figure 4.16 are:

- Weak positive relationship
- Weak negative relationship
- Strong positive relationship
- Strong negative relationship
- Nonlinear (curvilinear) relationship
- No relationship

Figure 4.16 Potential Scatterplot Relationships

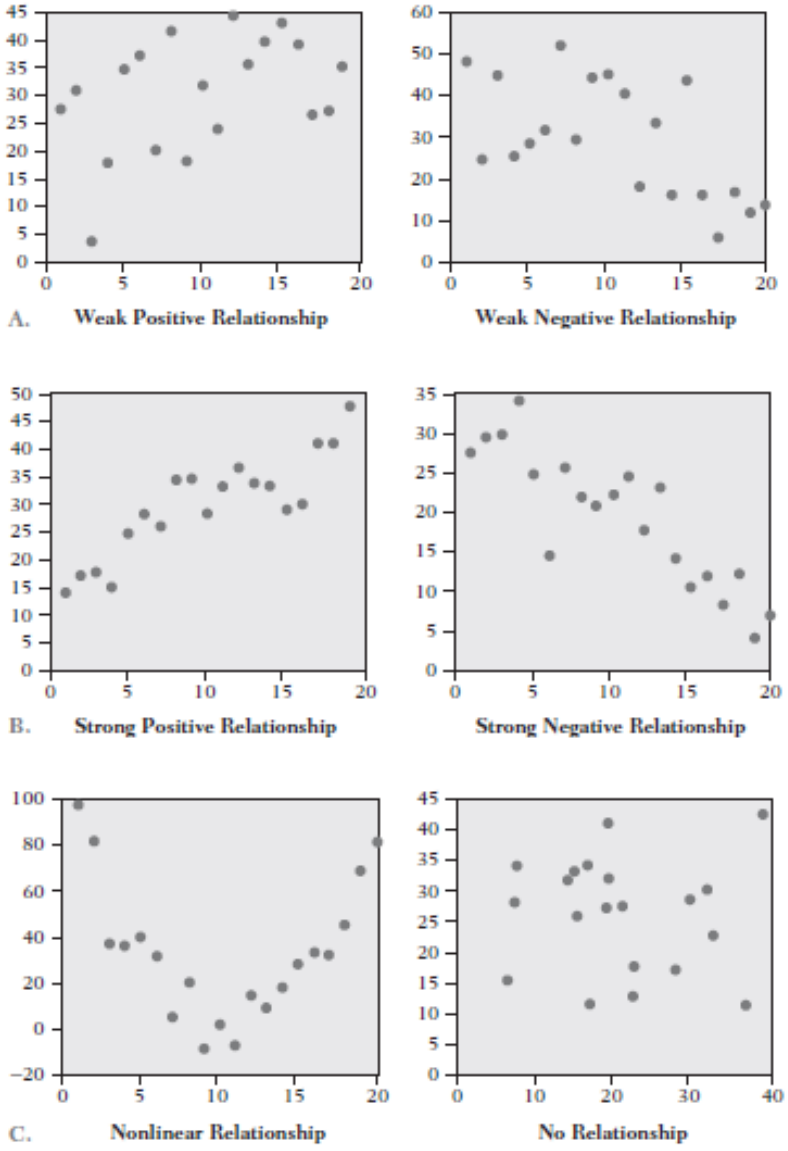


Figure 4.17 shows the relationship between MPG and weight of automobiles. As the weight of the car gets larger the the MPG decreases. Next we need to understand the fundamental basis for this relationship. In this case elementary physics tells us more energy is required to move a larger weight. Since larger cars typically weigh more and provide a smoother ride this plot shows that one is unlikely to get a smooth ride in a small (light weight) car. It is also important to note that experience has shown that weight is a dominant variable when studying auto characteristic variable effecting MPG. Other variables have an effect. but none have an effect as vehicle weight.

Figure 4.17 Scatterplot Example - Vehicle Miles/Gallon versus Vehicle Weight

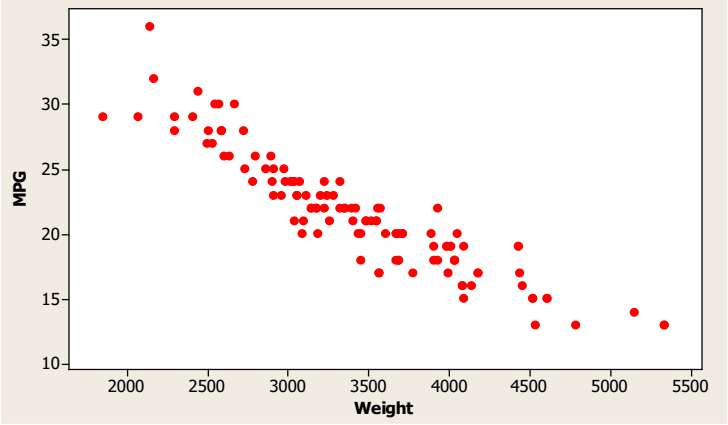
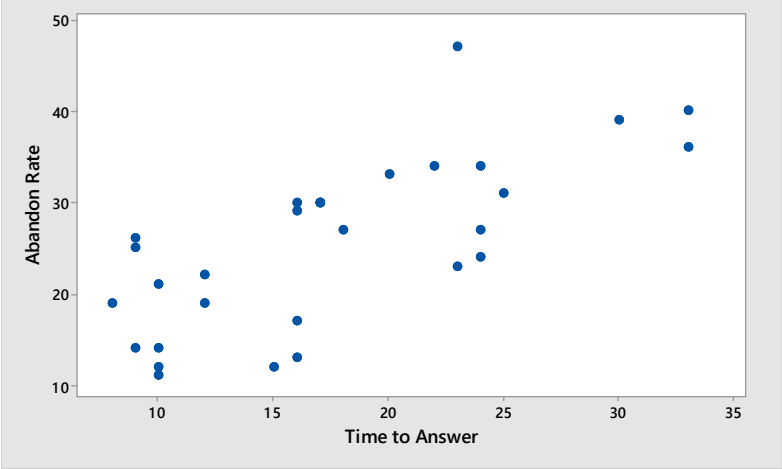


Figure 4.18 shows another example of the use of a scatterplot, call center call abandon rate versus time to answer. One would conjecture a positive relationship between these two variables as seen in Figure 4.18. In addition to the strong linear relationship, we see one data point that is atypical to the rest of the data. This relationship and other uses of scatterplot will be discussed further in Section 4.3.4.

Figure 4.18 Scatterplot Example – Call Abandon Rate versus Time to Answer



The Pareto Chart is very useful in studying the relationship between occurrence of events and the perceived causes of the events. In Table 4.7 we see a tabulation of data listing reasons why callers to a call center have to wait. Table 4.7 summarizes the data "Before" and "After" improvements were made.

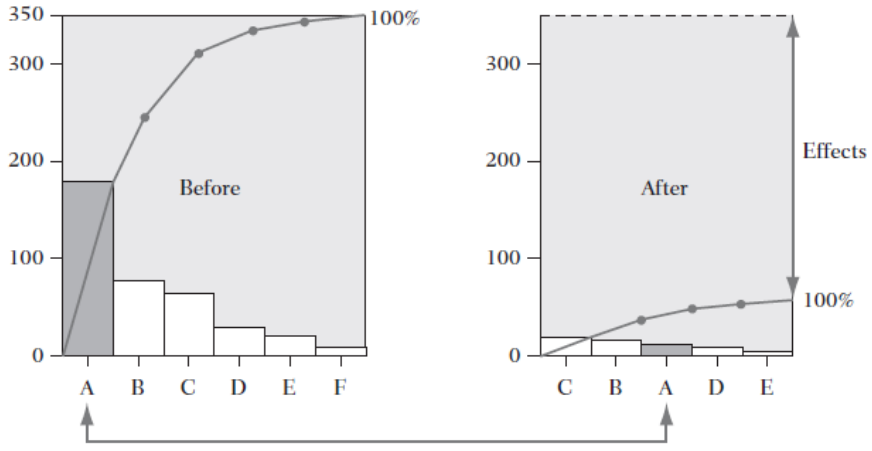
Table 4.7 Pareto Chart Example – Call Center Call Waiting Data

	Reasons why callers had to wait	Total number Before	After
A	One operator (partner out of the office)	172	15
B	Receiving party not present	73	17
C	No one present in the section receiving the call	61	20
D	Section and name of receiving party not given	19	4
E	Inquiry about branch office locations	16	3
F	Other reasons	10	0
	Total	351	59

The Pareto Chart displays the number of events (late answered calls) versus the cause of the event. The X-Axis of the plot is the cause in rank order with the largest cause shown on the left. In the analysis it is common to look for the "big bar on the Pareto" which indicates the cause with the largest effect. It is not unusual for a few causes to produce as much as 80% of the events. This observation has produced the "80-20" Rule: 80% of the events are caused by 20% of the causes. In figure 4.9 we see that Causes A and B account for approximately 70% of the events while Causes A, B, and C account for 87.2% of the events.

The first round of improvements focused on Cause A (one operator, partner out of the office) which accounted for approximately 49% of the events. We see in Table 4.7 and Figure 4.19, after improvements were made there was a significant drop in the number of events associated with Cause A.

Figure 4.19 Pareto Chart - Call Center Call Waiting Data



4.3.4.2 Visualization of Relationships between Variables

One of the most important activities when doing exploratory data analysis (EDA) is to look for relationships between variables. The useful result is the identification of candidate cause-effect relationships that will deepen understanding of the process that generated the data. The first step in such work should be to visualize the relationships. The critical tool of such visualizations is the scatter plot which was introduced earlier.

We will illustrate these uses using the data in Table 4.8 which provides data from a call center over a 30 month period on:

- X1 = Total Calls (thousands)
- X2 =Time to Answer (seconds)
- X3 = Virtual Response Unit (VRU) – Percent answered by VRU x 10
- Y = Abandon Rate (percent)

The relationships between these variables is visualized by three different types of scatter plots: Simple Y versus X, Y and X versus Time and X versus X.

Figure 4.20 shows plots of the response Y versus the three predictor variable X1, X2 and X3. In the figure we see positive relationships between Abandon Rate and Total Calls (call volume) and Time to Answer. VRU has a negative relationship with Abandon Rate. The relationship between Abandon Rate and Call Volume is the weakest of the three relationships.

Table 4.8 Call Center Data

Month	Total Calls	Time to Answer	VRU	Abandon Rate	Month	Total Calls	Time to Answer	VRU	Abandon Rate
1	135	12	142	22	16	172	17	261	30
2	150	8	156	19	17	174	16	278	17
3	109	9	187	25	18	165	12	272	19
4	109	9	178	26	19	177	9	290	14
5	128	18	189	27	20	183	10	304	21
6	125	17	148	30	21	179	23	282	23
7	124	17	155	30	22	191	33	323	40
8	141	20	200	33	23	170	33	292	36
9	133	22	175	34	24	174	24	289	27
10	153	24	192	34	25	187	24	281	24
11	155	25	177	31	26	198	10	290	11
12	294	30	108	39	27	161	10	309	14
13	295	23	132	47	28	150	15	302	12
14	305	16	165	29	29	149	16	283	13
15	221	16	251	30	30	139	10	302	12

Figure 4.20 Call Center Data – Scatterplots of Abandon Rate versus Total Calls, Time to Answer and VRU

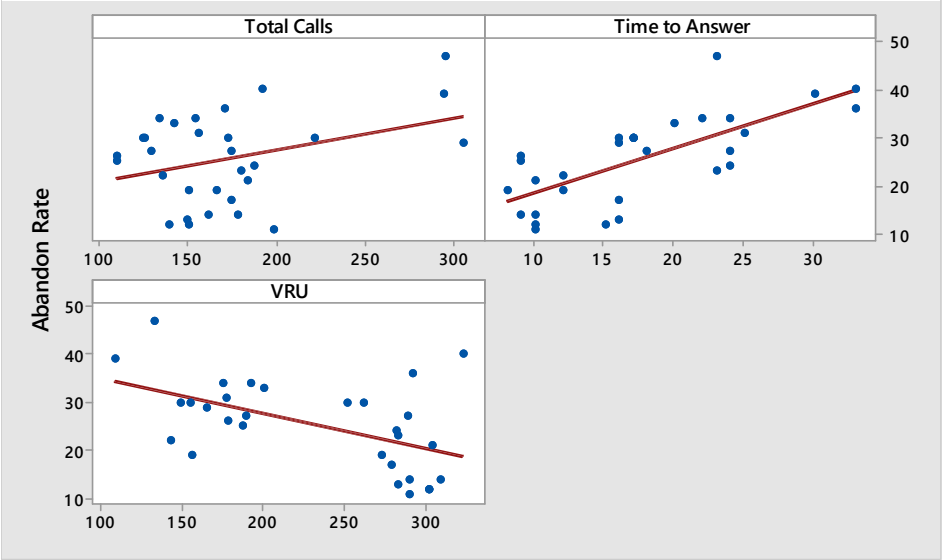
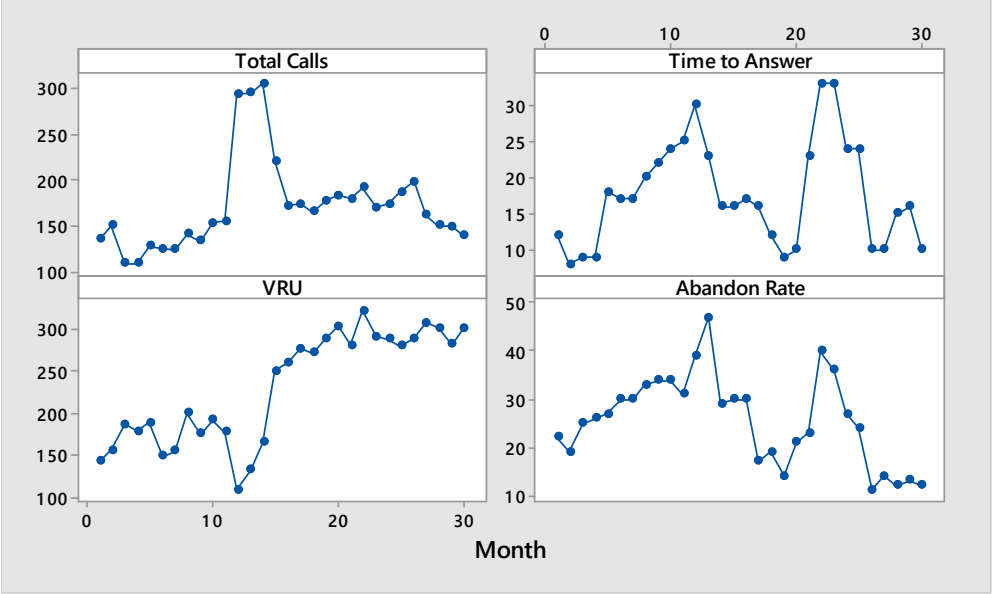


Figure 4.21 shows plots of the X1, X2, C3 and Y versus Month (Time). Of note is the similarity of the time plots for Abandon Rate and Time to Answer. The VRU plot shows a level change after Month 14. Recall that in the scatterplots (Figure 4.20) we saw strong positive relationships between Abandon Rate and Total Calls (call volume) and Time to Answer.

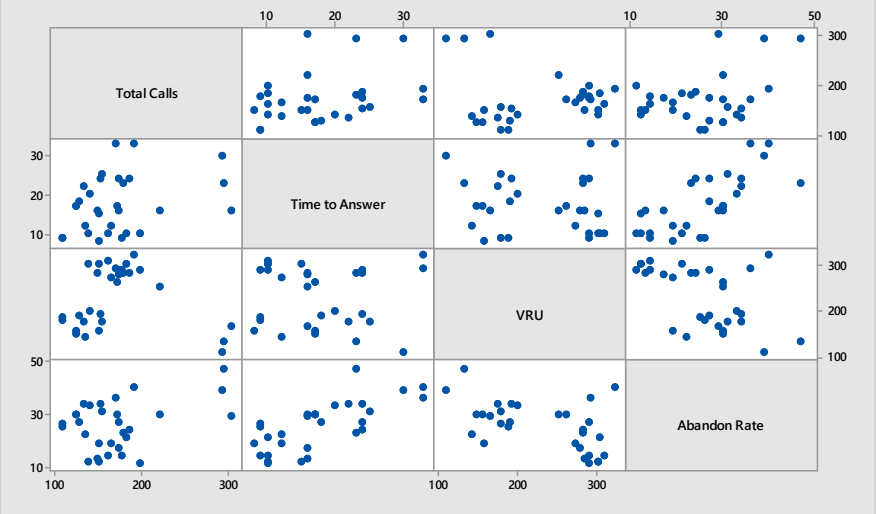
Figure 4.21 Call Center Data – Time Plots of Total Calls, Time to Answer, VRU and Abandon Rate



Plots of all the pairs of Xs are useful in identifying correlated predictor variables. In such cases it can be difficult to determine which the important predictor variables are. In general, multiple variables result in a proliferation of scatter plots. A succinct graphical display can be obtained

using a "Matrix Plot" such as that shown in Figure 4.22. Matrix plots of course get a bit difficult to interpret when the number of variables gets large. In Figure 4.22 we see that 3x's and 1 Y produces a total of 12 plots. Adding Month to the plots would produces 20 plots.

Figure 4.22 Call Center Data – Matrix Plot of S Total Calls, Time to Answer, VRU and Abandon Rate



A Labeled Scatter Plot is another plot that can be useful for visualization of relationships. In a labeled scatterplot, levels of a third variable are shown. The data in Table 4.9 are the sales (units sold) each quarter over a three year period in two sales regions. Figure 4.23 shows the relationship between Sales and Quarter over the three year period. A linear relationship is seen. The points have been "Labeled" with an "E" for the Eastern region and a "W" for the Western Region. The labels make it clear that the sales are consistently higher in the Western Region than in the Eastern Region. This difference is clarified further when straight line fits (Montgomery, et al 2001) for the two regions are shown on the plot as seen in Figure 4.24.

Table 4.9 Sales Data – Units Sold in Two Regions over a Three Year Period

Quarter	Eastern Region	Western Region	Region Difference	Quarter	Eastern Region	Western Region	Region Difference
1	986	1094	108	7	1090	1136	46
2	1017	1138	121	8	1124	1203	79
3	1075	1085	10	9	1113	1247	134
4	1053	1209	156	10	1126	1235	109
5	1076	1092	16	11	1218	1277	59
6	1064	1120	56	12	1162	1276	114

Figure 4.23 Sales Data – Labeled Scatterplot of Units Sold versus Quarter

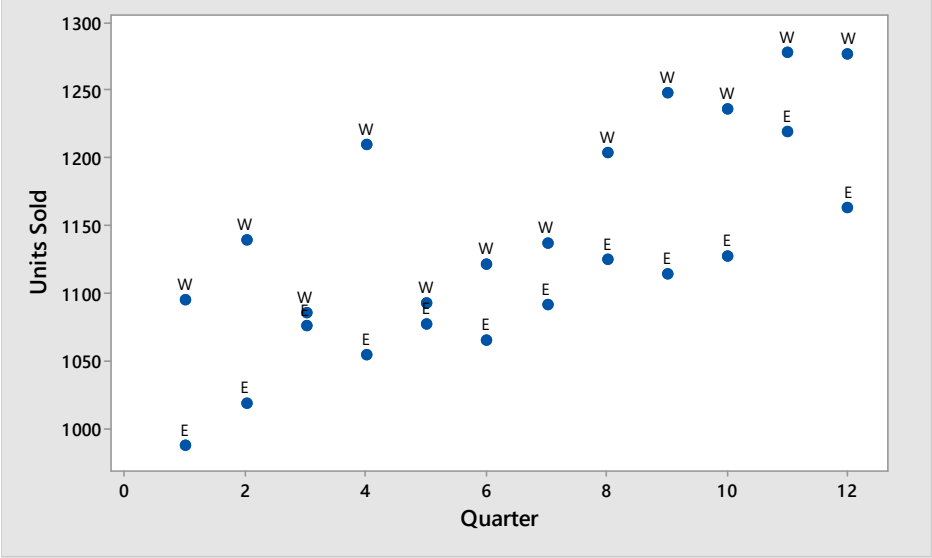


Figure 4.24 Sales Data – Scatterplot of Units Sold vs Quarter with Regression Lines for Each Region

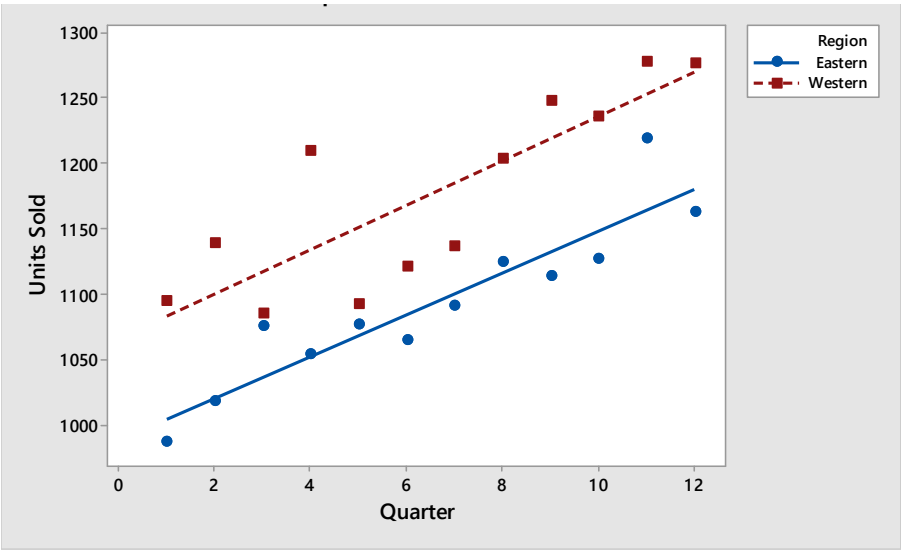


Figure 4.25 shows a slightly different use of the labeled scatter plot. The "labeled" variable in the sales data (Figure 4.23) was a qualitative variable. When the variable is quantitative it first converted to a qualitative variable and then the plot constructed.

In the case of the Call Center data time plots we saw in Figure 4.21 a shift upward in the VRU variable. We can see the effect of the VRU variable in the labeled scatter plot by categorizing the date in to "L" = VRU < 201 and "H" = VRU > 201.

The points are labeled with “L” and “H” in Figure 4.25. The labels make it clear that when VRU > 201 the Abandon Rate is consistently higher than when VRU < 201. This difference is clarified further when straight fits for the two groups of data are shown on the plot as see in Figure 4.26.

Figure 4.25 Call Center Data – Labeled Scatterplot of Abandon Rate vs Time to Answer for Low and High VRU

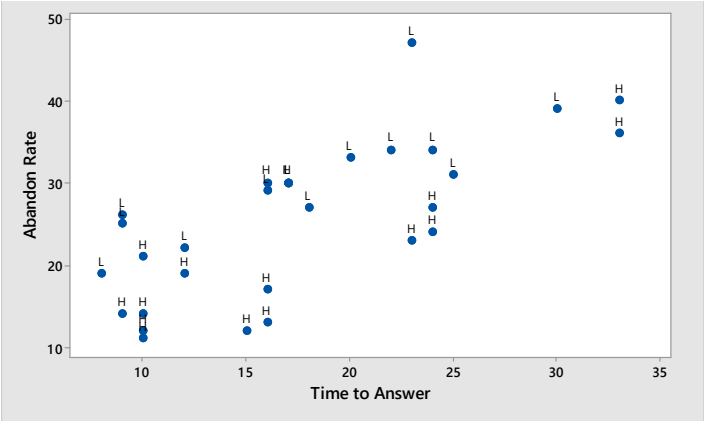
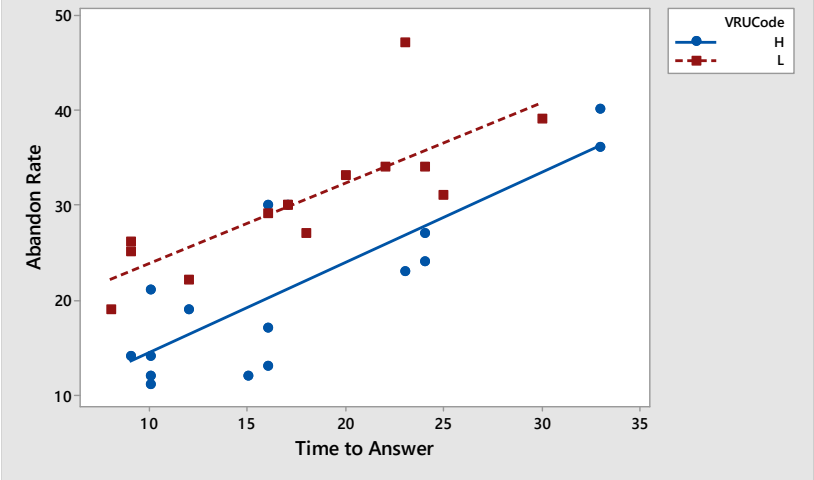


Figure 4.26 Call Center Data - Data – Scatterplot of Abandon Rate vs Time to Answer with Regression Lines for Low and High VRU



4.3.4.3 Visualizing Variation over Time

Almost all data are collected over time or in a sequence of some type. Plotting data versus time is a critical step in exploring the data, understanding the data and the process that generated it, and looking for variables that may not have been previously known to be important. By plotting the response versus time, we can identify clues as to which variables may be causing the variation in the response. In effect “time” acts as a surrogate for unknown important variables.

Time plots and the associated control charts also tell about the stability over time of the process

that generated the data. Stable processes are predictable and can thus be seen as "boring" devoid of any major causal variation. Assessing process stability is a critical activity in Statistical Process Control (SPC) Phase I studies (Montgomery 2013).

This role of time is illustrated in the following examples.

As a baseline we examine what a "stable" process looks like. In Figure 4.27 we see a process varying around a center value of 280 units. The variation is random, typical of a stable process. We see no evidence of non-random variation.

As noted in Section 4.3.4.1 a way to check for non-random patterns is to subject the data to a control chart analysis such as shown in Figure 4.28. Here we see all the points within the control limits indicating that the process is stable and predictable.

Figure 4.27 Bottle Weight Data – Time Plot

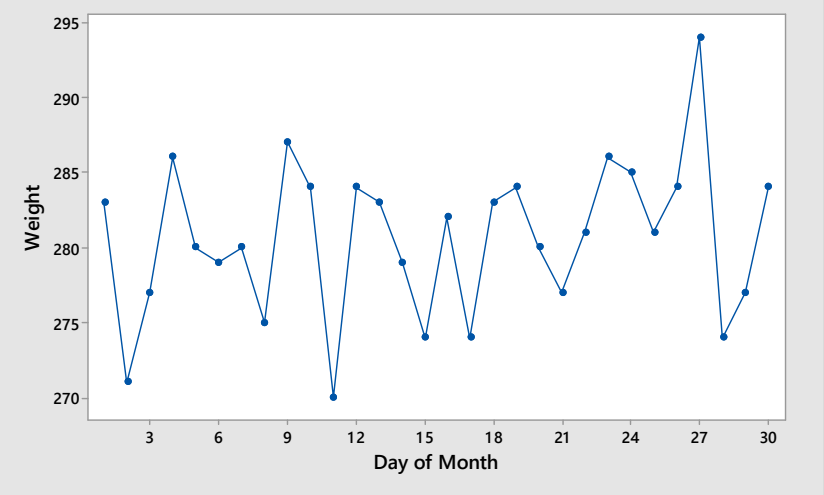
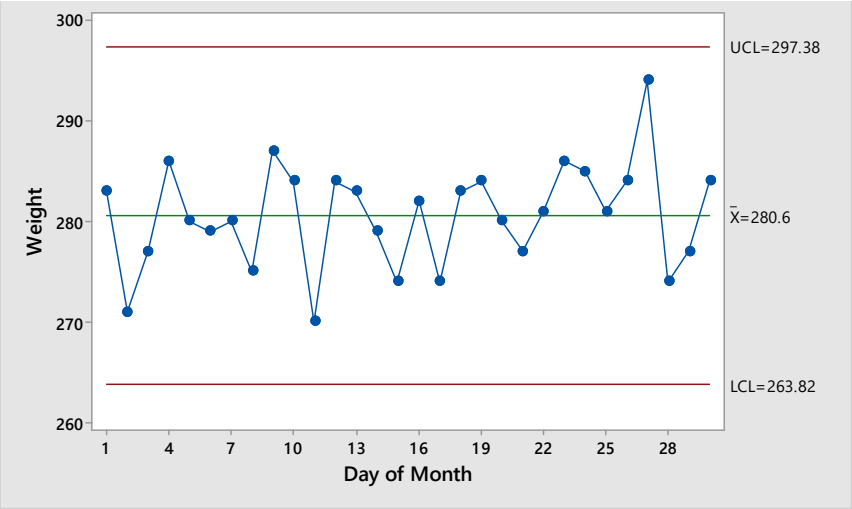


Figure 4.28 Bottle Weight Data – Control Chart



Process Yield. A company has a new product with a projected market of more than $500 million per year. The big concern is that the manufacturing process yield is too low to meet market demand. Manufacturing data are collected on the last 57 batches including yield (y) and approximately 30 process variables (Xs).

One of the first steps was to assess the stability of the process with respect to yield by constructing a control chart Figure 4.29. There we see that the yield for three batches was below the control limit, and we see a major downward shift after Batch 27. Clearly, the process is not stable. An investigation was initiated to identify the root causes of the problem.

A review of the data pedigree showed that three different batches of raw material had been used in the production of the 57 batches. A control chart with the batches segregated is shown in Figure 4.30. The low yields associated with Lot 3 were known, but the difference due to Lots 1 and 2 were unknown prior to this analysis. In fact, no analysis of the yield data had been done prior to this analysis.

Specifications were created for the raw materials and discussed with the supplier. When the tighter controls for the raw material were put into place, the batch yields increased by 25% and the market demand was met, resulting in a revenue gain of several million dollars.

Figure 4.29 Process Yield – Control Chart

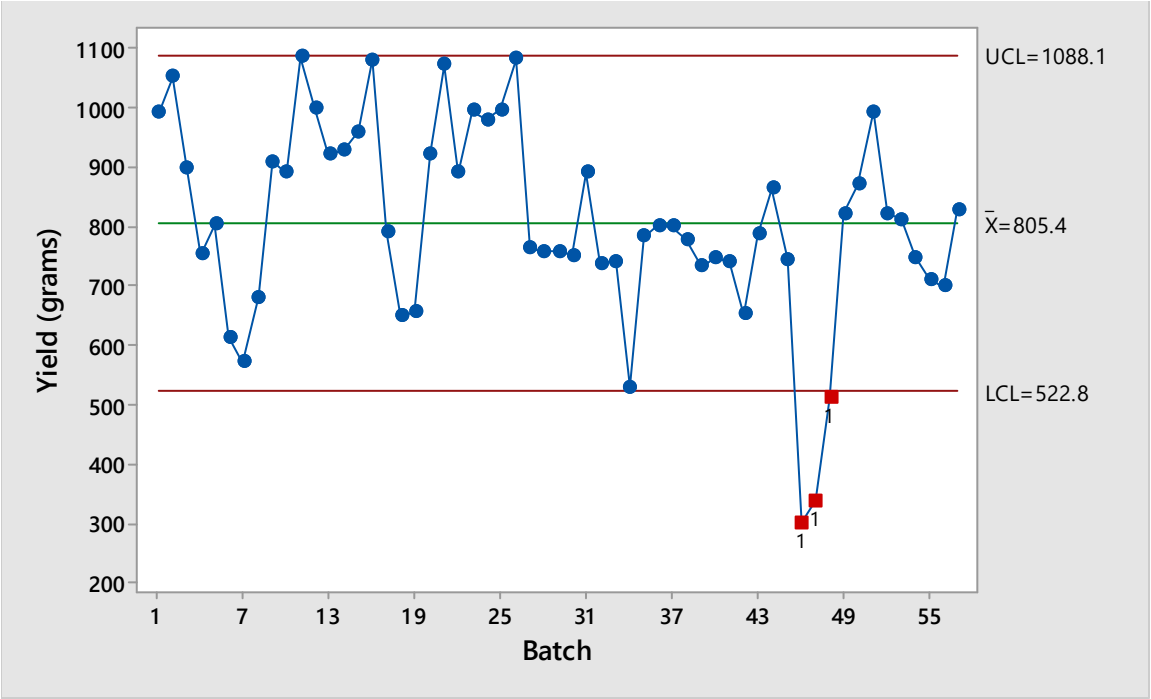
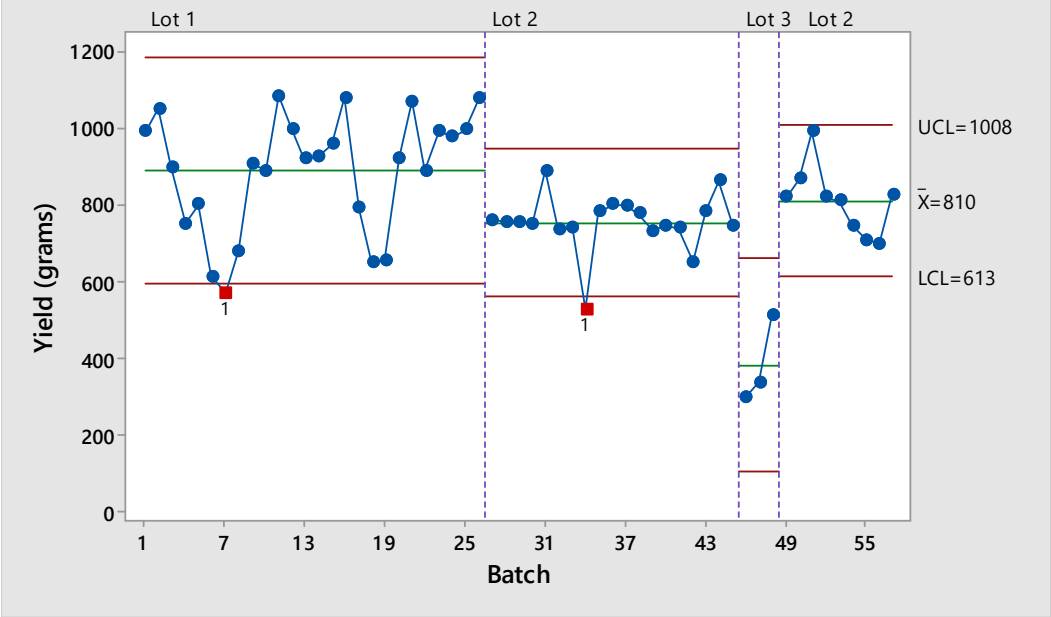


Figure 4.30 Process Yield – Control Chart Stratified by Media Lot Usage



Machine Differences. A pharmaceutical company is having problems with excessive variation of tablet thickness. Data were collected on 61 production lots. The first step was to look at

variation in tablet thickness over time using a time plot. We see in Figure 4.31 the large variation in tablet thickness. We also see a major bimodality with weights centered around 0.220 and 0.226. There is also a level shift after Lot 24. In examining the data pedigree it was learned that different tablet presses were used to manufacture the different lots. In Figure 4.32 we see that Press 120 is producing lower weight tablets than Presses 50, 90 and 110. Lots 1 – 24 were all produced using Press 120.

Figure 4.31 Pharmaceutical Tablet Thickness – Time Plot

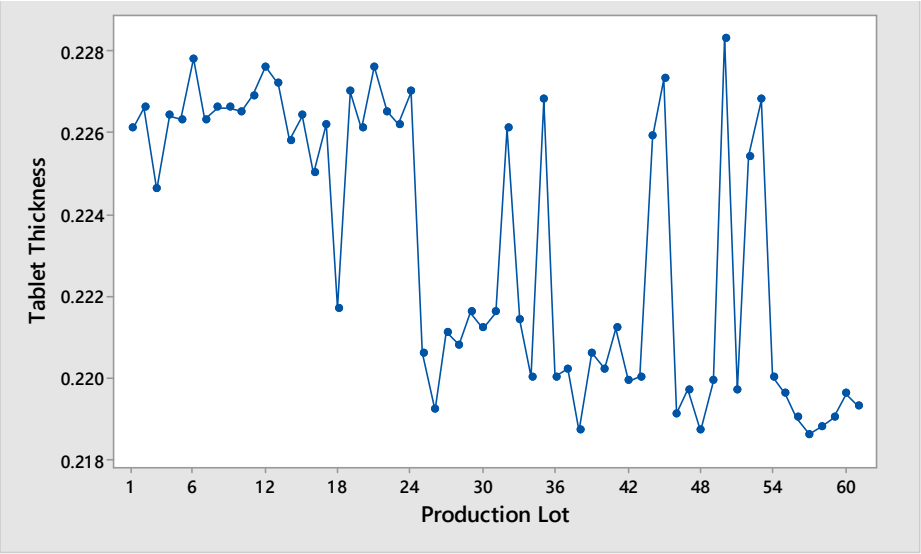
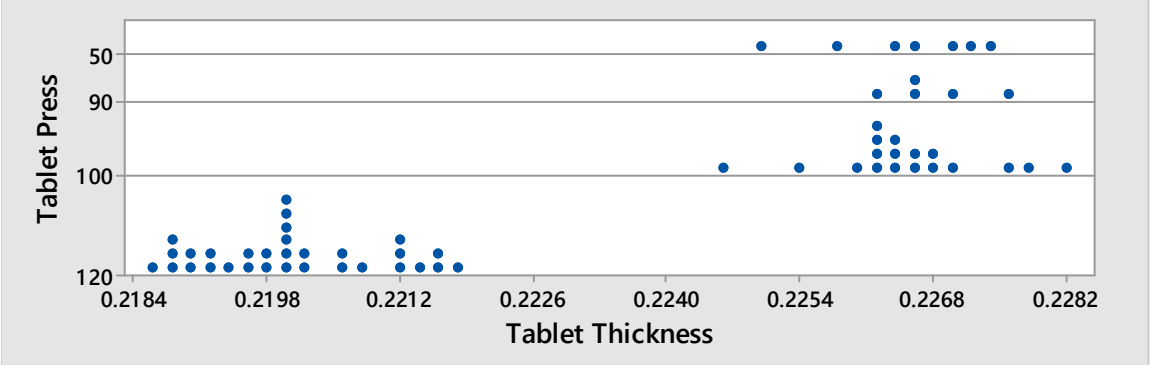


Figure 4.32 Pharmaceutical Tablet Thickness – Dot Plot of Thickness Stratified by Press



Analyst/Operator Differences. A pharmaceutical company was experiencing trouble with the active ingredient content of a tablet. A considerable amount of off-spec material was being produced and, as a result, much production was being rejected. Data were collected on the most recent 119 lots produced.

A time plot of the active ingredient was created as shown in Figure 4.33. There is considerable variation in the results with different amounts of variation at different points in time. Discussion ensued as to the sources/causes of the variation. Two prime candidates were tablet presses and raw material lots. Unfortunately, neither of these sources was found to correlate with the patterns observed in the data.

It was then recalled that different analysts were used to perform the test. When the analyst designations were superimposed on the data, it became apparent that Analysts A and C were major contributors of the variability and that the results of Analyst B were quite uniform (Figure 4.34). In Figure 4.35 which shows a dot plot for each analyst we also see that Analyst B is doing

most of the work (63%). Thus, by exploring the data using the data pedigree as well as time plots and dot plots uncovered the sources of the variation in the data.

Figure 4.33 Product Active Ingredient – Time Plot

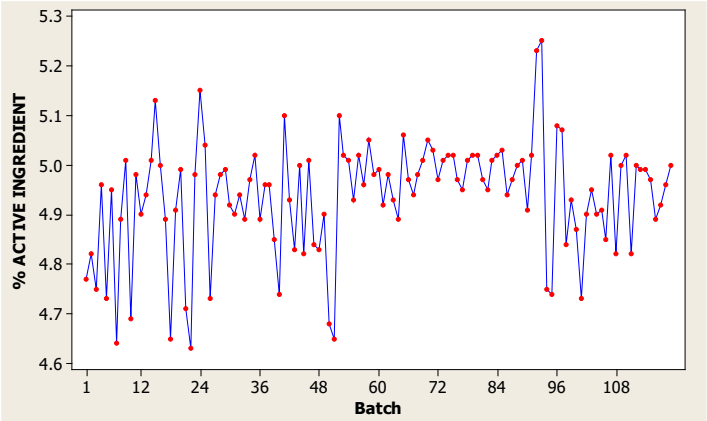


Figure 4.34 Product Active Ingredient – Labeled Time Plot

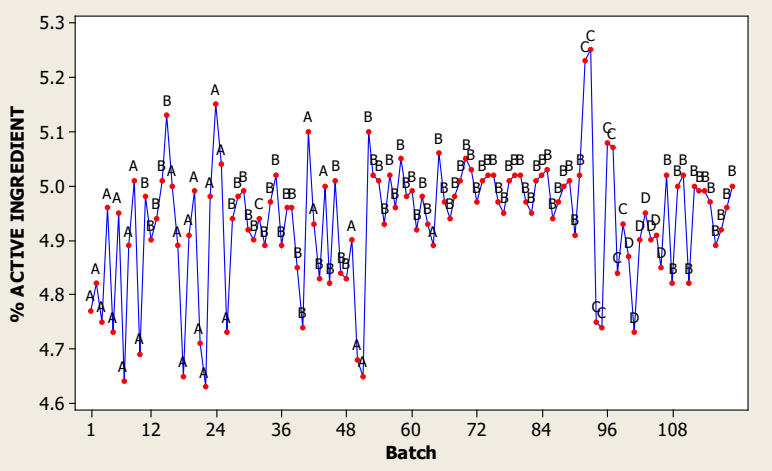
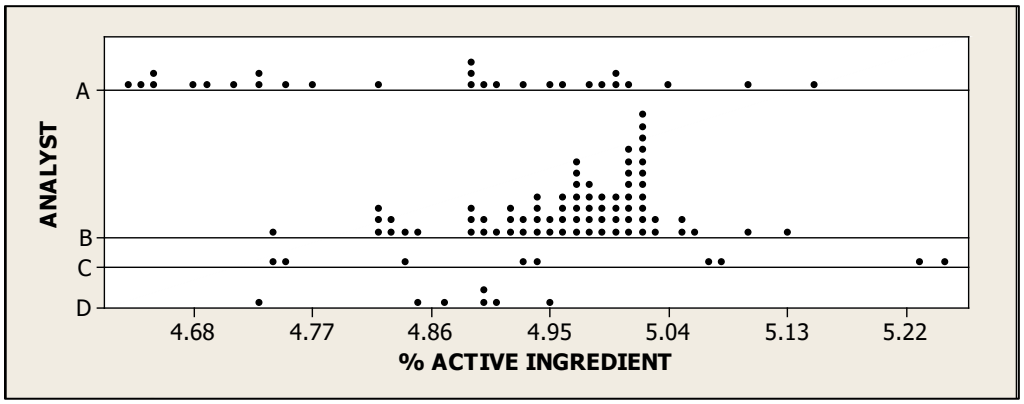


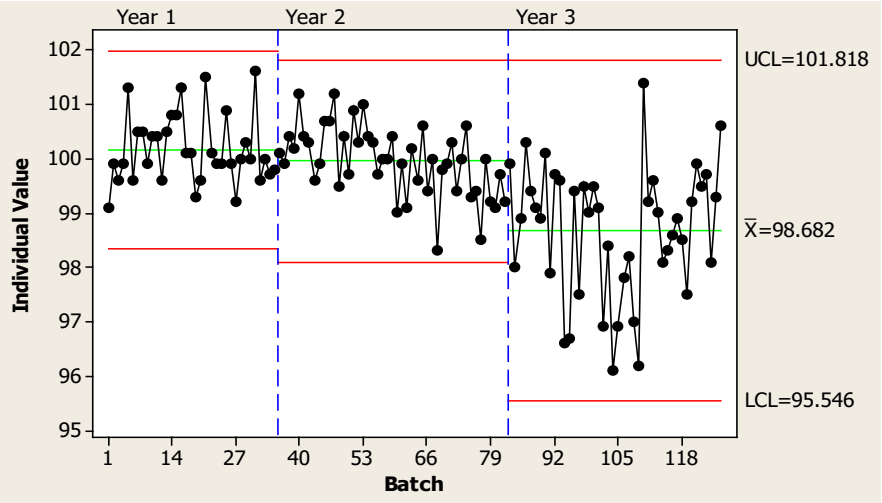
Figure 4.35 Product Active Ingredient – Dot Plot Stratified by Analyst



Process Stability. Stability is an important characteristic of a process. A stable process is predictable process in that we know limits within which the process will vary. The control chart is an excellent tool to assess process stability. In Figure 4.36 we see the batch assay values for a product produced over a three year period. During Year 1, the process is very stable. In Year 2, while all the points are within the control limits, a downward trend appears to start around the middle of Year 2. The trend continues into Year 3, around the middle of which a process adjustment is made increasing the assay value. Year 3 still shows considerably more variation than Years 1 and 2.

The assay specifications for the product are 95 – 105, so none of the assay values are out of specification. The lack of stability is still a concern as if not attended to may result in out-of-spec product in the future. For all processes, it is ideal to have an early warning devise that can detect the presence of shifts so quick corrective action can be taken.

Figure 4.36 Product Assay Value – Control Chart Stratified by Year of Production



4.3.5 Principles for Construction of Graphics

EDA makes varied and extensive use of graphics. One must be careful to use good graphics. It is not unusual to see graphics that are of poor quality for many reasons. So, what constitutes a good graph? Sage advice is offered by Tufte (1983) Cleveland (1985) and Hare (2020). Developers of graphics should consider the following principles:

- Communicate important findings using graphics. When possible, one graphic per important message or finding.
- An effective graphic is clear, concise, understood by developer and user alike.
- Follow Tufte's advice and avoid "chart junk"; that is information on chart that is not needed and does not add value. Direct reproduction of software output is often a source of chart junk. Frequently, there is information on computer output that is not relevant to the subject of the report or presentation.
- Clarity is paramount. Do not overdo tick marks; use them accurately and sparingly. Add reference lines when they enhance meaning. Avoid overlapping plotting symbols.
- Use figure legends (titles) to communicate important messages, but do not overdo legends. Explain graph features such as error bars. Make captions succinct.
- Fill the region almost entirely; avoid having too much white space. If adjacent panels are shown they should have the same scales. Include zeros if it makes sense, and use scale breaks only when necessary.
- Use color when possible.
- Graph development is generally an iterative process in which each succeeding graph represents an improvement in clarity over its predecessor.

4.3.6 Norms of EDA (Best Practices)

We summarize this subsection on Exploratory Data Analysis (EDA) by detailing some norms (best practices) one can use in implementing EDA. These include the following:

- You should adopt an inquisitive and skeptical mindset. Consider yourself a detective, forensic scientist, explorer, etc.
- Recognize that data exploration is a journey. You need to plan accordingly. You will learn about the data and the process that generated it as you proceed through the analysis process. You may need to collect more data or data of a different type.
- You should make extensive use of graphics along the way. You should expect graphical displays to reveal the unexpected as well as not to produce any useful information. Graphics do not always work.
- One of the first graphics to use is the time plot. Plotting the data in the sequence of generation should not be an option.
- It is critical at all times to be on the lookout for non-random patterns, things that do not look right. Important learnings are almost always in the non-random patterns.
- Focus on simplicity and parsimonious models. What are the critical few variables that are driving the process? Recognize that models are always imperfect; work to enhance models at every opportunity.
- EDA is inherently subjective; use subject matter knowledge and experience whenever possible. This includes understanding the data pedigree.

Using these best practices will make your EDA journey a pleasant one and increase the probably that your EDA work will be successful.

4.3.7 Note

The first section in this chapter identified five critical check points for exploratory data analysis: display, re-expression, residuals, resistance and iteration. This section is by necessity brief. As a result focus has been on display and iteration. The other check points are discussed in Chapter 5 of this Handbook titled, “Model Building.”

Summary of Chapter 4

This Chapter covers three topics: Theory of data exploration, data cleaning and using Exploratory Data Analysis (EDA).

Exploratory Data Analysis is defined and addressed at a high level providing the foundations for the methodology. The use of the resulting philosophy of data analysis is discussed and compared to other approaches. A "Global Positioning System" for an effective EDA journey is presented. The EDA journey is discussed along with competing models, understanding outliers, viewing EDA as a philosophy of science, EDA and other methods, and the norms of EDA.

Data cleaning is defined, highlighting five types of data cleaning problems and discussing methods for conducting data cleaning. The goal is to ensure that the available data contain as few problems as possible and are ready for analysis. Five types of "dirty data" are identified and discussed. Formal and informal methods of data cleaning are presented. Some best practices for data cleaning are discussed.

It is shown how Exploratory Data Analysis can be used to understand data prior to doing formal statistical analyses such as creating tests of significance, creating confidence intervals, developing statistical models, etc. EDA helps the analyst to be alert to unexpected patterns, relationships and extraordinary cases. Some tools useful in doing EDA are described and illustrated. We start with simple descriptive statistics that aid in characterizing the location and spread of the data. EDA is inherently graphical; widely useful graphical methods are discussed including the "Magnificent Seven" graphical tools. Next, the use of scatter plots in the visualization of relationships between variables is discussed. Special attention is paid to the use of time plots in studying the variation in the data over time. This often leads to the identification of important causal variables. Assessment of the stability of the underlying process is also a critical outcome of the analysis. The section concludes with a summary of the norms (best practices) one should keep in mind as EDA is used to explore data and better understand the process that generated the data.

Section 4.3 References

Cleveland, W.S. *The Elements of Graphing Data,* Wadsworth, Monterey, CA, 1985.

Hare, L. B. "Dodging Deceptive Depictions," *Quality Progress,* February, (2021): 36-44

Hoerl, R.W. and Snee, R.D. "Show Me the Pedigree," *Quality Progress*, January, (2019): 16-23

Hoerl, R. W. and Snee, R. D. *Statistical Thinking – Improving Business Performance*, 3rd Edition, John Wiley and Sons, Hoboken, NJ, 2020.

Jones, K.2020 https://www.visualcapitalist.com/media-consumption-covid-19/.

McDonald, L. "The Real Goods and the Oversell," *Significance,* April, Vol. 17, Issue 2 (2020).

Montgomery, D. C. *Introduction to Statistical Quality Control*, 7th Edition, John Wiley and Sons, Hoboken, NJ, 2013.

Montgomery, D. C., Peck, E. A. and C. G. Vining. *Introduction to Linear Regression Analysis*, 5th Edition, John Wiley and Sons, Hoboken, NJ, 2012.

Tufte, E.R. *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT, 1983.

Tukey, J. W. *Exploratory Data Analysis*, Addison Wesley, Reading, MA, 1977.

Wainer, H. "How to plot data badly," *The American Statistician*, May, Vol. 38, No. 2, (1984).

Chapter 5 – Drawing Inferences

Table of Contents

Preface

In this chapter we discuss a formal approach to drawing conclusions about an entire population or process of interest, based only on a sample, or subset, of this population.

Projecting from a sample of a population to the full population has risks associated with both bias and variation. Bias enters the picture when sampling units lack full and fair representation of the population. Sampling variation is the unavoidable difference between a sampled value and the true, but usually unknown corresponding value of the population. If sampling is carried out properly, avoiding pitfalls of bias and variation, we can derive accurate inferences about the entire population. This projection or inference from sampling to the broader population is what makes Statistics so useful to society.

This approach is generally referred to as *statistical inference*, although informally it is often referred to as *generalizability*. In this chapter we discuss the underlying theory of statistical inference, common reference probability distributions utilized in inference, and common methods of inference, such as confidence and prediction intervals, as well as hypothesis testing. We conclude with a discussion of common pitfalls in statistical inference and how they can be avoided.

Section 5.1 - The Theory of Statistical Inference

5.1.1 Objectives

The purpose of this section is to explain what statistical inference is, the general approach that is taken when applying it, and how it is relevant to statistical engineering applications.

5.1.2 Outline

We begin with an explanation of what is meant by the term *statistical inference*. This explanation involves discussion of the main steps taken to actually apply statistical inference. Next, we illustrate at a high level the underlying theory of statistical engineering, that is, how and why it works.

5.1.3 What is Statistical Inference?

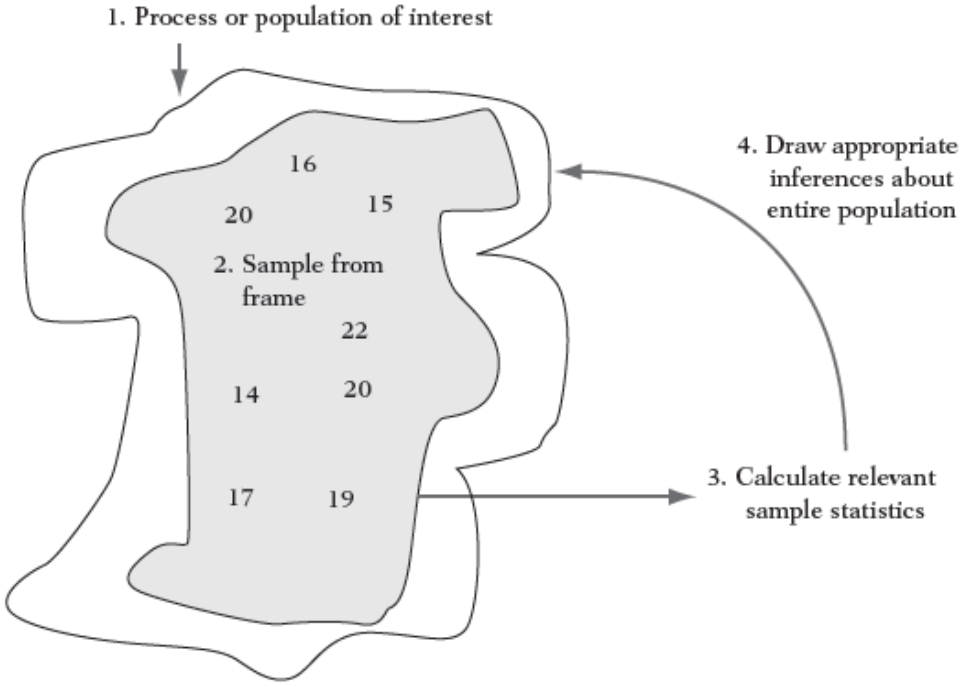
Fundamentally, the term *statistical inference* refers to drawing conclusions about an overall population or process of interest based on only a sample, or subset, of the population. For example, in election polling, pollsters do not speak with every single registered voter; this would be far too time consuming and expensive. Rather, they speak with perhaps 800 likely voters and use this information to make predictions or draw conclusions about what will happen in the election. Of course, one can never be 100% sure about the election from a sample of 800, but due to the power of probability theory, it can work surprisingly well. This same probability theory also documents the uncertainty, or "margin of error," in the inference.

The process of applying statistical inference is shown in Figure 5.1 (Hoerl and Snee 2020). This process consists of four main phases:

1. Identify and clarify the specific population or process of interest.
2. Define the "frame" from which one can actually sample and collect the amount of data needed. Document the data pedigree.
3. Calculate relevant statistics from the sample.
4. Draw appropriate inferences about the overall population using sample statistics in conjunction with subject matter knowledge, probability, and knowledge of the data pedigree.

Step 1, defining the population or process of interest, may sound trivial, but it is often the key to success. For example, in election polling, who specifically is in the population of interest? The obvious answer is: voters. However, ahead of time, how would we know who is going to vote? Registered voters could be used to define the population of interest, but some registered voters vote very infrequently. If registered voters do not actually vote, they will not impact the election. Therefore, professional pollsters generally use the concept of *likely voters*, which includes those who have voted in the past, as well as motivated newly-registered voters. Defining the population of likely voters is, in itself, challenging, and is done differently by different pollsters.

Figure 5.1 Statistical Inference Process



Once the population has been identified and clarified, such as defining *likely voter*, we still need to consider if we can sample directly from this entire population, or from only a subset, known as a *frame*. In most cases we cannot sample from the entire population of interest, due to practical considerations. Going back to election polling, some people will refuse to share for whom they plan to vote. In this case, these "refusers" cannot be sampled; rather, we sample from likely voters who are willing to share their intentions with us. Homeless people who are registered or people without phones would be other examples of people we are unlikely to be able to include in our sample. The sampling frame is that subset of the population of interest from which we can sample. This distinction between population of interest and sampling frame may seem trivial, but it often turns out to be a critically important component of the data pedigree.

As we will see later in this chapter, much of the theory of statistical inference is based on the principle of *random sampling*. A random sample requires that all items in the frame have an equal chance of being selected. More technically, it means that any sample of size n has the same chance of being selected as any other sample of size n. For example, a pollster may wish to send a detailed paper questionnaire to older likely voters, those who may not be proficient with computer technology, to better understand their priorities in an upcoming election. To minimize costs, they may decide to send the questionnaire to only 100 likely voters aged 60 or over. Taking the first 100 in alphabetical order on a list or 100 in a given zip code would not be random. These samples exclude voters in other zip codes and those whose names begin with letters such as w, y, or z. Such people could be sampled and are therefore part of the sampling frame. Having a person select the names generally produces a biased sample. Individuals often choose people they know or names they like. It is almost impossible for people to be totally objective in sampling. The most viable approach is to have a computer randomly select 100

numbers from one to the total number of people in the sampling frame and use this sample. This is conceptually equivalent to putting the likely voters' names on individual pieces of paper, placing these in a large hat, and selecting 100 out of the hat without looking—something that was actually done before personal computers with random number generators became commonplace! Random sampling helps ensure that the sample data we analyze are conceptually representative of the total population of interest, or at least of the sampling frame.

In practice, however, random sampling is rarely feasible. In election polling, it is well known that older retirees with land line phones are much easier to track down than young professionals with active social lives, who often do not own land lines. Young professionals might be reached; hence they are still part of the sampling frame, but it is less likely that they will end up being sampled. Professional pollsters understand this principle and tend to weight their results by demographics, knowing that some groups are likely to be oversampled and some undersampled. When sampling physical devices, such as motors sold over the past year to check for maintenance issues, some motors may be untraceable, some may have been resold, or some maintenance records may be incomplete, making it very difficult to obtain their information, even if selected in the sample. In practice, we attempt to sample in as random a manner as possible, or at least the most unbiased, and document any sources of non-randomness in the data pedigree.

The critical issues of obtaining the right quality of data, as well as the right quantity of data, were discussed in Chapter 2, Section 2.3. The approach used to collect the data, whether good or bad, should be carefully documented in the data pedigree.

It should be noted that in practice, we often decide on the process or population of interest, and even the type of inference desired, after seeing the data. As also noted in Chapter 2, the data collected often result in more questions than answers. For example, why do we see such high sales growth in the East, but not in the West? Why does closing the books take so much longer for 5-week months than for 4-week months? Surprises in the data will lead us to new hypotheses that we may wish to formally test via inference, even though the data were not originally collected for this purpose. In such cases, we should carefully consider what specific process generated the data; this gives the data a context and helps identify the population to which we can reasonably apply inference.

Once we have the data, we can calculate the statistics of interest, such as average, standard deviation, percentage, or perhaps a regression coefficient. Note that sample statistics are essentially "facts". That is, out of the 800 people polled, 425 said that they planned to vote for candidate Jones, producing a proportion of 0.53. This is factually true. However, we are not yet making any prediction (inference) about who will win the election. When we simply calculate sample statistics, without inferring anything about a larger population, these are usually referred to as *descriptive statistics*.

The calculated statistics also provide the input for constructing confidence or prediction intervals, hypothesis tests, or any other formal statistical inference technique, to be discussed in the following sections. These *inferential statistics* are not simply factual statements about the

sample, but are making inferences about the entire population, that is, about the data we did not observe. This is illustrated in step 4 of Figure 5.1.

What we are essentially doing in statistical inference is drawing conclusions about the *population parameters* of interest based on sample statistics. Population parameters, such as the population mean, variance, proportion, etc., are what we would calculate if we were able to perform a *census*, or 100% sample of the entire population. In practice, of course, we rarely have the opportunity to sample the entire population, so the population parameters are not observed. Elections are one counter-example; at the end of the election, we find out how all the voters actually voted. We get to see the entire population of voters.

To avoid confusion, textbooks typically use Roman letters for sample statistics, and Greek letters for population parameters. This is illustrated in Table 5.1. If we take a sample of 10 people from a large population, ask them their ages, and calculate the average age, this would be a sample statistic, or $\bar{x}$. If we were to use this sample average to infer about the average of the entire population of interest, we would be inferring about μ. Often, $\bar{x}$ is used to estimate μ; this is called a *point estimate*, because the inference involves only one value or "point," not an interval of uncertainty. It is important to keep in mind that these two symbols are not interchangeable; they do not represent the same thing. One is the calculated average of the specific sample that we selected ($\bar{x}$), while the other represents the average of the entire population (μ) and is not calculated from the sample.

Table 5.1 Sample and Population Symbols

	Sample	Population or Process
Average	$\bar{x}$	μ
Standard Deviation	s	σ
Proportion	p	π

Drawing conclusions about data we did not observe often seems counterintuitive or impossible, but statistical methods, if done properly, permit this. Of course, our inferences will never be 100% accurate. The technical challenge is to properly document our uncertainty, so that we know how far off our inferences might be. Confidence and prediction intervals are common methods of doing just this. As noted previously, modern election polls in the United States typically sample only about 800 likely voters, but these polls usually make accurate predictions of election results for the entire country. The trick is to sample carefully, and to accurately quantify uncertainty before forecasting the results of the election. In these polls, statements such as "This poll has a margin of error of +/- 4%" are typically made to quantify uncertainty. This "margin of error" is essentially a confidence interval.

A key conceptual question that needs to be answered at this point is how broadly the inference can be applied. For example, was the poll conducted across an entire state, within a given county, or only one city? If the poll was conducted in one county, then obviously one should not use these results to draw conclusions about the entire state. We should restrict our inferences to the actual frame from which we sampled. If, based on subject matter knowledge, we determine that

the difference between the sampling frame and overall population of interest is negligible, then we can be more confident in drawing inferences about the entire population.

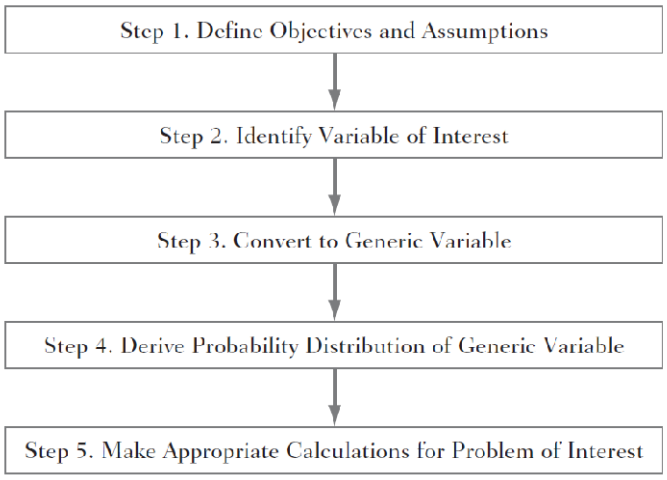
Such inferences beyond the sampling frame must be performed cautiously. A lack of careful consideration of the actual frame versus the population or process of interest is one reason why so many published studies seem to contradict one another. For example, news blogs may report a new study that indicates that some food or substance causes cancer. Very rarely is it actually mentioned that the study was conducted on rats that were given extremely large doses of the substance in question. Later, another study concludes that this substance does not cause cancer. Again, only in the fine print do you read that this second study was an observational study performed on humans, using levels of the substance found in typical usage. Neither study is wrong; the error occurs when the conclusions are inferred, or extrapolated, beyond the actual frame without careful thought. Further, the vital issue of the studies having totally different data pedigrees was ignored.

5.1.4 The Underlying Theory of Statistical Inference

Although there are many types of statistical inference tools, the theory underlying each of them is fundamentally the same. It is based on *mathematical statistics*, which has at its core the field of probability. A basic understanding of probability distributions helps apply statistical inference tools more thoughtfully. We present common probability distributions used in statistical inference in the next section. These are important because collecting sample observations from a larger population frequently follow one of these probability distributions.

Figure 5.2 (Hoerl and Snee 2020) shows an overall framework for the underlying theory of statistical inference. It illustrates the high-level steps conducted to draw quantitative conclusions about parameters of the population or process of interest from sample statistics. Note that these steps explain primarily step 4 of Figure 5.1, in which one infers about the population parameters on the basis of sample statistics. That is, we could consider Figure 5.2 an elaboration of step 4 in Figure 5.1. First, define the objective(s) and reasonable assumptions. Next, identify the key variable of interest and convert it into a generic format that is not dependent on the specifics of the current problem. This generic format makes it easier to determine or derive the appropriate probability distribution, which is then used to perform the calculations required for statistical inference.

Figure 5.2 Framework of Statistical Inference



Following is a closer look at each of these steps:

Step 1: Identify the objectives of the analysis and define the necessary and reasonable assumptions. For example, suppose we wish to estimate the population average (μ) from the sample average ($\bar{x}$). We would need to understand the pedigree of any data collected and make some assumptions about how the sampled frame relates to the entire process or population of interest. This determination would be based on subject matter knowledge, rather than statistics. Also, as we will see, it helps if we have some idea of the probability distribution of the population (normal, exponential, etc.).

Step 2: Identify the variable of interest. The variable must be something quantitative, that is, something that can be counted or measured. If the variable is not quantifiable, it cannot be mathematically analyzed. It could be a discrete (nominal) variable and be assigned values of 0 ("no") or 1 ("yes"). The variable of interest could be directly measured—such as dollars, temperature, time, and so on—or something calculated—such as profit, inventory turns, or price-to-earnings ratio. The variable of interest may be an average, a standard deviation or other parameter. In our case, it is the sample average, $\bar{x}$.

Step 3: Convert this variable into a "generic" or standardized variable to the greatest degree possible. Standardization allows categorization of numerous problems, which are all unique, into a few generic "buckets" that have standard solutions. This helps avoid having to invent a unique solution to every problem. In our case, $\bar{x}$ is in specific units of measurement and comes from a population with some mean and variance. We would not want to develop a different solution each time we estimate a population mean from a sample mean, depending on the units of measurement, the probability distribution, or the population mean and variance. Converting to a generic variable allows a standardized approach. In this case, we might convert the sample mean, $\bar{x}$, to a *t*-value, based the *t*-distribution (a common probability distribution). This *t*-value approach can be applied across a wide variety of individual problems, each with different units of measurement.

Step 4: Now that we have a generic variable, which is not problem-specific, we must determine or derive its probability distribution. In our simple case of $\bar{x}$, this is straightforward. However, for many problems, this step is much more difficult than one might think. Fortunately, the mathematical work required to derive the probability distributions has been done for the majority of standard inference problems, such as those given in this chapter. For an individual measurement, we can often estimate the probability distribution from a histogram of sample data. However, suppose the variable of interest is the standard deviation. What is the probability distribution of a standard deviation calculated from sample data? If we obtained another sample and calculated a standard deviation from the new data, would it exactly equal the standard deviation from the first sample? Of course not, because of random variation in the samples. Probability theory is the main tool used to derive these sampling distributions for averages, standard deviations, proportions, differences between averages, etc.

Step 5: Make the appropriate calculations to estimate a parameter (point estimate), document the uncertainty in such estimates (confidence interval), or determine if we have enough evidence to "call the election" (hypothesis testing). Modern statistical software typically makes this step fairly easy, but we still need to know how to interpret the results. For example, practitioners frequently confuse confidence and prediction intervals, to be discussed later in this chapter. The software uses knowledge of the probability distributions to make these calculations, which is why knowledge of the probability distributions involved is so critical.

While we have presented the underlying theory of statistical inference conceptually, we show practical applications of this theory in later sections. As you read about strange sounding variables, such as t, z, or F, recall that the purpose of using these generic variables is simply to allow a common solution to a diverse array of practical problems.

Section 5.1 References

Hoerl, R.W., and Snee, R.D. *Statistical Thinking: Improving Business Performance*, 3rd edition, John Wiley & Sons, Hoboken, NJ, 2020.

Parten, M. *Surveys, Polls, and Samples*, Harper, New York, 1950.

Hoerl, R.W., and Snee, R.D. *Statistical Thinking: Improving Business Performance*, 3rd edition, John Wiley & Sons, Hoboken, NJ, 2020.

Section 5.2 – Common Reference Probability Distributions

5.2.1 Objectives

This section provides basic information regarding fundamentals of probability distributions in most common use, together with basic building blocks of statistical inference.

5.2.2 Outline

The distinction between discrete and continuous probability distributions is made. This is followed by descriptions of binomial and Poisson distributions in the discrete case and normal, lognormal and exponential distributions in the continuous case. Next is an explanation of sampling distributions, including the distribution of averages, the central limit theorem, and the t-distribution. References for further reading appear at the section's end.

5.2.3 Discrete and Continuous Variables

Often the source of some confusion, the distinction between discrete and continuous variables must be considered before sensible inference may be undertaken. Generally, discrete means distinct or separate, and continuous means forming an unbroken whole or without interruption.

The language of science does not depart from these meanings, but its assigned connotations enhance specificity and utility. We refer to discrete variables as those that can take on only specific values. A compound is either carbon based or it is not. An automobile is domestic or foreign. A bird is a finch, a sparrow, a hawk, an eagle, or some other species. Of course, there are variations on the theme of each discrete variable. A domestic automobile may be largely composed of foreign parts, for example. Still, some kind of classification persists regardless of possible blending.

Continuous variables, on the other hand, may take on any number, some say an infinite number, of values between two points. Between any two rates of dissolution, there is another rate. Between any two temperatures, there is another temperature. This holds even if the "in between" numbers cannot be measured or perhaps even attained.

A rule of thumb is: if you count it, it is discrete; if you measure it, it is continuous.

Before we venture into details of some frequently used discrete and continuous probability distributions, some terminology and corresponding notation are needed. We reemphasize the distinction between a parameter and a statistic. A parameter is a true value of a population whereas a statistic is an estimate of that value. The value could be a mean, a median, a standard deviation or any other specific distributional characteristic.

As a reminder, Greek letters are usually used to denote parameters, and Roman letters are used to denote statistics. (See Section 5.1.3 and Table 5.1.) For example, if we were to find from a random sample of 500 Baroque music manuscripts, 150 had opening movements written in 6/8 time, we would estimate the true population proportion of all such manuscripts using that initial time signature as p = 150/500 = 0.3. Here, p is the parameter estimating the true proportion, π (pie). This is a discrete variable example.

By the same token, the gram weight of an aspirin tablet, x, randomly selected, may be used to estimate a property of a large population of aspirin tablets. If many such x-values are selected and their average is calculated, the average, $\bar{x}$ is an estimate of the population mean, μ (mu). This is a continuous variable example.

5.2.4 Common Discrete Distributions

While only two discrete distributions are discussed here, readers should know that there are many. The binomial and Poisson distributions are arguably those in most frequent use. However, occasions for other discrete distributions such as the hypergeometric, the negative binomial, and the beta binomial do exist in the world of practical application.

In what follows, we provide both means and standard deviations of the subject distributions. For many distributions, most of the observed values will fall within one standard deviation of the mean. This is only approximate. For more exact statements and fine distributional properties, see Johnson, Kemp and Kotz (2005).

5.2.4.1 Binomial Distributions

A political scientist ventures into a small town to feel the electoral pulse. She works to establish a focus group of six members, but in doing so, she wants to assure a reasonable mix of political persuasions. Historic data reveal that 30% of the city's population vote "R." Naturally, she would like a healthy mix of R voters (Rs) in her focus group. They should not be absent, but they should not dominate, either. The question is if she chooses participants at random and independently, how many Rs are likely to appear in the group?

Now the probability of an R is 0.3, so the probability of all six seats being filled by Rs is 0.3^6 which is very small, around 0.1%.

The probability of the first seat, but none of the remaining seats being filled by an R is 0.3 X 0.7^5 which is close to 5%. The same is true for the second seat, about 5% and so on for each of the six seats in the focus group. So the probability that exactly one of the six seats will be occupied by an R must be six times the probability of any single seat. That is 6 X 5% or about 30%.

So far, we have we have figured the probability that there would be no Rs, and we have determined the probability of exactly one R. The logic is the same throughout all the possible seat count combinations which are listed in Table 5.2. and shown in the form of a histogram in Figure 5.3.

Figure 5.3 Probabilities of Rs in a Six Person Focus Group from a Population with 30% Rs

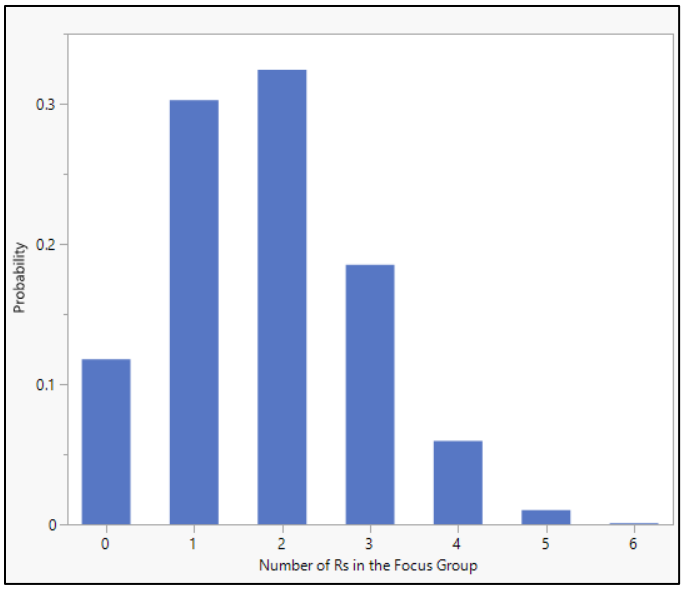


Table 5.2 Enumeration of Probabilities Associated with All Possible Outcomes of Selecting Dichotomous Events* with an Event of Interest having a Probability of 0.30

		Seat Number							
Number of Rs	Combination Number	1	2	3	4	5	6	Probability	Sum of Probabilities
0	1							0.1176	0.1176
1	1	R						0.0504	0.3025
	2		R					0.0504	
	3			R				0.0504	
	4				R			0.0504	
	5					R		0.0504	
	6						R	0.0504	
2	1	R	R					0.0216	0.3241
	2	R		R				0.0216	
	3	R			R			0.0216	
	4	R				R		0.0216	
	5	R					R	0.0216	
	6		R	R				0.0216	
	7		R		R			0.0216	
	8		R			R		0.0216	
	9		R				R	0.0216	
	10			R	R			0.0216	
	11			R		R		0.0216	
	12			R			R	0.0216	
	13				R	R		0.0216	
	14				R		R	0.0216	
	15					R	R	0.0216	
3	1	R	R	R				0.0093	0.1852
	2	R	R		R			0.0093	
	3	R	R			R		0.0093	
	4	R	R				R	0.0093	
	5	R		R	R			0.0093	
	6	R		R		R		0.0093	
	7	R		R			R	0.0093	
	8	R			R	R		0.0093	
	9	R			R		R	0.0093	
	10	R				R	R	0.0093	
	11		R	R	R			0.0093	
	12		R	R		R		0.0093	
	13		R	R			R	0.0093	
	14		R		R	R		0.0093	
	15		R		R		R	0.0093	
	16		R			R	R	0.0093	
	17			R	R	R		0.0093	
	18			R	R		R	0.0093	
	19			R		R	R	0.0093	
	20				R	R	R	0.0093	
4	1	R	R	R	R			0.0040	0.0595
	2	R	R	R		R		0.0040	
	3	R	R	R			R	0.0040	
	4	R	R		R	R		0.0040	
	5	R	R		R		R	0.0040	
	6	R	R			R	R	0.0040	
	7	R		R	R	R		0.0040	
	8	R		R	R		R	0.0040	
	9	R		R		R	R	0.0040	
	10	R			R	R	R	0.0040	
	11		R	R	R	R		0.0040	
	12		R	R	R		R	0.0040	
	13		R	R		R	R	0.0040	
	14		R		R	R	R	0.0040	
	15			R	R	R	R	0.0040	
5	1	R	R	R	R	R		0.0017	0.0102
	2	R	R	R	R		R	0.0017	
	3	R	R	R		R	R	0.0017	
	4	R	R		R	R	R	0.0017	
	5	R		R	R	R	R	0.0017	
	6		R	R	R	R	R	0.0017	
6	1	R	R	R	R	R	R	0.0017	0.0007

* Rs and not Rs

Life would be easier if there were a way to calculate these probabilities without having to stumble through the math or having to create a large table of all possible outcomes and their associated probabilities. As it turns out, there is. A formula that can be used to calculate binomial probabilities is:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, n = 1,2,3,\cdots$$

Here,

$f(x)$	is the probability of x, where
x	is the number of events. In the current example, x is an integer value from 0 to 6.
$\binom{n}{x}$	is the binomial coefficient. It serves as a counter of the number of ways we can take n items, x at a time. Mathematically, it is equal to $\frac{n!}{x!(n-x)!}$, where any integer followed by an exclamation point is the product of the integers up to and including that number. Zero factorial is defined as 1.
p	is the probability of a single event. In the current example, p is 0.30.

Following through with the example, suppose the political scientist is interested in knowing the probability that exactly 4 seats will be filled by Rs. We calculate using the formula for the binomial distribution, as follows.

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Becomes

$$f(4) = \binom{6}{4} 0.3^4 (1-0.3)^{6-4}$$

Note that the computed number of combinations for 4 seats is the same as that shown corresponding to 4 seats in Table 5.2, which also shows the corresponding probability of 0.0595 calculated the hard way.

And

$$\binom{6}{4} = \frac{6!}{4!\,2!} = \frac{6\ X\ 5}{2} = 15$$

So

$$f(4) = 15\ X\ 0.3^4\ X\ 0.7^2 = 0.0595$$

With all of this work so far, we have not quite met the concern of the political scientist. Recall, she wanted to assure that the Rs would be neither scarce nor dominant. Numerically, that might mean that there should be between 2 and 4 Rs. To respond to that concern, we would simply sum the binomial probabilities for x = 2, 3 and 4. You should get 0.5689. The conclusion is that the desired balance is no sure thing. An alternative recruiting strategy or perhaps a larger focus group might be considered.

Another useful form of the binomial distribution is defined by its cumulative statement.

$$F(x) = \sum_{x=0}^{k} \binom{n}{x} p^x (1-p)^{n-x}, k \le n$$

It accumulates the binomial probabilities from a starting point, usually 0 to a higher number, usually less than n.

Modern statistical software such as SAS, JMP, R and Minitab will do the hard work of calculations for you. Even software not specifically intended for statistical purposes contains provisions for basic statistical calculations. Excel is a case in point.

In Excel, the "BINOM.DIST" function is helpful with computations. Its arguments are the number of successes, the number of trials, the probability (of a success) and either "true" of "false." Use "true" if you want the cumulative probabilities from 0 successes to the desired high number of successes. Use "false" if you want an individual probability. In the case of the above example where we wanted between 2 and 4 successes, meaning Rs, we would use these Excel statements.

=BINOM.DIST(4,6,0.3,TRUE)-BINOM.DIST(1,6,0.3,TRUE)

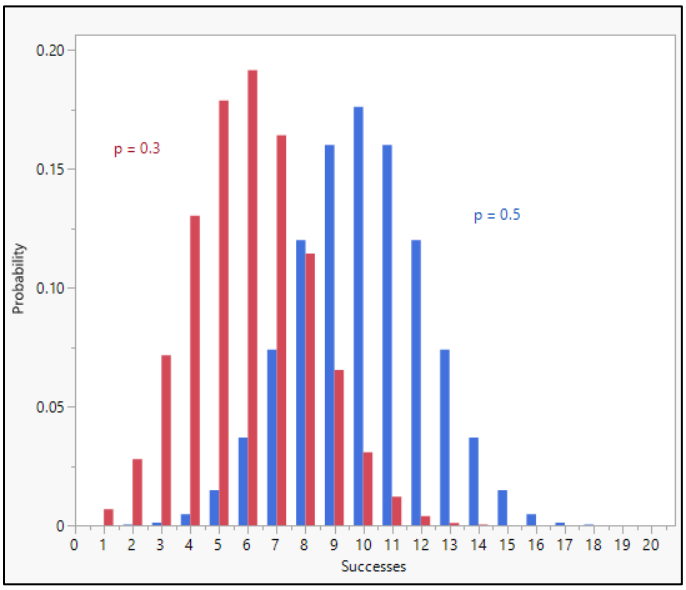
The result is 0.5689, and Excel does the math. In each Excel term, the first element is the number of successes, the second is the total number of trials, the third is the probability of an event, and the forth is either "true" for the cumulative distribution or "false" for an individual probability.

Two parameter estimates of the binomial distribution are important to know. One is the mean and the other is the standard deviation.

- The mean of the binomial distribution is simply np, and p= the probability of an event.
- The standard deviation is $\sqrt{np(n-p)}$.

As hinted by Figure 5.3, the binomial probability distribution is nearly bell shaped. As it turns out, if p = 0.5 as is the case with tossing an ideal coin, the binomial distribution is fully symmetrical. When p departs from 0.5, the shape becomes increasingly skewed and more peaked, as illustrated by Figure 5.4.

Figure 5.4 Binomial Probabilities of Success from Distributions of 20 Trials Each with Probabilities of Success of 0.3 and 0.5



5.2.4.2 Poisson Distributions

Suppose the same political scientist we met in the previous section wants to address a different problem. This time, she is concerned about the prospects of her focus group continuing uninterruptedly. Her champions back at head office have told her to expect at least one interruption per focus group session, caused by such events as cell phone users failing to obey the rules they agreed upon, emergencies in the building, wrong room deliveries and other sources of annoyance.

Assuming the one interruption per session rate applies, what are the chances that her session is interrupted? A mathematical genius, Siméon Poisson, derived a generalized solution to this and similar problems by examining the limits of the binomial distribution when the sample sizes are very large but the number of successes remained constant. His solution gave rise to the distribution named in his honor:

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,3\cdots$$

Figure 5.5 Poisson Probabilities Corresponding to Means of 1, 2, 5 and 10

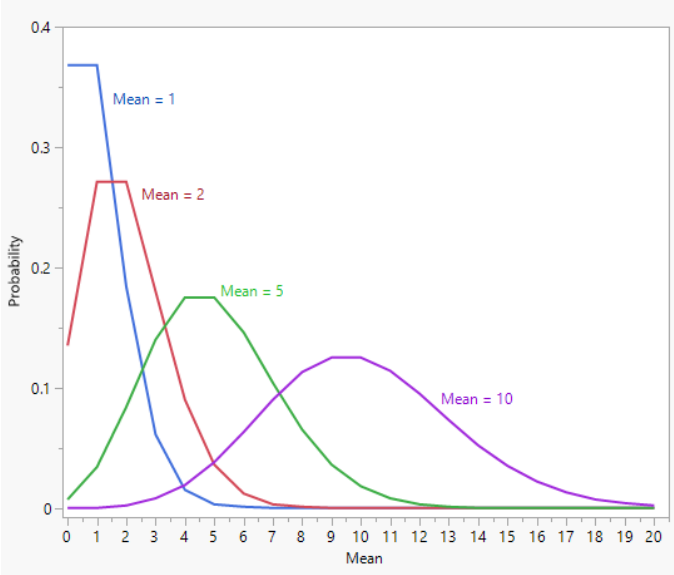


Table 5.3 Poisson Probabilities of Interruptions for an Expected Mean of 1

Interruptions:	0	1	2	3	4	5	6
Probability:	0.368	0.368	0.184	0.061	0.015	0.003	0.001

From this table, we can see that the probability her session will be interrupted at least once is the sum of the probabilities corresponding to 1 or more interruptions. This is the same as 1 minus the probability of no interruptions, or 0.632.

Excel will calculate these probabilities for you. The statement is,

=POISSON.DIST(J$6,1,FALSE)

in which the first argument is the location of the number of the event, the second is the mean and the third specifies the choice of cumulative probabilities (true) or individual probabilities (false).

The Poisson distribution has a single parameter, λ which represents the mean. The variance is also λ. This means that the standard deviation is $\sqrt{\lambda}$.

Examples of applications of the Poisson distribution abound. A few examples are:

- The number of consumer compliments or complaints coming into a call center in a fixed time period
- The number of particles emitted by a radioactive source in a fixed time period
- The number of customer accesses to an ATM during a fixed time period
- The number of blemishes on automobile hoods of a fixed model exiting the production line.

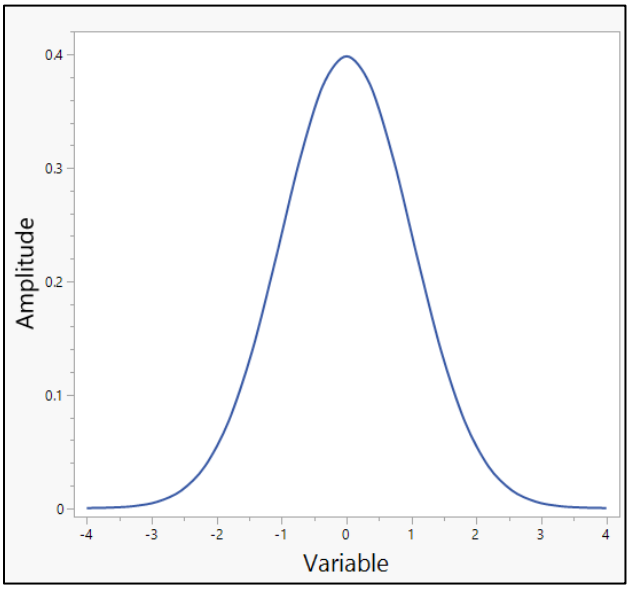
5.2.5 Common Continuous Distributions

As is the case with discrete distributions, there are many continuous distributions. We discuss only three here; the normal, the lognormal and the exponential distributions. For each continuous distribution, the response lies along a virtually infinite scale. It is possible for a process to run for 13.75824 hours, for example, even though that duration might be extremely difficult to measure with any reasonable degree of precision. Still, the duration exists in theory if not in practice.

Another discrete-continuous distinction is elucidated by the way we interpret event probabilities. As a case in point, the histogram in Figure 5.4 shows bars whose heights correspond to the probabilities of each of the seven events represented; that is, from 0 to 6 participants in the focus group. This is not the case involving displays representing continuous distributions.

What we see in the image of a continuous distribution is the amplitude of a curve corresponding to an infinite number of values along the horizontal axis as in Figure 5.6. Because the number of values is infinite, the actual probability of each is 0. The area under the curve is 1, as is true for the sum of all the probabilities represented by the bars depicting a discrete distribution.

Figure 5.6 A Continuous Probability Distribution



5.2.5.1 Normal Distributions

There is elegant simplicity to the distribution in Figure 5.6. It is a standard "normal" distribution developed by German mathematician, Carl Friedrich Gauss (1777-1855) and popularized by Karl Pearson (1857-1936), a pioneering British statistician. Pearson gave it a generic name in order to make it independent of country of origin (The drums of war had been beating.), and he called it

"normal" meaning it is the "norm" or usual distribution. He did not intend to mean that other distributions were abnormal.

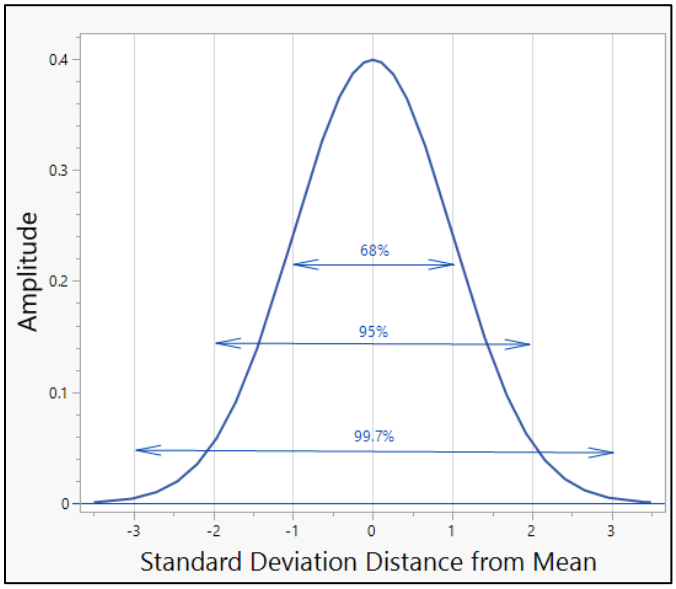
To be sure, it is the distribution in most common use, as it stands at the base of the most frequently used statistical computations, analyses and tools such as regression analysis, t-tests, the analysis of variance and control charts – reasons follow later in this section. In addition, it is the underling distribution of many sampling units measures, especially those in which a central value or target value is pursued. Examples include container net contents, tablet weights and automobile parts dimensions, just to name a few. Applications abound.

The normal distribution is uniquely defined by its mean (μ) and its standard deviation (σ). And while its probability distribution is given by a formidably appearing equation,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty,$$

certain distributional characteristics make it easier to understand and put to advantage. For example, see Figure 5.7. It shows that in a normal distribution, the interval of ±1 standard deviation about the mean will contain 68% of individual observations, ±2 standard deviations from the mean will contain 95% of individual observations, and ±3 standard deviations from the mean will contain 99.7% of individual observations. Of course, the total area under the normal curve is 1.0 or 100% of the observations. In many practical applications, that is all you need to know.

Figure 5.7 Percentage of Individual Observations within 1, 2 and 3 Standard Deviations of the Mean in a Normal Distribution



For example, if you have a manufactured automobile part measuring 10.0 mm on average with a standard deviation of 0.1 mm, you should expect approximately 95% of the parts to fall between 9.8 and 10.2 mm. 99.7% would fall between 9.7 and 10.3 mm.

Regardless of the shape of your continuous distribution, you can apply the equation below to convert it to a standard distribution with mean 0 and standard deviation 1.

$$z = \frac{x - \mu}{\sigma}$$

The numerator in this equation effectively slides the original numbers across the number scale so they are centered (mean) on zero. Dividing the numbers by their standard deviation generates a new set of numbers with a standard deviation of 1. The newly formed variable, z, thereby has mean 0 and standard deviation 1.

If the original x data follow a normal distribution then the new z data will also. z is often referred to as the standard normal variate.

Returning to the example of the automotive part whose mean is 10.0 mm and whose standard deviation is 0.1mm, suppose a part from the process is measured at 9.6 mm. Is there reason to believe it may not have come from the original population? Calculate

$$z = \frac{x-\mu}{\sigma} = \frac{9.6-10.0}{0.1} = -4.0,$$

which is to the far left of the number scale in Figure 5.7, suggesting the part likely does not belong to the original population. It is simply too many standard deviations away from the mean.

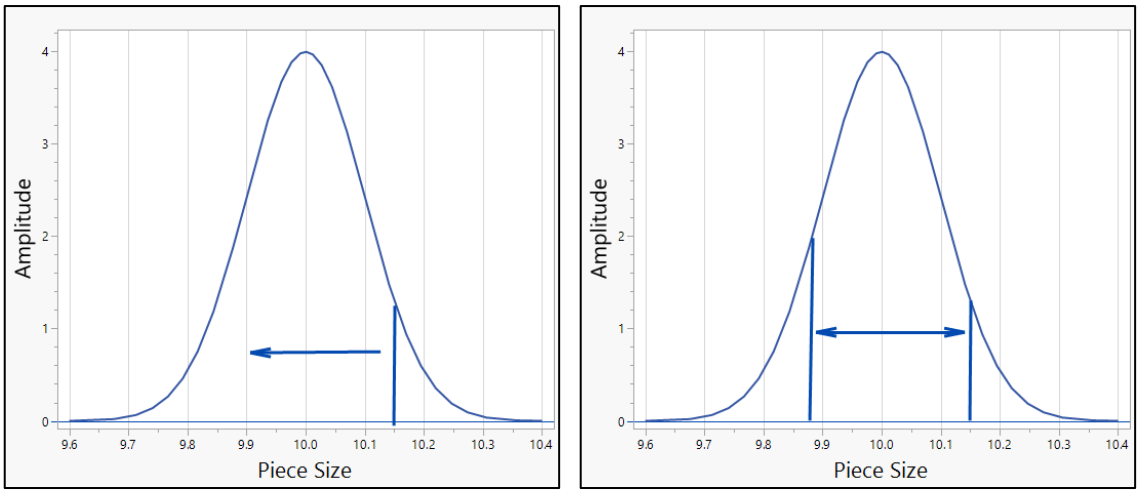
More specific questions can be answered by converting x data to z data and looking up the area under the normal curve corresponding to z in a table of the normal distribution. These tables may be found in the back of most statistics texts. Many students of statistics have, out of anger and frustration, converted their texts to ash. Excel comes to the rescue of the repentant.

Suppose, from the problem above, we want to know the percentage of parts at 10.15 mm or below. Go to Excel and choose: NORM.DIST from the function library. Fill it in with the value in question, the distribution mean, the standard deviation and "true," because you want the cumulative distribution from negative infinity all the way up to 10.15 mm.

Of course, we are aware that there can be no negative sizes and even if there could be, they would not go all the way down to negative infinity. But the practicality of the situation is that the areas under the curve at the extreme lower end do not amount to much. We willingly apply the normal distribution with "a wink and a nudge".

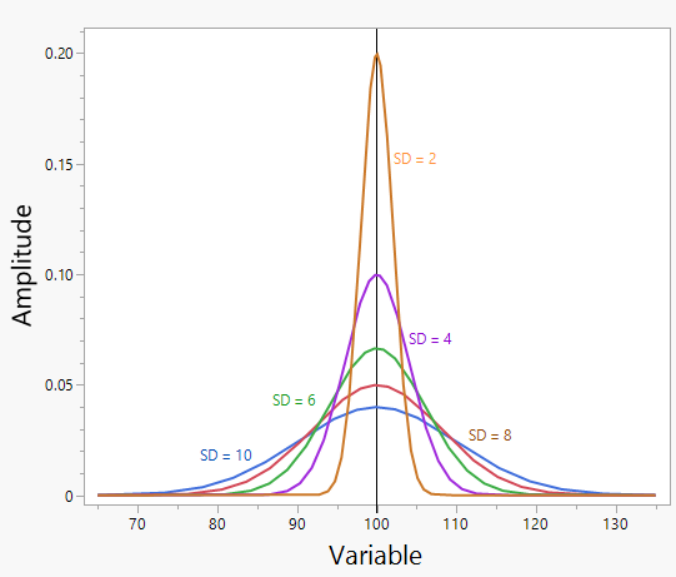
The Excel statement is =NORM.DIST(10.15,10,0.1,TRUE) , and it shows a result of 0.933 or 93.3%. And if you wanted to know what percentage of parts falling between 9.85 and 10.15, you could enter =NORM.DIST(10.15,10,0.1,TRUE)-NORM.DIST(9.85,10,0.1,TRUE). The result is 86.6%.

Figure 5.8 Finding Areas Under the Normal Distribution Curve, (a) Cumulative for Percentage of the Distribution below 10.15 and (b) for the Percentage of the Distribution between 10.15 and 8.95



(a) (b)

Figure 5.9 Normal Distributions Centered on 100, with Standard Deviations of 2, 4, 6, 8 and 10



5.2.5.2 Lognormal Distributions

Q. What do microbiologists, epidemiologists, and seismologists have in common, other than "ists"?

A. They all deal commonly with lognormal distributions.

Microbiologists, for example, plate out samples in a sequence of dilutions, 1 in 10, 1 in 100, 1 in 1000, and so on, deliberately so they can find a dilution with an organism count easily obtained. Then they "un-dilute" mathematically, so they can arrive at a count of the organisms in a linear scale. If they find 35 organisms in a 1 in 1000 dilution, they would declare 350,000 organisms or 35×10^4.

Epidemiologists recognize that the spread of organisms follows certain doubling times. Seismologists characterize the strength of tremors and earthquakes using the Richter scale which is set out in powers of ten. A quake of 6 on the Richter scale has amplitude of its tallest wave on a seismograph ten times higher than one of 5.

Simply put, these three sciences and many others deal with distributions whose logarithm is normally distributed. If x is positive and lognormally distributed, then y = log(x) is normally distributed.

Technically, the probability density function for the lognormal distribution is

$$f(x;\mu,\sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\log x-\mu)^2}$$

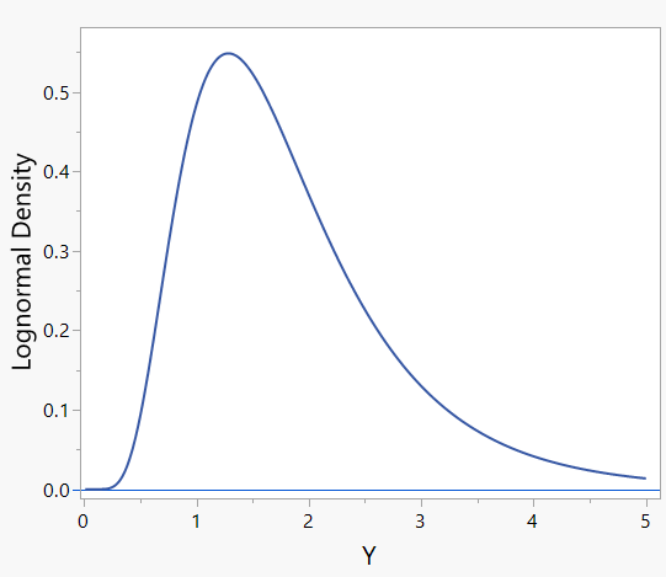
where $-\infty < \mu < \infty$ and $\sigma > 0$. If the lognormal distribution is standardized so that the location parameter is zero and the scale parameter is 1, then the mean is

$$e^{0.5\sigma^2}$$

and the standard deviation is

$$\sqrt{e^{\sigma^2}\left(e^{\sigma^2}-1\right)}$$

Figure 5.10 A Lognormal Probability Density for a Variable with a Location Parameter 0.5 and Scale Parameter 0.5



For many practical applications, users simply take logs of data demonstrated to derive from a lognormal distribution, perform the needed calculations and back-transform in order to estimate parameters of the original distribution. Doing so lacks statistical rigor, but it is often expedient.

5.2.5.3 Exponential Distributions

Like lognormal distributions, exponential distributions are favorites among those engaged in reliability studies. In those applications, they are used to model data with a constant failure rate.

The probability density function is

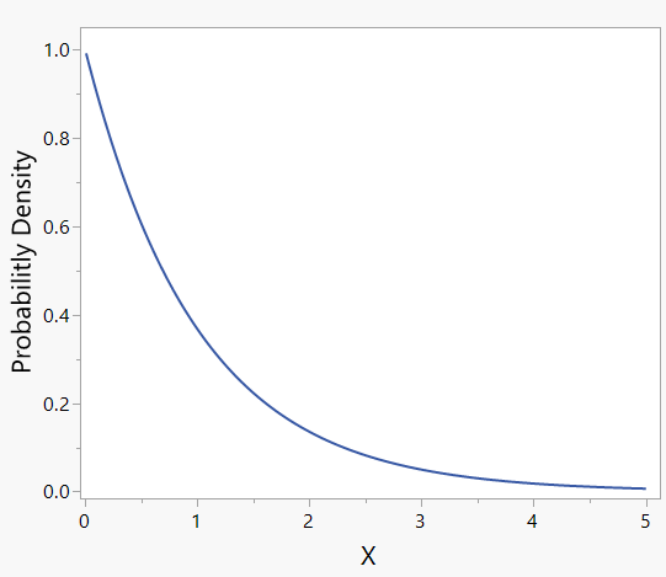
$$f(x) = \frac{1}{\beta} e^{-(x-\mu)/\beta}, x \geq \mu; \beta > 0$$

Here, μ is the location parameter and β is the scale parameter.

As a mild digression, notice that part of this density function looks suspiciously like a component of the normal distribution density function. That is because the normal distribution density function is back to back exponential distributions modified so that the area under the resulting curve integrates to 1.0.

By itself, the exponential distribution is not wholly adequate to describe or predict reliability in all situations. It lacks full flexibility to accommodate situations where, factors such as the effect of age or subject memory influence reliability. Its major flaw in this regard is that it assumes a constant rate of decay or failure over the life of the subject.

Figure 5.11 An Exponential Probability Density for a Variable with a Location Parameter 0, Scale Parameter 1



However, it is a good start for probing into the reliability field. Early applications include animal studies of chronic and infectious diseases and electronic component failures.

5.2.6 Sampling Distributions

Suppose you were to sample from a population, and you were to examine a summary statistic. It could be the sample mean, or the median, or the 82nd percentile or anything else that captures your interest. If you were to do that repeatedly a large number of times, you would end up with data representing the sampling distribution of that chosen statistic.

Of course, you are not likely to do that, but you can conceptualize it. Likewise, you can conceptualize the sampling distribution of some statistic. The notion, though seemingly impractical, is important to the concept of statistical inference.

5.2.6.1 Distributions of Averages

Doubtless, the most common sampling situations are those engaged in pursuit of the mean. In that endeavor, many samples are taken and the sample mean is calculated. The intent, of course, is to estimate the true but ever elusive population mean. It would seem only fair that the more samples taken the better the mean is known. In other words, larger sample sizes should yield more precise estimates of the mean.

It turns out that when means are taken, the collective means form an estimate of the population mean but the standard deviation among the collective means estimates the population standard deviation divided by the square rout of the sample size going into each mean.

Figure 5.12 Distributional Frequencies of Individual Observations and of Means of Samples of Size 4 from a Population with Mean 100 and Standard Deviation 5

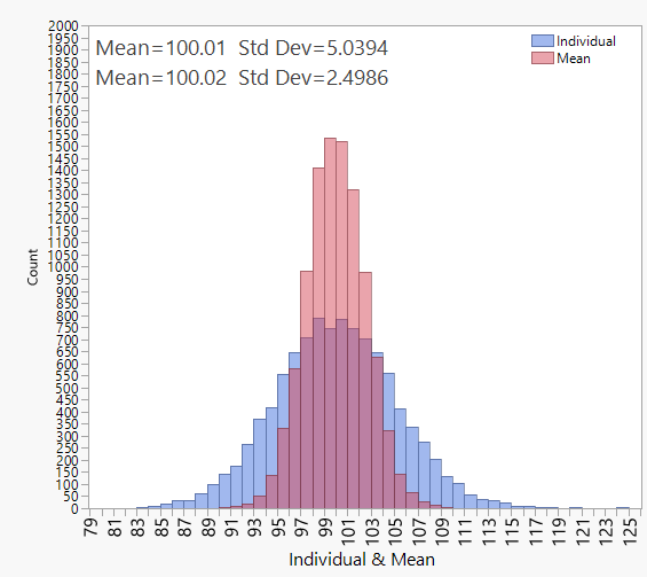


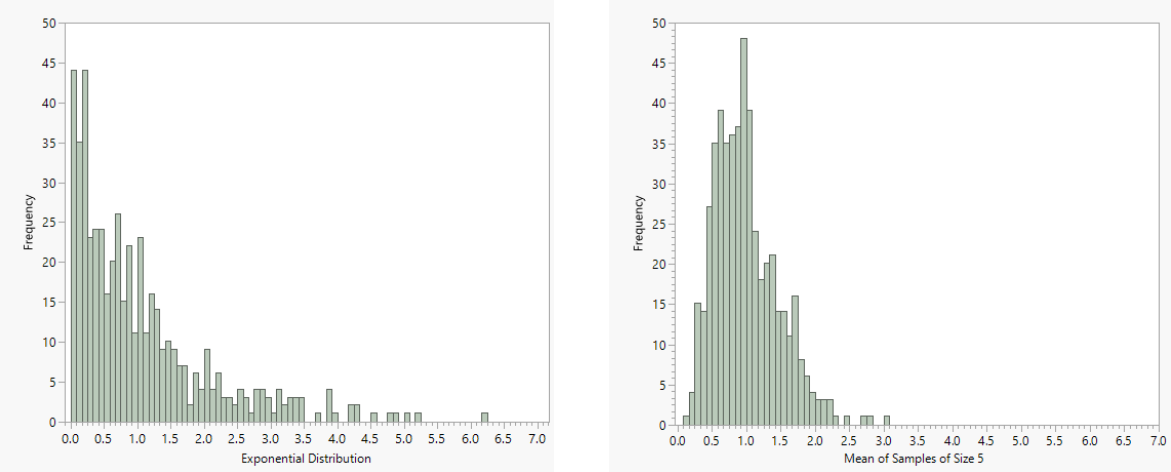
Figure 5.12 illustrates this point. The population described in the caption contains 10,000 observations. The distribution of individuals, as shown in blue, is more dispersed than the distribution of means shown in red. Predictably, its standard deviation is half the size of the population standard deviation: $\sqrt{4} = 2$.

5.2.6.2 The Central Limit Theorem

If you sample from any distribution repeatedly and calculate sample averages, the averages will tend toward normality. Additionally, the mean of the averages will tend toward the mean of the population.

This is true even for a basic distribution as skewed as the exponential distribution as illustrated in Figure 5.13.

Figure 5.13 Histograms Resulting from Single Sampling of an Exponential Distribution (Left) and from the mean of 5 Samples (Right)



Generally, $s_{\bar{x}} = s/\sqrt{n}$ where $s_{\bar{x}}$ is the standard deviation of the mean, s is the standard deviation and n is the sample size. The sample mean of means tends toward the population mean.

In terms of the ideal, using Greek notation, we write

$$\mu_{\bar{x}} = \mu,$$
$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

in place of the Roman letter notation which designates estimates as explained in Section 5.2.3.

5.2.6.3 The t-Distribution

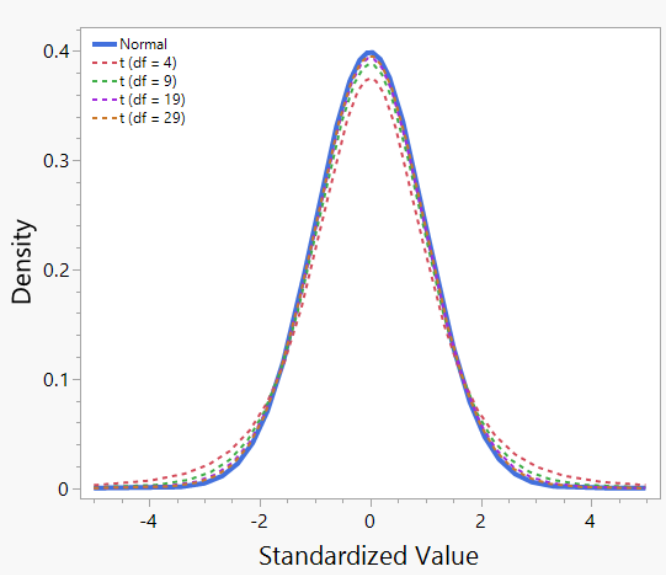
We leverage the central limit theorem for statistical inference to detect differences that might be important to success, however measured. Increasing sample sizes clears the fog of variation, enabling us to see those differences more clearly.

At the dawn of modern day Statistics, there was heavy reliance on the normal distribution and the central limit theorem, and samples upwards of 200 were not unusual. In fact, they were welcome because large sample sizes were needed for precise estimates of the standard deviation.

As applications of statistical inference increased, the need for accurate decision making based on small sample sizes did also. The problem was how to get past the uncertainty associated with the standard deviation estimate. William Gossett (1876-1937), a mathematically gifted chemist, came up with a solution. He developed the Student's t-distribution, using a pen name rather than his own because of his employer's ban on publications. (For more on this interesting story see Salsburg, 2001.)

Gossett's t-distribution accounts for the uncertainty caused by not fully knowing σ, but instead estimating it via s, the sample standard deviation. Our measure of knowledge of the standard deviation is "degrees of freedom" which is the number of observations minus 1. As the number of degrees of freedom increases, knowledge of σ increases, and so Gossett's t-distribution approaches the normal distribution. For smaller sample sizes, the density broadens, reflecting increased uncertainty.

Figure 5.14 The Normal and Selected t-Distributions with Corresponding Degrees of Freedom



In practice, to find the standard distance away from center in a t-distribution we would use

$$t = \frac{\bar{x} - \mu}{\bar{s}}$$

Or in the two sample case which is far more common, representing two groups or populations,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

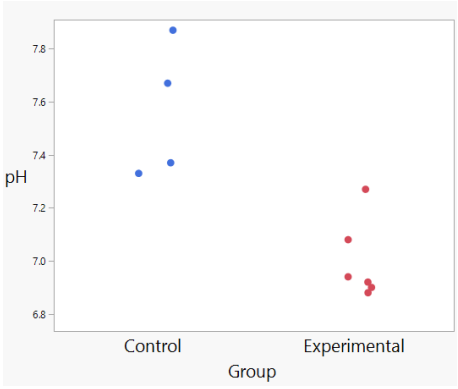
Here, s_p represents the "pooled" standard deviation under the assumption that the standard deviations among replicate observations of the two groups in question come from the same population of standard deviations.

For example, suppose we want to compare two types of dairy blends for pH. We have 4 control and 6 experimental batches taken at random with data shown in Table 5.3.

Table 5.4 pH of Two Types of Dairy Blends

Run	Control	Experimental
1	7.67	6.88
2	7.33	6.94
3	7.37	6.92
4	7.87	6.90
5		7.08
6		7.27
Mean	7.56	7.00
Std. Dev.	0.26	0.15

Figure 5.15 Plot of pH Data in Table 5.3



The pooled standard deviation is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(4 - 1)0.26^2 + (6 - 1)0.15^2}{4 + 6 - 2}} = 0.20$$

And the t-statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{7.56 - 7.00}{0.20\sqrt{\frac{1}{4} + \frac{1}{6}}} = 4.41$$

What does that mean? Suppose we went into this test thinking the pH could go either way; that is, we would be just as surprised if the treatment exceeded the control as if it did not. In that case, we would be looking at the size of the combined tails of the t-distribution at ± 4.41 standard

deviations from the mean. (In Excel, use =T.DIST.2T(4.41,8).) It is tiny: 0.002. It is not likely that this difference happened by chance.

The world is never as simple as merely comparing two groups. R.A. Fisher (1890-1962), a mentor of Gossett's developed methodology for analyzing and interpreting data sets with increased treatment levels and increased numbers of factors. Please refer to Chapter 2, Section 6.

Section 5.2 References

Hoerl, R. and Snee, R., Statistical Thinking, 3rd ed., Wiley, Hoboken, NJ, 2020.

Johnson, N.L., Kemp, A.W. and Kotz, S. Univariate Discrete Distributions, 3rd Ed., John Wiley & Sons, Hoboken, N.J, 2005.

Johnson, N.L., Kotz, S. and Balakrishnan, N. Continuous Univariate Distributions, 2nd Ed., John Wiley & Sons, Hoboken, N.J, 1994.

http://www.itl.nist.gov/div898/handbook/2013

Salsburg, David. The Lady Tasting Tea, Holt, New York, 2001.

Section 5.3 – Inferences on Parameters and on Predictions

5.3.1 Objectives

The purpose of this section is to differentiate between the main types of statistical inference, including point estimation, interval estimation, and hypothesis testing.

5.3.2 Outline

We begin by revisiting the concept of statistical inference. Next, we explain the main types of statistical inference: point estimation, interval estimation and hypothesis testing. We then provide further details on point estimation. The following sections in this chapter go into detail on interval estimation and hypothesis estimation; hence they are not covered further in this section.

5.3.3 The Three Main Types of Statistical Inference

Recall that the term *statistical inference* refers to drawing conclusions about an overall population or process of interest based on only a sample, or subset of the population. Figure 5.1 illustrated this concept. As a practical example, we referred to election polling, in which pollsters may interview only 800 or so likely voters, and then try to make inferences about how the election will go when all those who are actually going to vote do so.

The three main ways that people typically want to draw inferences about an entire population are:

- Point estimation
- Interval estimation
- Hypothesis testing

These are the three main types of statistical inference utilized in applications of statistical engineering. While these are the main types, they are certainly not the only ways in which one might try to draw inferences about an entire population from a sample.

For example, we might want to use a sample to determine the specific relationship between variables in the population. In this case, we do not have an initial functional form that we suspect, but rather we wish to determine this through analysis of the sample data. This problem is certainly an inference problem, in that we want to infer about the relationship between variables in the entire population. However, it is not point or interval estimation, nor hypothesis testing. In addition, it is common in machine learning applications to develop models from sample data that we hope will predict well within the entire population, but we do not care about the form of the model. That is, in such applications we do not actually care about the exact nature of the relationships, as long as the model predicts well. This is also a form of inference.

The term *point estimation* refers to the use of a sample statistic, a single value or "point," to estimate a population parameter. As we have seen previously in this chapter, sample statistics are typically represented by Roman (Latin) letters, while population parameters are typically represented by Greek letters. So, we might calculate the sample mean, $\bar{y}$, and use this to estimate the mean of the entire population, μ. In election polling, we typically calculate the proportion in our sample who say they will vote for candidate Jones, represented by the letter p, and use this to estimate the proportion of people in the entire population of actual voters who will vote for Jones, represented by π.

Point estimation is not always quite so straightforward. For example, we might wish to estimate the difference between the proportion of votes for Jones versus candidate Smith, or perhaps the number of Electoral College votes Jones will receive. Since both of these examples involve estimating a single number, they are also point estimates. Note that, by definition, point estimates only involve a single number, not a measure of uncertainty on that number. This is where interval estimation comes in.

Interval estimation provides not only a single number, but rather an entire range of numbers, or interval, in which we think a population parameter is likely to fall. As a common example, when election polls results are presented in the media, the result is often qualified by saying something similar to: "This poll has a margin of error of +/- 4%." What exactly does this statement mean? It suggests that the point estimate, let us say it was 52% for Jones, could be inaccurate by as much as 4% either way. In other words, while 52% is our point estimate for the percentage to vote for Jones, the poll suggests that the actual percentage Jones will receive on election night could be anywhere from 48% to 56%. As in this case, interval estimates are generally more informative, because they document how far off we might be in our inferences about the population. In this case, we are not yet ready to call the election for Jones, because 48.5% is a plausible election night result, based on this poll.

There are several types of interval estimates, which are explained in more detail in the next section. These include *confidence intervals*, which document uncertainty on our point estimates of population parameters, *prediction intervals*, which document uncertainty in predicting new values outside our sample, and *credible intervals*, which are similar to confidence intervals, but integrate prior information into the estimation process, using a methodology known as Bayesian estimation.

In some cases, the key inference we need to make about the population is not a point estimate or even interval, but rather a binary conclusion about the population. For example, suppose in a sample of ten faculty members at a major university, the six male professors in the sample make an average of $5,000 more than the four female professors. Does this prove beyond a reasonable doubt that the university is discriminating against women? After all, whenever comparing male versus female salaries in a sample, there is a 50/50 probability that, just by chance, the female average salary will be lower. Hypothesis tests typically provide a quantitative measure of how unlikely a result is, assuming in this case that there is no discrimination. That is, if the university were completely fair is granting salaries, how unusual is it to find a discrepancy this large in a sample of 10? This is typically determined using probability theory. If the result would be extremely unusual assuming fairness, this constitutes evidence of discrimination. Conversely, if

the result would occur frequently just by chance, then the evidence is thin, and does not lead to a conclusion that the university is discriminating.

Such evaluations are referred to as *hypothesis testing*, because we are considering a hypothesis, such as fairness in salaries ($\mu_{male} = \mu_{female}$), and then evaluating the degree to which the available data are or are not consistent with the hypothesis. Such hypotheses do need to be put into numerical terms, and then can be rigorously evaluated using the methods discussed in the hypothesis testing section.

5.3.4 Point Estimation

As noted in the previous section, point estimation means estimating a population parameter with a single value, or "point." In some cases, the best way to do this is obvious. For example, the most common point estimate for the population mean, μ, is the sample mean. We typically designate the point estimate of a population parameter by adding a "hat" (French accent symbol) on top of it. So, our sample point estimate of μ is written $\hat{\mu}$. If the variable of interest is y, then the sample mean is written $\bar{y}$. Therefore, $\hat{\mu} = \bar{y}$. The sample mean, of course, is simply the sum of all the observations (assume we have "n" of them), divided by n:

$$\bar{y} = \sum_{i=1}^{n} y_i/n$$

There are other situations where choice of the best point estimate is not as obvious. For example, the point estimate of the population variance, σ^2 is typically the sample variance, calculated as:

$$s^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)$$

Note that the divisor is n-1 and not n. One can estimate σ^2 by dividing this numerator by n, but it has a disadvantage. The *expected value* of s^2 using a divisor of n, that is, the long-term average value of s^2 that we would obtain if we took a large number of samples, estimated s^2 from each, and then averaged all the s^2 values, is biased on the low side. That is, the expected value of s^2, written as $E[s^2]$ is smaller than σ^2. Mathematically, we would say $E[s^2] < \sigma^2$. When the expected or average value of a point estimate is not equal to the population parameter we are trying to estimate, it is referred to as a *biased estimate*. Fortunately, if we divide by n-1, then $E[s^2] = \sigma^2$. We therefore say that the point estimate of σ^2 using the sample variance, dividing by n-1, is *unbiased*.

Finding an unbiased point estimate is desirable, all other things being equal. Even if the point estimate is unbiased, we would still like the uncertainty in the estimate to be small. That is, we would like the confidence interval for the population parameter to be as narrow as possible. Sometimes, we can obtain a narrower confidence interval by using a biased point estimate. This complicates selection of the "best" point estimate.

In formal models, such as those discussed in Chapter 4, including regression and analysis of variance models, selection of point estimates is a little more complicated. For example, suppose we wish to estimate the population parameters (β's) in the following linear equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \varepsilon,$$

where y is the response of interest, and ε represents the random error. The most common method used to obtain point estimates for such linear models is called *least squares*. That is, we obtain that set of estimates of the population parameters, the $\hat{\beta}_i$, that minimize the squared deviation of predicted values of y ($\hat{y}_i$) from the actual values of y (y_i). That is, we minimize the sum of squared residuals. The predicted y values are calculated as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots \hat{\beta}_k x_k$$

The sample residuals are then calculated as:

$$e_i = y_i - \hat{y}_i$$

The residual sum of squares, which we select the point estimates ($\hat{\beta}_i$) to minimize, is:

$$\sum_{i=1}^{n} (e_i^2) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Calculus is typically used to solve for these least squares estimates. See Montgomery et al. (2012) for details on obtaining least squares point estimates for population parameters in such linear models.

For more complex models, including non-linear models, another method used for obtaining point estimates is called *maximum likelihood*. This approach utilizes probability functions to obtain a *likelihood function* for the data, and then solves this equation for that set of point estimates that would maximize this likelihood function. In essence, it solves for the set of point estimates that would make the data actually observed most likely to have occurred. For example, if we observed 5 heads out of 10 flips of a coin, the value of π, the probability of a head, which would make this result most likely to occur is .5, so this is the maximum likelihood estimate. See Montgomery et al. (2012) for more detail on maximum likelihood estimation.

5.3.5 Summary of Key Points

- The term *statistical inference* refers to several individual approaches to drawing conclusions about a population from a sample.
- Three of the most common methods used in inference are *point estimation*, *interval estimation*, and *hypothesis testing*.
- Point estimation refers to obtaining an individual value (point) that in some sense provides the best estimate of a population parameter, such as a mean.
- Interval estimation refers to obtaining an interval of values within which we expect to observe a population parameter or future sample statistic.
- Hypothesis testing refers to determination of how likely (probable) the sample results would have been if, in fact, our original hypothesis were true. If this probability is very small, i.e., unlikely, this provides evidence that the hypothesis is in fact false.

Section 5.3 References

Montgomery, D.C., Peck, E.A., and Vining, C.G., *Introduction to Linear Regression Analysis*, 5th ed., John Wiley & Sons, New York, 2012.

Section 5.4 – Statistical Intervals

5.4.1 Objectives

This section guides the development of intervals to quantify the uncertainty associated with parameter estimates derived from summary statistics derived from data gleaned from populations.

5.4.2 Outline

There are many different types of statistical intervals, including confidence intervals and prediction for various distributions and associated parameters. This section covers only a few which are believed to be in greatest practical use. A thorough treatment is provided by Hahn and Meeker (1991). Here, for continuous distributions, we discuss briefly confidence and prediction intervals for the mean, confidence intervals for the standard deviation and for combinations of means. We move on to confidence intervals for proportions. This is followed by a discussion of an entirely different approach, namely the use of Bayesian statistics to arrive at "credible intervals." Finally, we cover tolerance intervals.

5.4.3 Confidence Intervals for the Mean

It would seem that inferences made consistent with the provisions on the opening section of this chapter should be accompanied by a statement of precision.

To that end, a well-accepted statement comes in the form of a confidence interval. It first appears in a publication by Jerzy Neyman (1937), a highly influential Polish mathematician who worked in Warsaw, University College, London and finally at the University of California, Berkeley. (See Chapter 2, Section 6.)

The confidence interval is the sample mean plus or minus the t-statistic times the standard deviation, all divided by the square root of the sample size. Algebraically:

$$\bar{x} \pm \frac{ts}{\sqrt{n}}$$

Terms are defined in Section 5.2. Note that the t-statistic is chosen to correspond to the desired level of confidence and degrees of freedom.

For example, a look back at Table 5.4 shows the Experimental group of pH readings has a mean of 7.0 and a standard deviation of 0.15 based on 6 observations. A 95% confidence interval about the mean is

$$7.0 \pm \frac{(2.571)(0.15)}{\sqrt{6}} = 7.0 \pm 0.16, \text{ or the interval from 6.84 to 7.16.}$$

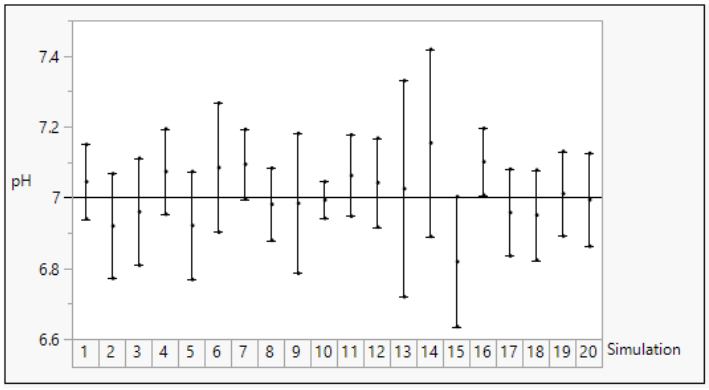
If you follow along, you will notice that the t-value of 2.571 comes from the t-distribution with 5 degrees of freedom and with 2.5% of the area under the curve in each tail.

We can do the math, but the question of interpretation remains. What does the pH interval "from 6.84 to 7.16" mean? The convention is to say that we can be 95% certain that the "true" pH lies between these two numbers. This is commonly in use, and it serves most practical purposes.

So the reader is aware, there are some technical difficulties with the conventional statement. In theory, the 95% confidence interval means that if we were to rerun the experiment many times and do the math to create the interval each time, 95% of the intervals would include the true mean. This can be seen in Figure 5.16 which was created by simulation as the caption describes. Simulation 16 misses the true value. A few others squeak by. As chance would have it, 95% of them cover the true mean!

Of course, this logic is pure folly. We are not going to rerun the experiment many times. We are not going to rerun it at all. So we are forced to follow convention.

Figure 5.16 95% Confidence Intervals for the Mean Produced from 20 Simulations of pHs with Mean 7.0 and Standard Deviation 0.15



5.4.4 Prediction Interval for One Observation

Occasionally it is useful to know about a confidence interval for a predicted future observation from a population previously sampled. The expression is

$$\bar{x} \pm t\sqrt{s^2 + \frac{s^2}{n}}$$

Remaining with the data from Table 5.3, Experimental group, to predict an interval for one additional observation, we would have

$$\bar{x} \pm t\sqrt{s^2 + \frac{s^2}{n}} = 7.0 \pm 2.571\sqrt{(0.15)^2 + \frac{(0.15)^2}{6}} = 6.58 \text{ and } 7.42.$$

This interval is necessarily wider than that shown above because of the original uncertainty of the mean, coupled with the uncertainty of the future observation, itself.

5.4.5 Confidence Intervals for Combinations of Means of Variables

In a linear combination such as

$$y = c_0 + c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

where the cs are constant and the μs are parameters to be estimated by statistics, we know that

$$\mu_y = c_0 + c_1\mu_1 + c_2\mu_2 + \cdots + c_n\mu_n, \text{ and that}$$

$$\sigma_y^2 = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_n^2\sigma_n^2.$$

$$\text{So, } \sigma_y = \sqrt{c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_n^2\sigma_n^2}.$$

This tells us that the mean of the sum of random variables, like the means discussed above, is the sum of the means. It follows that the difference between two random variables is simply $\bar{x}_1 - \bar{x}_2$. The cs in the first equation above may be positive or negative.

The variance or square of the standard deviation of random variables is the sum of their variances, and they are all positive whether the cs are positive or negative because when the terms are squared; the result is always positive.

All the above assumes the variables are independent, meaning that knowledge of one imparts no knowledge of the other.

The discourse above has direct applicability to creating a confidence interval for the difference between two means.

$$(\bar{x}_1 - \bar{x}_2) \pm ts_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

From the data in Table 5.5 and the calculations that follow it, we have

$$(\bar{x}_1 - \bar{x}_2) \pm ts_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (7.56 - 7.00) \pm (2.365)(0.20)\sqrt{\frac{1}{4} + \frac{1}{6}} = 0.27 \; and \; 0.86$$

Note that the t-statistic is computed with 8 degrees of freedom and with 2.5% of the distribution in each of the two tails.

The conclusion from Section 5.2 is that the difference in pH between the two groups likely did not happen by chance alone. With the current effort we have a 95% confidence interval about that difference. It is more informative than the simple declaration that a difference exists. Of course, it is subject to the cautions of interpretation discussed in the previous sections.

5.4.6 Confidence Interval for the Standard Deviation

Variances, the squares of standard deviations, follow a distribution of the sum of squares of random variables such as those appearing in the exponent of the normal distribution probability density function (Section 5.5.5.1). The distribution function, itself, is not a pretty sight and is not displayed here, but it can be found in texts on statistical theory. Its probability points for associated degrees of freedom can be found in most beginning statistics texts and in statistical software. It is referred to as the Chi-squared distribution (a.k.a. the Chi-square distribution), and it has many applications, only one of which is pursued here. That is to serve as a base of the confidence interval for the standard deviation.

The confidence interval defining inequality for the standard deviation is

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]^{\frac{1}{2}} \leq \sigma \leq \left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]^{\frac{1}{2}},$$

where the notation used is the same as that in preceding sections with the addition of the χ^2 values at upper and lower probability points, indicated by subscripts.

As an example, we can follow through with the Experimental group in Table 5.3. Its standard deviation estimated from 6 observations is 0.15. Requesting 95% confidence ($\alpha = 0.05$) and applying the inequality above, we have

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]^{\frac{1}{2}} \leq \sigma \leq \left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]^{\frac{1}{2}} =$$

$$\left[\frac{(6-1)(0.15)^2}{12.83}\right]^{\frac{1}{2}} \leq \sigma \leq \left[\frac{(6-1)(0.15)^2}{0.831}\right]^{\frac{1}{2}} or$$

$$0.094 \leq \sigma \leq 0.368$$

You have probably noticed that the interval about the estimated value is asymmetrical or skewed to the high side. That is because the Chi-squared distribution is skewed as a result of its addressing the behavior of squared values. Squaring small numbers makes them larger, to be sure, but squaring larger numbers makes them much larger yet, and the exaggeration is not fully corrected by extracting the square root.

5.4.7 Confidence Intervals for Proportions

Confidence intervals for proportions follow the same format as those for the mean and standard deviation. Intervals are composed of the statistic surrounded by distributional probability points times the variation estimate. For proportions, especially when the sample sizes are large, the distribution of p, the estimate, is nearly normal.

For a single proportion we use

$$p \pm z\sqrt{\frac{p(1-p)}{n}},$$

where z is the standard normal deviate. For 95% confidence it is 1.96.

In Section 5.5.3, we discussed a sampling of 500 Baroque music manuscripts in which 150 contained opening movements written in 6/8 time. The resulting proportional estimate is p = 0.3, and a 95% confidence interval is

$$p \pm z\sqrt{\frac{p(1-p)}{n}} =$$

$$0.30 \pm 1.96\sqrt{\frac{0.3(0.7)}{500}} \; or$$

0.26 and 0.34

To place a confidence interval on the difference between two independent distributions we use

$$p_1 - p_2 \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Be careful with this expression. It cannot be used to compare electoral survey of two opposing candidates, for example, tempting as that may be. Those results are not independent.

5.4.8 Credible Intervals

The term Credible Interval refers to an uncertainty interval based on a Bayesian procedure. This contrasts with the classical confidence interval discussed in the previous sections. One important characteristic of the Bayesian calculation of a credible interval is that it conveys a direct coverage probability for the parameter of interest. The interval is conditional on the observed data, a data model and an assumed prior distribution of the parameter.

The Bayesian view considers the parameter as a random quantity, whereas the classical view considers the parameter as a fixed quantity and a confidence interval on the parameter as a random outcome. The classical confidence level expresses a long term coverage probability of repeated calculations of a confidence interval from similar experiments. The confidence level does not imply a probability applying to any particular calculated confidence interval. It is beyond the scope of this section to delve more deeply into the differences between the classical and Bayesian inferential procedures. Suffice it to say that the two are based on different philosophies which lead to different interpretations and different calculations of their respective interval estimates.

Under certain limited cases, the calculations can lead to identical interval bounds, but that is not true in general. Both the classical and Bayesian approaches are valid inferential procedures and have important advantages and disadvantages in practice.

A necessary requirement for the calculation of a credible interval is a proposed prior distribution for the parameters in the data model. The prior distribution is intended to express our best guess or judgment about the values which we believe the parameter might reasonably take. If we have sufficient information from previous experiments, literature references, or expert knowledge, we might propose an "informative" prior distribution to capture that knowledge. If we lack prior knowledge, we can propose a "vague" or "non-informative" prior distribution for the parameter.

The role of the prior distribution is critical and can have substantial effects on the calculations, so the choice of prior distribution must be handled with care. The Bayesian paradigm then incorporates the prior distribution together with the likelihood from the data model according to Bayes' rule to yield a "posterior" distribution of the parameter. The posterior distribution is conditional on the data (the likelihood) and the prior distribution of the parameter. In this sense, Bayes' rule is a mechanism for updating our current knowledge of the state of nature with the data we have just observed.

Bayes' rule as the fundamental basis for Bayesian calculations can be expressed succinctly as follows:

$p(\theta|y) \propto p(y|\theta)p(\theta)$ = Likelihood x Prior distribution, where

$p(\theta|y)$ is the posterior density of the parameter θ conditional on the observed data y which is considered fixed now in the Bayesian paradigm

$p(y|\theta)$ is the sampling distribution and
$p(\theta)$ is the prior distribution of the parameter θ

The grand superstructure of Bayesian methodologies rests on this deceptively simple but elegant expression.

The posterior distribution is found through an integration across the entire parameter space. Unfortunately, this has a closed form solution only for a small set of relatively simple models.

For most models that we typically encounter in practice, for example nonlinear models and mixed-effects models, the posterior distribution has no exact integration. So simulation methods known as Markov Chain Monte Carlo (MCMC) must be employed. In these cases, no exact inference is possible. However the numerical random sampling methods to carry out the integration have been shown to yield reasonable approximations to the posterior distribution. The MCMC algorithms are by no means trivial or guaranteed since the simulation must converge to the stationary (target) distribution (posterior distribution) in order to yield valid estimates. Assessing the convergence is very important in practice and requires some care.

In the remaining part of this section we will use the data of Table 5.4 to compare the Bayesian credible interval under two different scenarios with the confidence interval calculated previously. The data model contains 3 parameters: the means of groups 1 and 2, say θ_1, θ_2, and the common within group variance σ^2, assuming normality. We can write the statistical model as follows:

$$y_{j(i)} = \theta_i + \varepsilon_{j(i)}, i = 1,2; j = 1,2,...,n_i,$$
$$\varepsilon_{j(i)} \sim N(0,\sigma^2).$$

where n_i is the sample size of the i^{th} group, where i=1=Control, and i=2=Experimental group. . The classical procedure relies on the likelihood function to derive estimators of the model parameters of interest subject to the distributional assumptions of the residual term. Those estimators lead directly to the calculated values shown in Table 5.5.

We turn now to the Bayesian approach for comparison purposes. Let us begin with a Bayesian credible interval calculation assuming some information on the group means but less prior information on the residual variance corresponding to measurement method uncertainty. In this scenario, suppose the scientist who ran the experiment reported that the mean pH across a series of similar studies is 7.25. It is also known that the overall range of pH is limited to [0 – 14]. A relatively uninformative prior distribution on the group mean parameters would include a large variance to span the full pH range. We can express this prior distribution on the group means as a normal distribution with mean 7.25 and variance 2. The prior distribution for the error variance can be expressed through an uninformative distribution through an inverse gamma distribution with scale and shape parameters 2 and 2, respectively. (See Gelman, 2006 and Gelman, et.al., 2013 and their references.)

Given these assumptions regarding the prior distributions of the 3 parameters in the data model, and conditional on the observed data, the Bayesian calculation leads to the values given in the column headed Bayesian 1 in Table 5.5. Note that the means are similar but the credible interval is much for all parameters, reflecting the influence of little prior information on the likely range of the variance parameter prior distribution. A comparison of the prior distribution of the treated mean is shown in Figure 5.18 and a similar comparison of the variance parameter distributions are shown in Figure 5.19

Figure 5.18 Comparison of prior distribution with posterior distribution with a realtively uninformative prior distribution for the Experimental group mean of Table 5.4

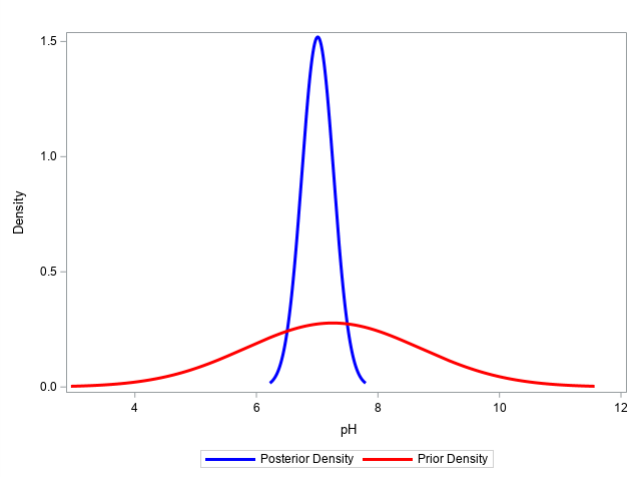
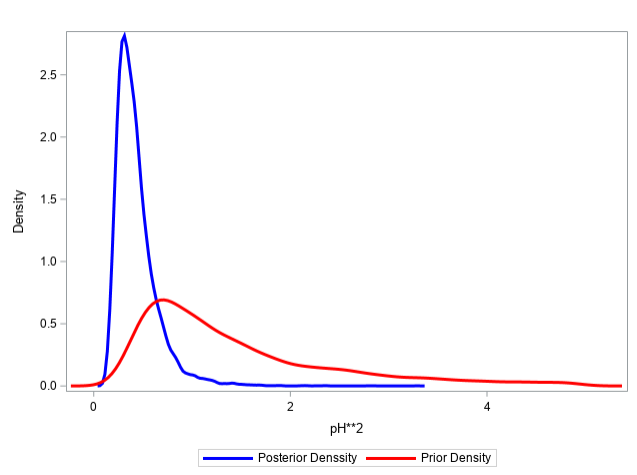


Figure 5.19 Comparison of prior distribution with posterior distribution with a realtively uninformative prior distribution for the residual variance parameter of the Experimental group of Table 5.4



We turn now to a scenario where we have good prior information about an expected range for the group means and the measurement uncertainty. This is especially important in this case because of the limited data collected, where we have only 4 observations in the control group and 6 observations in the experimental group. In cases like this, prior knowledge can benefit the statistical results. Suppose that prior elicitation of scientific judgment from the experimenter reported that the group means are not likely to vary by more than 3 units in either direction from the mean 7.25 again based on similar previous experiments. In addition, the experimenter reports that method variability is not expected to exceed 0.5 standard deviations in pH units. This information can then be captured through more informative prior distributions. The prior on the group means can be expressed through a normal distribution with mean 7.25 and variance 1, The error variance can be expressed through an inverse gamma distribution with scale and shape parameters 4 and 0.4 respectively, related to the sample size and residual variance observed from prior experimentation.

Given these assumptions, and conditional on the observed data, the Bayesian calculation leads to the values given in the column headed Bayesian 2 in Table 5.5. Note that the means are similar to the classical estimates, and the credible intervals are not much wider, again reflecting the influence of better prior information on the likely range of the parameter distributions. A comparison of the prior distribution of the treated mean is shown in Figure 5.20 and a similar comparison of the variance parameter distributions are shown in Figure 5.21.

Figure 5.20 Comparison of prior distribution with posterior distribution with a realistic prior distribution for the Experimental group mean of Table 5.4

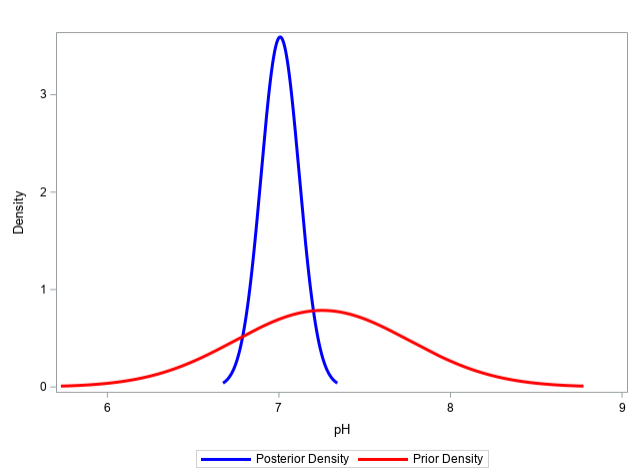


Figure 5.21 Comparison of prior distribution with posterior distribution with a realistic prior distribution for the for the residual variance parameter of the Experimental group of Table 5.4

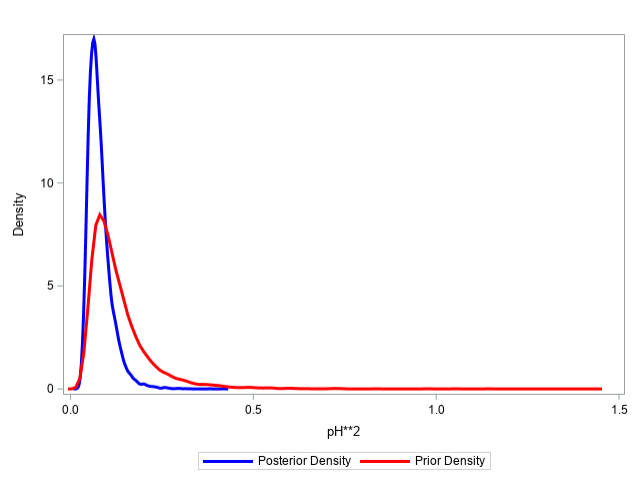


Table 5.5 Comparison of 95% Credible Intervals with 95% confidence Interval

	Classical		Bayesian 1		Bayesian 2	
Parameter	Mean	Interval	Mean	Interval	Mean	Interval
Control	7.56	7.40 – 7.72	7.54	6.89 – 8.18	7.54	7.26 – 7.80
Experimental	7.00	6.84 – 7.16	7.01	6.52 – 7.56	7.01	6.79 – 7.23
Difference	0.56	0.27 – 0.86	0.53	-0.32 – 1.33	0.53	0.18 – 0.87
Variance	0.039	0.018 – 0.142	0.431	0.153 – 0.827	0.079	0.33 – 0.143

5.4.9 Tolerance Intervals

Tolerance intervals cover a fixed proportion of a population with a desired level of confidence. They come in three varieties.

- You may want to know what interval will contain a fixed proportion of the population.
- You may want to know what interval guarantees that a fixed proportion of the population will not fall below a certain limit.
- You may want to know what interval guarantees that a fixed proportion of the population will not fall above a certain limit.

In other words, tolerance intervals define upper and/or lower bounds within which a specified proportion of a population exists, with a fixed level of confidence.

They assume normality, and they are based on two concepts, coverage and confidence. Coverage refers to the desired interval or bounds, and confidence admits to a level of certainty.

The underlying mathematics is complex and incorporates the Chi-squared distribution which we have already ducked, but most statistical software does the calculations for you.

If, for example, we were to examine all 120 observations from the simulation data of Section 5.4.3 and request an interval within which 90 percent of observations will fall with 95% confidence, our statistical software would give the interval from 6.72 to 7.30. These data are listed in Table 5.6, and are made available for reader's confirmation of the tolerance interval shown.

Table 5.6 Simulated pH Data. Bold numbers correspond to individual simulations.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
7.05	6.93	6.79	6.97	7.07	7.36	6.92	6.95	7.13	7.03
7.02	6.67	6.88	7.00	6.82	6.92	7.05	6.83	6.82	7.02
7.19	6.90	6.88	7.19	6.84	7.05	7.15	7.07	7.29	6.97
6.88	7.04	7.19	7.19	6.82	6.91	7.15	7.06	6.88	7.01
7.07	6.91	6.95	7.14	7.14	7.21	7.18	7.05	6.83	6.90
7.05	7.06	7.06	6.95	6.83	7.06	7.11	6.92	6.95	7.02
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
7.15	7.21	7.21	6.74	6.99	7.10	7.03	7.01	6.92	6.79
7.06	7.04	6.47	7.10	6.50	6.94	7.09	6.83	7.05	7.05
7.07	7.15	7.06	7.14	6.77	7.14	6.96	7.10	7.22	7.14
6.89	7.01	7.03	7.39	6.82	7.19	6.74	6.89	6.90	6.91
7.20	6.95	7.08	7.45	6.93	7.07	6.95	7.05	6.98	7.04
7.01	6.90	7.29	7.11	6.90	7.16	6.97	6.82	6.98	7.03

Section 5.4 References

Gelman, Andrew. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis Vol. 1, Number 3, (2006), 515–533.

Gelman, Andrew, Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. Bayesian Data Analysis, 3rd ed., Chapman & Hall/CRC, Boca Raton, 2013.

Hahn, G.J, and Meeker, W.O. Statistical Intervals, John Wiley and Son, Hoboken, NJ, 1991.

Neyman, J. Outline of a theory of statistical estimation based on the classical theory of probability, Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236(767), (1937), 333-380

Section 5.5 – Hypothesis Testing

5.5.1 Objectives

The information presented here is intended to aid in the understanding of the basic thinking involved in testing simple hypotheses and to expand from there to hypothesis testing in the more complex Statistical Engineering world.

5.5.2 Outline

The basic thinking of hypothesis testing is divided into two schools of thought, both productive with one highly mathematical and the other more relevant to success in Statistical Engineering. Readers are encouraged to examine Chapter 4 on modeling and Chapter 6 on data collection.

5.5.3 Basic Principles of Hypothesis Testing

A beginning statistics text could hardly be considered legitimate if it failed to include a chapter on hypothesis testing. Bewildered Stat 101 students' heads are filled with anxiety-inducing lore of Type I and Type II errors and corresponding alphas and betas. By the time they mature to the point of actual application, they only remember the alienation; the concept is forgotten.

Without a doubt the concept and the kind of thinking behind it are important to data driven decisions, but it would be a mistake to get hung up on hypothesis testing as a keystone of statistical thinking and statistical engineering. It is not.

Hypothesis testing is important to the notion of decision making in the shadows of uncertainty. Uncertainty clouds the path, causing some decisions to be in error. A formal structure for minimizing the probability of that error is needed. Jerzy Neyman and Egon Pearson (1933) formulated a structure, building on, but in heated contention with, the ideas of Karl Pearson (Egon's father), William Gossett and Ronald Fisher (see Chapter 2, Section 6). Their differences were never really resolved, but the Neyman-Pearson system receives attention during one's early statistical education, only to recede as statistical experience and education increase.

What is the difference?

5.5.3.1 The Neyman-Pearson School

Neyman and Pearson, in a simple situation, posit two hypotheses, H_0 and H_A, with a stated α, the probability of rejecting H_0, the "null," when it is true, and β, the probability of accepting H_0, when H_A, the alternative, is true instead. Corresponding sample sizes are fixed in advance based on prior beliefs, available resources, etc. Students are taught to determine rejection regions for the hypotheses. If the test statistic falls into the rejection region of H_0, H_A is accepted. Otherwise H_0 is believed to be true unless and until more data prove otherwise. (Not to worry: most statistical software calculates rejection regions based on α, β and the sample sizes.)

In a nutshell, the Neyman-Pearson procedure pays very close attention to mathematical rigor. We discuss tests for the mean, but the procedure is the same for tests involving other parameters.

Its steps are:

1. State the null hypothesis, H_0. Usually, it is set up as something the user is attempting to deny such as, "There is no difference in the active ingredient between these two suppliers."
2. Choose your α-level. This is the probability that you are in error when you reject the null hypothesis. If the consequences in terms of loss to the organization resulting from declaring no difference when, in fact the difference is real and is great, α should be selected at a very low level. Many users simply default to $\alpha = 0.05$.
3. Find the t-value or z-value corresponding to the α-level. Some deliberation is in order here. If you would be just as surprised if, for example, Supplier 1 had higher active ingredients or it had lower active ingredients than Supplier 2, then your test is two-sided, meaning that your test statistic, t or z, would have equal probabilities in both tails of the distribution. Otherwise, your test statistic should be chosen as one-sided, with all the probability in one tail.
4. Calculate the test statistic as t was calculated following Table 5.4. If the calculated value falls in the critical region determined in Step 3, reject the null hypothesis. Otherwise accept it under the condition that it will be believed to be true unless and until data are available to prove the contrary. The probability that you are in error is limited to the declared α-level.

Table 5.7 Neyman-Pearson Style Hypothesis Testing Strategy

		Conclusion	
		Accept the Null Hypothesis	**Reject the Null Hypothesis**
Truth About the Population	**True**	**Correct Decision**	**Type I Error, Probability α**
	False	**Type II Error, Probability β**	**Correct Decision**

5.5.3.2 The Fisherian School

By contrast, the Karl Pearson, William Gossett and Ronald Fisher approach is to set up a null hypothesis such as: The "effect of adding 'eye of newt' to this formula increases viscosity." They report the level of significance attained, and do not hassle about the difference between α = 0.4, 0.5 or 0.6. Instead, they focus attention on low probability events, and hold judgement on others until more information is gained. Rather than exerting a focus on α-values, they seek large main effects, capitalizing on them to drive progress.

At this point, the difference might seem small, but it has large consequences in driving projects forward toward improvement through a strategy of experimentation involving sequences of designed studies. Those situations predominate, and they rarely break down to a decision between two alternatives. Instead, if we think of it in these terms, the null hypothesis would be written

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

against the alternative that at least one mean is different from the rest. It gets more complicated than that when we consider that there may be interactions, quadratic effects and other model factors. (See Chapter 4 for modeling and Chapter 2, Section 6 for an overview of experimental design considerations.)

We ignore these issues at our peril. For example, in recent years, the pharmaceutical industry has had difficulty reproducing the findings of some of their clinical trials. Could it be that trial designs are too simple to accommodate differences among the way doses interact with various population subgroups? Or could it be that the very probability base we have come to rely upon for much decision making is on shaky ground? We know that the foundations of probability are firm when applied to such things as games of chance with attendant randomization. But are they really firm when applied to large population studies, big data sets, etc.

It is little wonder that the Fisherian School refused to get bogged down in the strict mathematics of hypothesis testing à la Neyman-Pearson.

For more on this topic, see Salsburg (2001) and Weisberg (2014).

5.5.4 More on hypothesis testing

5.5.4.1 Testing variation

Careful testing of differences between means should be accompanied by a comparison of the variation among observations within each of the groups represented by those means. It is not necessary that the means be equivalent to test for differences between variances or their corresponding standard deviations. Fisher devised a variance ratio test, later named for him as the F-test and based on ratios of Chi-squared distributions.

We calculate $F = s_1^2/s_2^2$, with related degrees of freedom for each variance estimate. The test is generally two-sided, so it does not matter which variance is numerator or denominator. Most put the larger in the numerator and then appeal to tables of the F-distribution to determine if the difference is real or could have happened by chance alone.

Revisiting the pH data of Table 5.4, we have $F = \frac{0.256^2}{0.151^2} = 2.891$. [Your findings may differ, depending on how and when you round, but it is always best to preserve as many decimals as practicable and round at the end.] The numerator has 3 degrees of freedom while the

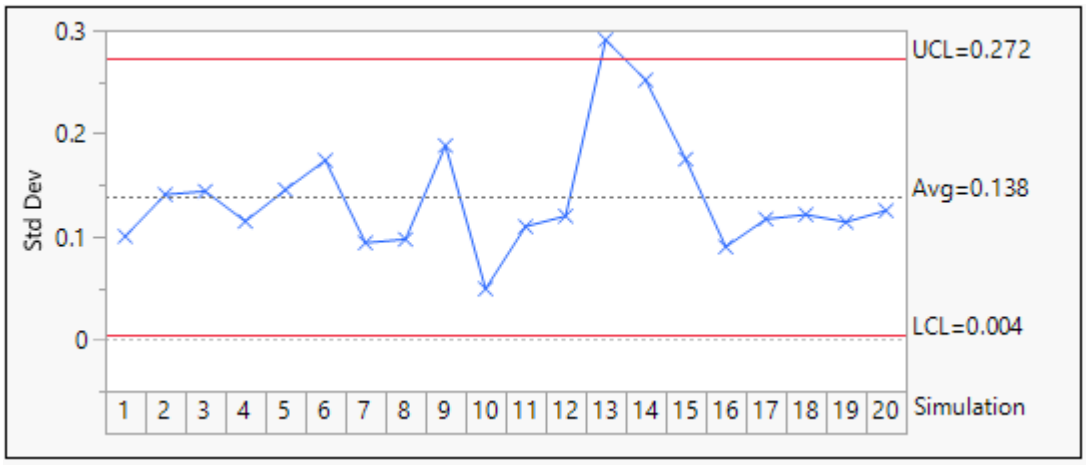
denominator has 5. If we conduct a 2-sided test with $\alpha = 0.05$, we see from the F tables that $F(3,5)_{0.025} = 0.067, and\ F(3,5)_{0.975} = 7.76$. Because the calculated F statistic falls neatly between these two values, we would not reject the hypothesis of equality. We are left to conclude the variances are the same until more data are available.

It is likely that Fisher, himself, would have examined the area under the F-distribution above the calculated value, found it to be 0.28 and moved on to more important conjectures.

Just as there are formal tests for differences among multiple means, there are formal tests for differences among variances. Bartlett's Test (Bartlett, 1937; Mason, et. al., 2003) and Levene's (1960) test are popular among them.

Such formal tests are mandatory if findings are to appear in peer reviewed publications and other formal reports. In the heat of pursuit, a quick visual test to help determine if variance estimates are ill behaved is a simple control chart of the standard deviations.

Figure 5.22 Standard Deviation Chart of Simulated pH Data in Table 5.6



Did something strange happen at observation 13?

For more on control charts, see Chapter 3, Section 3, and for technical details of control chart theory, see Montgomery (2013).

5.5.4.2 Non-parametric hypothesis testing

All data follow some kind of distribution, and all distributions have parameters. Some distributions and their corresponding parameters are unknown. So the term "non-parametric" is misleading, although it has fallen into common parlance. It does not mean that there are no parameters. It simply means we do not know the distribution or its attendant parameters.

"Distribution free" is a more fitting term. It means that the involved testing does not rely on estimated parameters. In those situations, especially where data appear to be ill behaved, distribution free testing can come in handy.

Thinking behind these tests has been around since the early 1900s. Both Karl Pearson and Ronald Fisher, mentioned earlier, engaged in some simple distribution free tests. More elaborate exposition and innovation came from Frank Wilcoxon (1947), a chemist who thought more tests should be available and, finding none, developed his own. The field has expanded and come to a level of maturity over the decades since.

Originally, these tests were intended to aid in decision making with renegade data and to simplify calculations during an age when calculations were often major obstacles. Of course, the latter has been resolved through wide availability of statistical software to do the job.

It would seem that there should be great losses of information via the use of distribution free testing, especially if the underlying data distributions are normal or nearly so. Studies of their asymptotic relative efficiencies have shown surprising positive results. (Pitman, 1948)

We present only a few of the simple tests and refer readers to other sources such as Hollander, et al. (2014).

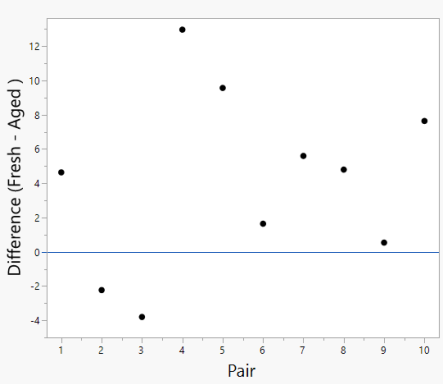
5.5.4.2.1 Paired Replicate Analyses by Signed Ranks

A simple test worked out by Wilcoxon is the distribution free signed rank test. Shown in Table 5.8 and in Figure 5.23 are data generated by repeated measures of zeta potential, an electronic measure related to the strength of an emulsion.

Table 5.8 Zeta Potential Difference with Product Age

Sample	Fresh	Aged	Difference	Absolute Difference	Rank	Indicator	Rank x Indicator
	a	**b**	**c**	**d**	**e**	**f**	**g**
1	30.57	25.93	4.64	4.64	5	1	5
2	37.47	39.70	-2.23	2.23	3	0	0
3	35.97	39.77	-3.80	3.80	4	0	0
4	30.80	17.83	12.97	12.97	10	1	10
5	27.90	18.33	9.57	9.57	9	1	9
6	27.97	26.33	1.64	1.64	2	1	2
7	41.27	35.67	5.60	5.60	7	1	7
8	40.30	35.50	4.80	4.80	6	1	6
9	37.17	36.63	0.54	0.54	1	1	1
10	31.97	24.33	7.64	7.64	8	1	8
						Rank sum:	48

Figure 5.23 Zeta Potential Differences by Sample



Proceed as follows:

1. List the data in columns (a) and (b), and show their differences in column (c).
2. Calculate the absolute value of the differences and put them in column (d).
3. Put the rank, from smallest to largest of the absolute values of the differences in column (e).
4. Form an indicator variable, using a 0 if the value in column (c) is negative and a 1 if it is positive.
5. Multiply the indicator, column (f), times the corresponding rank, column (e), and put it in column (g).
6. The sum of the numbers in column (g) is the rank sum, (T^+).

Critical values or tail probabilities for this statistic are calculated by exhaustive enumeration of the possible combinations and permutations of ranks and indicators for each number of observations – all carried out by hand or rotary calculators during the 1950s.

Resulting tables of upper tail probabilities for the null distribution are shown in references on distribution free statistics. In the case of this example, the tail probability corresponding to 10 observations and a rank sum of 48 is 0.019. We would conclude that age reduces zeta potential.

When the number of samples is very large and the null hypothesis is true,

$$T^* = \frac{T^+ - [n(n+1)/4]}{[n(n+1)(2n+1)/24]^{1/2}}$$

where n is the number of pairs, is nearly normally distributed with mean 0 and standard deviation 1. In those situations, T^* can be used as a z-value in normal distribution statistics.

5.5.4.2.2 Distribution free one-way layout testing

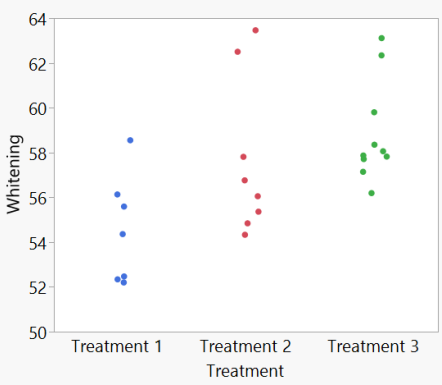
A next level of sophistication is the one-way classification. Data representing whitening effects of three detergent treatments are shown in Table 5.9 for purposes of illustration.

Table 5.9 A One-Way Classification Study – Whitening Effects of Detergents

	Treatment 1	Rank 1	Treatment 2	Rank 2	Treatment 3	Rank 3
	a	**b**	**a**	**b**	**a**	**b**
	52.47	3	56.05	9	62.34	22
	56.13	10	54.33	4	57.82	16
	52.20	1	56.76	12	56.19	11
	52.34	2	54.84	6	59.80	21
	54.36	5	62.50	23	63.11	24
	58.55	20	55.36	7	57.71	14
	55.59	8	57.81	15	58.06	18
			63.46	25	57.87	17
					57.14	13
					58.35	19
c	Number	7		8		10
d	Sum	49		101		175
e	Average	7.000		12.625		18.500

Observations are shown graphically in Figure 5.24

Figure 5.24 One-Way Classification Data from Table 5.9



A distribution free evaluation for treatment differences is the Kruskal-Wallis (1952) test.

Its steps are as follows:

1. List the data (a) and corresponding ranks (b) of each observation with respect to the entire data set.
2. Count the number of observations (c) in each treatment category.
3. Sum the ranks in each category (d).
4. Calculate the average rank for each category (e).

Calculate H:

$$H = \left(\frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right) - 3(N+1),$$

where N is the total number of observations in the table, and n_j is the number of observations in each of the j treatments.

We get $H = \frac{12}{(25)(26)}\left[\frac{49^2}{7} + \frac{101^2}{8} + \frac{175^2}{10}\right] - 3(26) = 8.411$

For relatively low sample sizes, critical values of H are tabled in various texts on distribution free statistics. For large sample sizes H is distributed as a Chi-squared statistic with degrees of freedom equal to the number of treatments, k, minus 1. In this case, we find that the area above the calculated H-value is 0.0149, so we would declare that the whitening power differs among treatments.

As might be expected, if the H test shows significance at some reasonably low probability level, it would be desirous to learn where differences lie. A multiple comparison test is given by Dunn (1964) as

$$\left|\bar{R}_i - \bar{R}_j\right| > \frac{q_\infty, k, \alpha}{\sqrt{s}}\sqrt{\frac{N(N-1)}{12}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

where q is defined as Tukey's (1949) studentized range.

5.5.4.2.3 Distribution free two-way layout testing

Building on the preceding techniques, we now consider an additional level of complexity in the form of a randomized block experiment. This is similar to the one-way classification except that observations are set out in separate blocks to reduce the cloud of variation and allow treatment effects to become more visible.

This example concerns the effect of four different treatments on protein levels in livers of laboratory rats. Note that instead of assigning the animals at random to treatments, litter mates were chosen in order to minimize what would otherwise been replicate-within-treatment variation. Data are listed in Table 5.10 and shown graphically in Figure 5.25.

Table 5.10 Randomized Block – Lab Rat Organic Protein Levels Resulting from Different Treatments

Litter	A	Rank A	B	Rank B	C	Rank C	D	Rank D
1	56	3	64	4	45	2	42	1
2	55	3	61	4	46	2	39	1
3	62	4	50	3	46	2	45	1
4	59	4	55	3	39	1	43	2
5	60	4	56	3	43	2	41	1
Rank Sum		18		17		9		6
Rank Avg.		3.6		3.4		1.8		1.2

S = 12.6 Chi-squared = 0.00559

Data source: H.P. Andrews (1967)

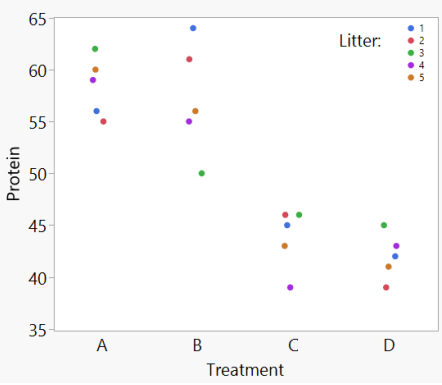
The distribution free test is commonly called the Friedman test. Its steps are as follows:

1. Rank the responses, lowest to highest within each block.
2. Sum the ranks for each treatment.
3. Calculate the average ranks for each treatment.
4. Calculate

$$S = \left[\frac{12}{nk(k+1)}\sum_{j}^{k} R_i^2\right] - 3n(k+1)$$

As with the Kruskal-Wallace test, for small sample sizes, critical values of S may be found in texts on distribution free statistics. For large sample sizes, S is distributed as Chi-squared with (k-1) degrees of freedom.

Figure 5.25 Lab Rat Organic Protein Levels Resulting from Different Treatments



Section 5.5 References

Andrews, H. P. Design of Experiments course notes, (1967).

Bartlett, M. S. "Properties of sufficiency and statistical tests," *Proceedings of the Royal Statistical Society*, Series A 160, (1937), 268–282.

Dunn, O.J. Multiple comparisons using rank sums, Technometrics, 6, (1964) 241-252.

Hare, L.B. The Foundation of Statistical Engineering, Quality Progress, August, 2019.

Hollander, M., Wolfe, D. A., and Chicken, E. Nonparametric Statistical Methods, 3rd. Ed, John Wiley and Sons, Hoboken, NJ, 2014.

Kruskal, W.H. and Wallis, W. A. Use of ranks in one-criterion variance analysis, J. Amer. Statist. ASS. 47, (1952), 583-621.

Levene, Howard, "Robust tests for equality of variances". In Ingram Olkin; Harold Hotelling; et al. (eds.). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, (1960), 278–292.

Mason, Robert L., Gunst, Richard F., Hess, James L. Statistical Design and Analysis of Experiments, with Applications to Engineering and Science 2nd Edition, John Wiley and Sons, Hoboken, NJ, 2003.

Montgomery, D. C. Introduction to Statistical Quality Control, 7th Edition, John Wiley and Sons, Hoboken, NJ, 2013.

Neyman, J. and Pearson, E. S. On the Problem of the most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society A*, (1933) 231.

Pitman, E.J.G. Notes on non-parametric statistical inference, Columbia University (1948).

Salsburg, D. The Lady Tasting Tea, W. H. Freeman and Co., New York, 2001.

Tukey, J. Comparing individual means in the Analysis of Variance, Biometrics, 5 (2), (1949), 99-114

Weisberg, H.I. Willful Ignorance, John Wiley and Sons, Hoboken, NJ, 2014.

Wilcoxon, F. Some Rapid Approximate Statistical Procedures, American Cyanamid Co., Stamford Research Laboratories, Stamford, CT, 1947.

Chapter 6 – Solution Identification and Deployment

Table of Contents

Section 6.1 - Theory of solution identification and deployment

6.1.1 Objectives

After the statistical engineer (SE) has learned about the problem it is time to search and select the most prevalent influences and to generate and ultimately select the most adequate and elegant solutions. The process of solution identification and deployment is an iterative process whereby creativity and an entrepreneurial attitude are pivotal. This process is characterized by phases of divergence (generating as many possible influence factors that are needed for a solution) and convergence (converging to the most prevalent influence factors which solve the problem adequately). To successfully arrive at the best solutions collaborations are crucial. Consultation of experts in the problem area, hints and clues about possible influence factors and solutions from subject matter experts, and early understanding of the managerial appetite for anticipated investments are crucial for successfully navigating through the phase of solution identification and deployment.

Selection of the best solution requires understanding what is fundamentally causing the problem. Therefore, this phase begins with the identification of possible factors causing the problem and defines a process for identifying the most prevalent factors that can be controlled, compensated or eliminated. Finally, in this phase the SE starts to anticipate what is needed for the deployment of possible solutions. Once possible solutions are revealed, initial considerations about what is needed for deployment starts. Thereby, the process of preparing the organization and key leaders for deployment is commenced. Possible hurdles can be identified and timely action to prevent or mitigate deployment obstacles can be taken.

6.1.2 Divergence: Finding possible influence factors

The task of identifying many possible influences is challenging. Several techniques can be of value for the SE to assure a complete and exhaustive consideration of possible influence factors (also see De Mast et al., 2012).

Consultation of process know-how

Ask people in the direct vicinity of the problem to contribute their experience and expertise. This is typically done through direct meetings and consultation, but it is just as easy to call or email the persons the SE deems knowledgeable about the problem. The basic idea is to pool knowledge: many people see different aspects of the process or problem under study; integrating and combining all the perspectives will start the shaping of possible influence factors. The sources that this approach exploits are specific (technical) knowledge and tacit know-how. To help contributors structure their theories about possible influence factors, a cause and effect diagram is a convenient tool. To construct it, the SE keeps asking questions (“5 times why”) insisting that contributors make explicit what they insinuate. A useful taxonomy is given by the Ishikawa diagram with the perspectives of man, machine, material, method, measurement, and Mother Nature (6Ms) in manufacturing environments. In service environments, think of causes related to employee, computer, information, working method, customer, and external factors.

Technical literature and experts

Besides relying on local process know-how, the SE can consult the literature on the subject or an expert to identify potential influence factors that cause the problem.

Exploratory data analysis

To identify influence factors, the SE can search the data for patterns and other salient features. Nonrandom patterns in data are symptoms of disturbances and nuisance variables. In trying to explain these patterns, the SE identifies potential influence factors. Since the SE does not know in advance what he is looking for, graphical techniques are especially powerful tools for exploratory data analysis, because they have the power to reveal the unexpected. Control charts are often used for this purpose, as are scatter plots and boxplots.

Lessons from historical cases

It is usually very insightful to make a close examination of a limited number of past failure investigations. In these retrospective studies, the SE tries to reconstruct what went wrong there, and thus gains insight in the causes and nature of problems. In a similar manner, the SE may want to investigate past instances where a change or new product commercialization went exceptionally well. During such a study, the SE selects several cases that went well (low lead time or processing time, or very good quality) and several cases that went poorly. Next, the SE closely compares the successful and unsuccessful cases, writing down all the differences that she sees. This investigation and documentation process continues until a pattern emerges in the differences between the success and failures.

Lessons from analogous situations

The SE could try to figure out how others have solved similar problems at other sites or in other companies. It is important to realize that most problems have been solved by others in a different context. By enquiring into what worked and what did not for a similar problem, the SE can obtain valuable insights for solving the problem at hand.

Eliminate and zoom-in

Advanced strategies for problem solving are almost always variants of the principle: eliminate and zoom-in. The SE observes how the problem behaves. Statistical Engineering applications of eliminate and zoom-in include localization of the problem in place and time. Observing where in the process the problem first manifests itself, the SE knows that the cause of the problem must be located prior to this point. From these observations she can eliminate whole directions as impossible, zooming-in on directions that appear to be more likely. Alternatively, the SE determines the dominant sources of variation.

Failure modes and effects analysis (FMEA)

Also, disturbances – the more or less frequent events that derail the regular process – could be identified following the approaches explained above. An approach designed especially for the identification and prioritization of disturbances is the failure mode and effect analysis (Stamatis, 1995). This approach looks like a brainstorming session as described above (consultation of process know-how): the SE invites a group of people who are involved in the process to identify potential disturbances. A FMEA goes further, in that the cause and the effect of each disturbance are determined. Rating the frequency of the disturbance (on a scale of 1: rare to 10: frequent) and the impact of its effect onto the problem (also on a scale of 1 to 10), the SE determines its priority as frequency times impact. Sometimes a third dimension (detectability) is included to determine the priority.

6.1.3 Convergence: Selecting important influencing factors

An exploratory mindset is helpful when working in the divergence phase to identify influencing factors. In this phase the SE converges to the most important influencing factors and solutions. This requires a different way of working, being methodical, rigorous and objective. To study the effects of all the listed influencing factors, several types of statistical investigation can be applied.

Design of experiments

In the design of experiments (see Chapter 3, Section 5), the factors of interest are studied according to a well-defined plan, while other factors are kept as constant as possible. This is called an experimental design. The simplest form of an experiment is one factor at the time, in which the experimenter studies a factor at only two levels. The experimental design is very simple in this situation. A number of randomly chosen items is processed using one level of the factor, while other items are processed using the alternative level. Each time, the response is measured. By comparing the results of one group of measurements to the other group, the SE can study the effect of the factor on the response.

Factors often have more than two levels (the current application and two or three alternatives). Mostly, the experimenter wants to study more factors simultaneously. It is not efficient or more reliable to study these factors one at a time. Experimental designs in which all factors are manipulated simultaneously give better results (although the analysis is more involved). The statistical theory of design of experiments (usually abbreviated to DoE) describes optimal schemes for such experiments. These comprehensive theories are treated in this book. The reader is also referred to Box, Hunter and Hunter (2005) and Montgomery (2012) for a clear overview.

Statistical modelling, hypothesis testing and goodness of fit

Analysis procedures to establish the effect of influencing factors on the problem to be solved consist of three steps: model the effect of the factor, test whether this effect is significant, and study whether the model fits the data well. Most software packages offer integrated procedures, which execute these steps together and provide a graphical display in addition. The first step is to give a mathematical description of the effect(s) of the factor on the response(s). The second step is to test the hypotheses. The last step is to verify that the model fits the data well (statisticians speak of the goodness of fit). To study goodness of fit we examine the random scatter of the so-called residuals (observations minus the fits of the model). A good fit implies that this random scatter is patternless noise.

The statistical sciences offer an overwhelming number of techniques for modelling, hypothesis testing and goodness of fit. One should bear in mind, though, that all these techniques perform the three tasks listed above — the differences are in the mathematical details, not in their function. The appropriate technique depends on the type of data (categorical or numerical) and assumptions about statistical properties of the data (e.g., independence and normal distribution or other).

Establish relationships

Having figured out which factors have the highest impact on the response, the SE needs to know how large each effect is. This follows directly from the data analyses done in the previous step. For categorical factors, a list of means for each level of the factor quantifies the effect of the factor. For numerical factors, a regression equation represents the relationship between the influencing factor and the problem.

6.1.4 Preparing for solution deployment

By now the key influencing factors requiring solutions become clear and the SE starts with making the necessary preparations for deployment. Pivotal in this phase is to:

- Identify the persons or stakeholders involved in the deployment of the feasible solutions
- Estimate the timelines needed for deploying the most feasible solutions
- Specify the impact of the feasible solutions on the problem.

These activities are all part of managing the deployment process. In this phase early involvement of key stakeholders is required. The anticipated impact and effort of solution deployment is jointly determined. Thereby, ownership for solution deployment is created in the organization. This owner is fundamental for successful project discharge after solution deployment. Anticipating solution deployment comprises three main dimensions that need consideration and preparation:

- **Technical dimension:** In this dimension the SE and key stakeholders specify what is needed to deploy the solution. Typical deliverables for this dimension are a business case, a detailed design of the solution and the development of success criteria.
- **Organizational dimension:** Here the SE and the key stakeholders detail what is needed from the organization to make the solution work. Deliverables for this phase include roles and responsibility specifications, a deployment roadmap, the assurance of needed resources (such as on-site support) and a hand-over to business as usual plan.
- **Political dimension:** Apart from the solution and the deployment activities needed, the SE should be aware of other interests and needs from key influencers in the organization. Logic does not always triumph, and for that political considerations are needed. Typical deliverables for this dimension are stakeholder analysis, a guiding coalition of influential proponents, definition of a sense of urgency for improvement and a vision about how the solutions deployed are making a difference. Most important is agreement on how political issues that will arise in the solution deployment process are to be managed.

6.1.5 Conclusion

In the solution identification and deployment phase the following steps must be completed:

- **Determine the most important influencing factors:** Statistical methods such as hypothesis testing and design of experiments have been used for screening many factors to identify the important influencing factors (control variables and nuisance variables). Disturbances with a high risk have been identified via an FMEA.
- **Determine relations between the characteristic of interest and influencing factors:** For the most important influencing factors, a statistical model had been built (describing the mathematical relation between the characteristic of interest and the factors).
- **Anticipate solutions deployment:** For the most plausible solutions, preparatory activities are performed and deliverables in three dimensions are considered.

Section 6.2 - Holistic solution deployment

6.2.1 Objectives

Solving a problem by advanced statistical inference requires more than a mere "technical solution and deployment." To solve a problem and make a solution work, a broader perspective needs to be adopted. This makes the role of a SE challenging. One must be an expert in advanced statistical engineering techniques and must also be an effective organizational leader (Senge, 2014). This later role requires that the SE also has organizational awareness and excellent influence and negotiation skills. Part of this role as organizational leader is creating commitment and educating the workers in the near vicinity of the problem and its solutions. Beyond bringing a one-time solution, an effective SE brings a solution and improves the system that gave rise to the problem in the first place. Thereby future problem manifestations are prevented and better process control, early issue resolution, and improvement capabilities are developed. Improvement of the system directly related to the problem and its solution is discussed in the sections "Adjust the Quality Assurance System" and "Statistical Process Control."

Another important part of this role is finding the right balance between solutions that have impact and are simultaneously feasible. Impact is determined by the scale of improvement a solution can bring. Feasibility is about the likelihood or ease of successful implementation of the solution. The key question a SE must ask himself in this phase is about the effort and investment that are needed in relation to the impact of the solution that is to be deployed. This section discusses two prevalent topics related to these two concepts.

6.2.2 Impact: Applying a Systems Thinking approach

Apart from understanding the influencing factors that are directly related to the problem, the SE should understand the broader systems in which problems emerged to truly value the impact of the possible solutions. Systems thinking is about understanding the root causes for the problem the SE tries to solve. Often these root causes are a consequence of the fundamental design of the organization. These are root causes that are not directly related to the problem at hand but do have a fundamental connection to the emergence of the problem. While understanding these fundamental and systemic factors that cause unwanted consequences and manifest themselves by problems, what the SE is trying to solve is pivotal for sustainable solutions. This makes systems thinking especially relevant, as it is concerned with finding the organizational root causes underlying the problem. The SE should be aware that solving such fundamental systemic issues is ultimately a task for the executive management team.

Example
Image your washing machine is broken, and an engineer arrives to fix it. The way the engineer behaves in your home depends on the system she is working in. Imagine the engineer does not have the part to fix the machine, two things can happen:

1. She says she cannot fix your machine and tells you to phone the helpdesk to tell them you need another appointment. She fills in a form to tell someone what part is needed.
2. The engineer calls logistics to see when the part is available and then tells you when she can come back to fix your machine, checking if the date suits you.

The problem is the same. However, outcome 2 is clearly to be preferred. What makes the difference? The system, not the engineer! In outcome 2, the engineer could make a commitment because the system supported her.

Understanding the broader system

Sustainably solving a problem requires a broader understanding of the system in which the problem manifests itself. A good start, after the problem root causes are identified, is to try to identify what systemic factors might have been fundamental to this root cause. Here the SE asks himself, what management behavioral routines or management thinking might have caused this problem in the first place (Figure 6.1). Identification of such causing factors is a first step in ensuring sustainable solutions. This can be a complex exercise; for those interested further reading on the topic is advised (see Senge, 2014).

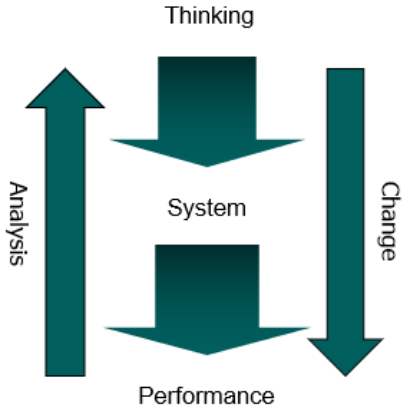


Figure 6.1 Conceptual model for Systems Thinking (Seddon, 2019)

A good start to understanding the system in which the problem could manifest itself is to assess on several dimensions how the organizational system can be classified. A method to do so is provided by the framework in Figure 6.2 (Seddon, 2019). The framework separates ten dimensions in which organizations can be assessed, to provide the first clues about what systemic conditions and underlying system thinking are present. Basically, the framework separates a “command and control” organization on the left from a Systems Thinking organization on the right.

Example

After having established the root cause of the software malfuntioning, the SE digs a little deeper and asks himself a few more times “why” this problem has occurred. hHe finds that contractual agreements caused the problem: the developers were forced to document the code manual for the software in a mandatory format. That format left little room for elaboration of certain design features. As a result, the software deployment engineers were unable to fully understand the design. As the deadline for the project was tight, and management would not accept delays, the engineers decided to work with the incomplete documentation, ultimately leading to the problem.

Figure 6.2 Framework for analyzing the system (Seddon, 2019)

6.2.3 Feasibility: Selecting feasible solutions

Determining the feasibility of a solution begins with a straightforward analysis of the means and resources needed for a solution to be deployed, as discussed in previous sections. Apart from applying this primary level of analysis, an effective SE needs to be aware of common hurdles and hindrances in the solution deployment processes. Based on these insights the SE can determine how this affects the feasibility of the proposed solutions. Key questions the SE asks himself are about the most common reasons for solution deployment failure, how these reasons might impact the proposed solutions, and how they affect the feasibility of specific solutions.

Example In an industrial process it is found that poor product quality can be traced back to a specific production line responsible for making components. One solution is to improve the knowledge and skills of the line workers, whereas another solution is about further automation efforts. Both solutions require similar amounts of investment in terms of time and budget. To assess the feasibility, that is the chance for successfully deploying the solution and have an impact on solving the problem, the SE reflects on common reasons for solution deployment failure. She finds that technical complexity of solutions is a known common reason for failure. Based on discussions with several stakeholders she estimates that the organization is indeed too unfamiliar with the automation software available and she decides that, for now, developing the knowledge and skills of the workers is a more feasible solution.

Common causes for solution deployment failure

By understanding the internal and external factors that affect solution deployment failure, the SE can mitigate the consequences or even prevent such failure from happening. One might suggest that failure to deploy solutions should never occur, but often solutions that have the most impact also have a higher risk of failing. By understanding the factors that influence failure the SE can reduce the likelihood of failure or, after solution deployment initiation, make early termination decisions to mitigate the downside of the failure. The categories that are discussed are the result of a systematic literature review of research specifically focused on identifying the most prevalent solution deployment failure factors. Figure 6.3 reveals the most widely acknowledged and recognized causes for failure (Lameijer et al., 2019a). Each of the failure categories mentioned in Lameijer et al. (2019a) are discussed below:

No.	Solution deployment failure categories
Managerial	
1	Ambiguous rationale, scope and objectives
2	Incomplete requirement analysis and delivery
3	Incorrect project management methodology application
4	Insufficient change management
5	Insufficient stakeholder management
6	Unavailability of project team and lacking skills
7	Unclear project team roles, responsibilities and relationships
8	Insufficient sponsorship and commitment
Technological	
9	Insufficient experience and technological novelty
10	Unforeseen problems and technological complexity
11	Incompatible existing technological infrastructure
Organizational	
12	Obstructive organizational culture
13	Complexity and stability of the organizational structure
External	
14	Changing regulatory requirements
15	Effects of public justification
16	Effects of alliances and collaborations

Figure 6.3 Categorized solution deployment failure factors (Lameijer et al. 2019a)

Project managerial failure factors

Several failure factors that originate from the way the SE project is being managed before, during and when nearing completion can lead to failure.

Rationale, scope and objectives: Ambiguous rationale, scope and intended objectives of SE projects can result in reduced commitment from the project manager, reduced management attention, inappropriate allocation of technical and organizational resources and budget over-runs. Insufficient scope definition allows for "scope creep" and increases the likelihood for conflicts and failure due to differing views. In an uncertain business environment, changes in scope can occur. Scope management techniques should be incorporated at the project planning stage and any necessary changes should occur through the formal control procedures within the predetermined time and cost.

Requirement analysis and delivery: In early project phases a sound requirement analysis and involvement of key stakeholder(s) is important to ensure clarity on functional

performance and reliability requirements. Proposed mechanisms comprise product breakdown structures and Agile development structures. Clarity on the project requirements minimizes the risk of confusion, conflict, and delays.
Project management methodology: Ineffective project management methodologies and incompetently applying project management methodologies are recognized as important failure factors. More specifically, primary failure factors include incorrect project planning, inadequate risk analysis and management, incomplete dependency management, insufficient progress monitoring and project control, unclear process instructions, insufficient quality assurance and inadequate internal project member communication.

Change management: Where project management is process related, change management is people related and consists of clear communication and coordination of roles and responsibilities especially for larger projects, managing partnerships, and maintaining clear linkage to corporate strategy.

Stakeholder management: Improper management of stakeholders, such as subject matter professionals, end-users and senior management, can result in conflicting interests and expectations. Stakeholder attitudes, expectations, interplay and influence must be managed, monitored and assessed periodically, taking the cultural background of stakeholders into account. Stakeholders deliver the appropriate resources. Known reasons for weak participation are engagements in existing operational activities or geographical distance from the project location.

Project team availability and skills: Unavailability of a knowledgeable project team, unwillingness to share knowledge, an insufficiently knowledgeable project manager, unavailable specialized subject matter experts, or departure of critical members are known to cause project failure.

Project team roles, responsibilities and relationships: The need for active participation from project team members is acknowledged. The quality of relationship, cohesive behavior, effective conflict management, as well as the distance of the physical working locations from project members are important reasons for project failure. Team member self-reflection, shared intention and corresponding egalitarian forms of responsibility, progress monitoring and conflict resolution mechanisms are important for project teams to function well. The project member's job satisfaction is not directly associated with adhering to deadlines or cost objectives. Moreover, employing a cross-functional team will mitigate project failure as teamwork encourages people with varied skill sets to work together as opposed to working in isolation.

The leading role and people skills of the project manager are recognized, though their impact on projects failure remains inconclusive. Recent results indicate that more participative decision structures have a positive effect on project failure prevention compared to hierarchical decision structures. Project leaders must hold a facilitator position in the organization in order to ensure management commitment and appropriate allocation of resources.

Sponsorship and commitment: Employees tend to focus on activities that their management deem important. Senior managers have an important role in safeguarding: (1) projects from excessive business pressure and loss of autonomy, (2) realization of the business changes resulting from the project, (3) ensuring alignment with corporate strategy, and (4) providing

the necessary resources and authority to the project. Prior research indicates that the main role of senior managers is to lead and monitor projects, provide the resources for their implementation and establish work policies for the improvement teams. At the same time, management must also carry out a process for integrating the different departments, enabling everyone to share common objectives.

The strength of sponsorship is determined by the importance of the project to the strategic objectives of the organization and the stability of the senior management positions. It is important to identify the right sponsor from the beginning and secure active participation throughout the life of the project.

Technological failure factors

Project failures due to complexities that are rooted in technology fall into three categories:

Technological novelty: New technology is known to create risks and has caused many project terminations. Prior experience with the technology decreases the chance for failure. Therefore, the decision to apply new technology must be taken carefully. If existing alternative technologies exist, it is recommended to first explore how existing technology (for the firm or for the industry) can solve the problem.

Technological complexity: Unforeseen problems due to complexities, caused internally or externally, can surface in the design or when building the deliverable. When not corrected, they may cause project failure.

Technological compatibility: Lack of compatibility of project deliverables with existing IT infrastructure or data models and software are known to cause project failure.

Organizational failure factors

Organizational culture and structure are recognized failure factors.

Organizational culture: Culture can be supportive or obstructive towards the intended project outcomes. Failed projects are related to organizational cultures that are characterized by an internal focus on resistance to change. Prior research suggests that some form of reward and recognition system is necessary for employees to be motivated and engaged in the execution of SE projects. The incentive or reward system fosters a sense of achievement and company recognition, thus generating greater employee motivation and commitment in future SE projects, producing an upward spiral effect.

Organizational structure: Complexity caused by differing organizational units, the change resistance caused by the organizational structure, and corporate headquarters design which limits local innovation create risks for project failure.

External failure factors

The final category of project failure factors that are recognized in the literature originates from outside the project and the organization.

Regulatory requirements: Changes, absence or incomplete legal frameworks and standards can lead to ambiguity and conflict that contributes to project failure.

Public justification: External justification by stakeholders influences the commitment of actors involved in the project. Even for outcomes that have low expectations, public justification significantly influences the willingness to invest.

Alliances or collaborations: Collaborations with parties outside the organization affects project outcome due to potential limited availability of essential knowledge, especially for SMEs, potential difficulties in partnerships or cooperation, or conflicts in the supplier-buyer relationship.

Mitigation strategies for project failure

Knowledge of what project failure factors are likely to impact solution deployment raises the need for mitigation strategies. A systematic review of the literature has identified mechanisms that allow for before, during and after the project failure mitigation (Lameijer et al. 2019a).

Before-the-project failure mitigation strategies: Known preventive strategies for project failure are skill gap identification and training programs for project sponsors, managers, members and stakeholders to ensure technical and intercultural competency. For preventing project failure when engaged with external suppliers, it is advisable to use contracts wherein suppliers agree on the costs and penalty (calculated based on the probability of failure) when they fail to deliver as promised. Clear and shared understanding of the project scope must be in place through transparent and effective communication in the early stages to reduce the chances for project failure.

During-the-project failure mitigation strategies: While the project is in execution mode, close monitoring of progress allows for learning at regular intervals and following significant events. To do so, feedback loops on sub-tasks that quickly deliver sub-deliverables of the bigger project's end result are advised. In a typical project, a reporting system is designed to meet the needs of the organization. Strategies related to user involvement and project planning and communication are most influential in preventing failure.

When project failure is imminent and commitment to project success is failing, actions to turn troubled projects around are redefinition of the project and its objectives, improvement of the project management methods applied, and a change in project leadership. For project leadership, it was found that when project managers believe the failing project is under their control it is unlikely they recommend alternative courses of action other than continuation (Jani, 2008).

After-the-project failure mitigation strategies: The elements of learning and the execution of retrospectives are used often as an after-the-project failure mitigation strategy. It includes elements as cognitive and causal mapping and decomposition of a project in a complex set of linear and non-linear interactions to identify interactions and dependencies. Learning at the individual, team and organizational levels is essential for the sustainable deployment of solutions.

Prior research suggests that the ability to learn from failed projects is negatively influenced by the intensity of the emotional reactions. While delayed project termination does provide the time needed for learning from failure, a negative side effect is that negative emotions have more time to grow. Finally, be advised that what works well in one situation may not

work in another, and therefore engagement in after-the-project learning should be focus on generic and specific lessons learned.

6.2.4 Conclusion

In the holistic solution deployment phase, the following steps must be completed:

- **Applying a systems thinking approach:** Understanding the broader systems in which problems emerged to truly value the impact of the possible solutions.
- **Selecting feasible solutions:** Awareness of common hurdles and hindrances in solution deployment processes. Based on these insights the SE can determine how this affects the selection of alternative solutions.

Section 6.3 - Incorporating Human Factors

6.3.1 Objectives

Sustainable solutions are the results of thorough analyses, adequate and realistic solution deployment planning, and the commitment of all team members who remain related to the problem and its solutions. Ensuring commitment and involving those key players in solution development and deployment is an important task for the SE. The SE needs to demonstrate change leadership, a behavioral routine especially focused on ensuring participation and commitment of everyone in the immediate vicinity of the problem or its solutions.

6.3.2 Demonstrating change leadership

Identification of significant influences and proven solutions is the basis for effective improvement. In many situations, part of the improvement is about deploying the solution in the operational process and instructing or training the workers in that process on how to work with the solution. In this situation the concept of commitment and action is important. The SE is not only responsible for the technical solution, but also carries the responsibility to implement the solution. For that, the SE needs to manage the direct stakeholders and ensure a state of action. A useful conceptual model to understand commitment and attitude of stakeholders is the ACCA framework (Colley, 1961), see Figure 6.4.

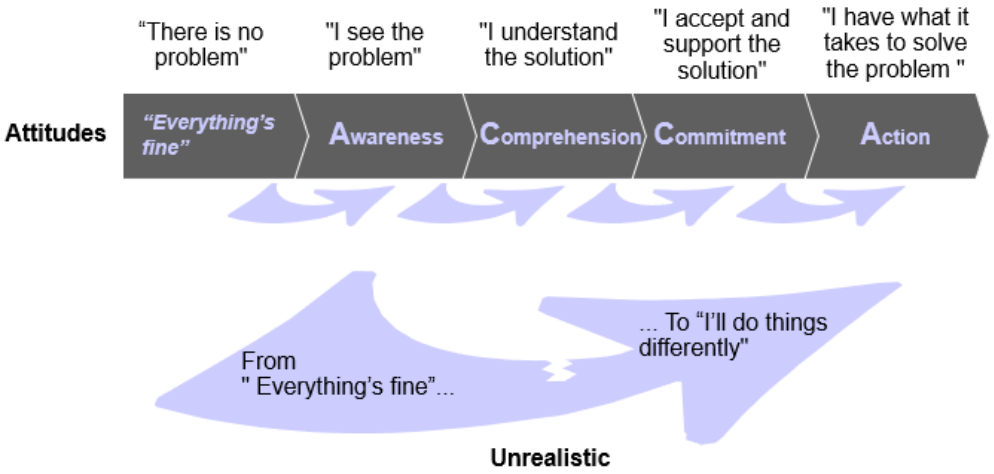


Figure 6.4 ACCA change leadership framework

Bringing about change means understanding where the stakeholders are in terms of their change process. People who believe there is no problem have a hard time accepting a solution, and rightfully so! It is the SEs task to estimate where his stakeholders are in terms of understanding and commitment and design appropriate actions accordingly. It is hard to accept and commit to change for someone when one does not even see that there is a problem.

Example
A higher education institute had problems related to late publication of student grades. After rigorous analysis, the SE found that these problems were largely due to configurational

settings in the administrative software. To solve the problem, several technical modifications were proposed. For this, extensive resources from the IT support team were needed. After consultation, the SE found out that these resources were not available for at least six more months. After further discussions with the head of the IT support team, the SE discovered that his request was prioritized as being very low. It appeared the head of IT support was not aware of any problems with the publication of student grades, let alone with the software supporting this process. By better understanding the problem and its implication the head of IT support saw that indeed a solution was needed quickly and was able to reprioritize. Additionally, he decided to install a monitoring mechanism to make sure the problem stayed under control.

6.3.3 Meeting behavioral change needs

Changing behavior is often an important part in bringing effective solutions. Effective change of human behavior has several prerequisites, for which the influence model provides useful guidance. This influence model was developed and popularized by McKinsey (McKinsey, 2015), see Figure 6.5.

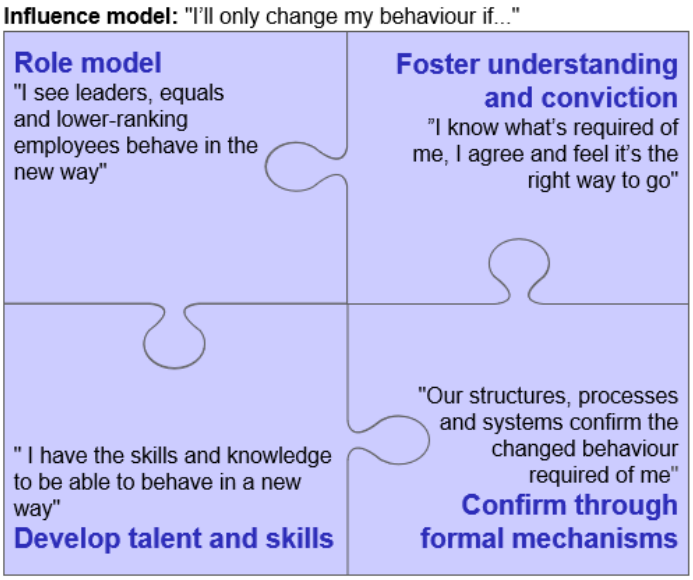


Figure 6.5 Influence model for designing effective change

Role modeling

For any solution to work it is essential that the directly and indirectly involved leaders demonstrate exemplary behavior. For the SE this means that for any solution she must determine what the desired behavior is that is needed for the solution to work. Asking leaders or influential employees to role model and showcase the desired behavior will likely accelerate employee adoption.

Foster understanding and conviction

Understanding the logic and reasons behind a solution are crucial for employees to accept and adopt the solution. For the SE this means attention must be paid to explaining and creating a compelling story as to why the solution is right and necessary. This also includes listening to what is not yet clear about the solution and if needed, include alternative views in the solution design if this fosters conviction.

Developing talent and skills

Having the skills needed to work with the solution is an important part of getting solutions to work. The SE is responsible for facilitating the learning needed for the designed solution. Thereby employees are better able to understand what the solution means for their situation and how they can contribute to making a solution work.

Reinforcing change through formal mechanisms

Formalizing the behaviour that is needed for the solution to work is the final important aspect for effective solution deployment. Formal mechanisms are structures, systems and processes that support or influence employee behaviour. For the SE this means she must consider formal organizational mechanism changes to support solution adoption, such as performance goals or financial and nonfinancial incentives.

6.3.4 Understanding the needs for behavioral changes

The SE needs to be aware that change comes gradually and must bear in mind that people need time to go through the change process that comes with solution deployment. When bringing changes, there can be differences among what people say they would like to see happen (the solution and corresponding behavior), what people think this means (the solution), what actions they perceive as open to them given their skills and knowledge, what people do at their desks, and the final outcome of individual and group actions.

Hence, often we know where we are right now and what we want to achieve in terms of solution deployment. Therefore, we have to identify the changes in our operating practices that are needed to help achieve that vision. To do so we need to ‘uncover’ mindset changes to cause lasting change in operating practices and sustainably solve problems. In that process there is a fundamental difference between what is visible in terms of behavior and what is invisible in terms of thinking and feeling, values and convictions and needs that are or are not met. For a SE to bring effective change, she needs to be aware of these fundamental concepts and closely monitor in the solution deployment process where possible causes for delayed impact might be rooted. This can be a complex exercise. For those interested further reading on the topic is advised (see Goodman, 2002).

6.3.5 Conclusion

For solutions to work, it is important that the human aspects associated to the changes are considered. Several techniques are presented that can clarify what human aspects need to be addressed and how this can be done.

Section 6.4 - Standard improvement directions for piloting solutions

6.4.1 Objectives

When the statistical engineer (SE) knows the important influencing factors and understands their relationships with the characteristic of interest, he can undertake various actions to improve the behavior of the characteristic of interest. The next sections discuss the possible directions for improvement.

6.4.2 Increase or decrease the mean value

To bring the mean value of the characteristic of interest to a target value, or to maximize or minimize a characteristic of interest, the SE can utilize the effects of the control variables. The statistical model enables him to choose combinations of factor settings that optimize the characteristic of interest.

Example
In a coffee decaffeination process, the relation between caffeine percentage and extraction time is negative (see Figure 6.6), i.e., the longer the extraction time the lower the caffeine percentages. This is confirmed by the correlation coefficient of -0.927. Extraction machine-to-machine variation appears to be an influencing factor of caffeine percentage. Furthermore, all batches processed on Extractor Machine 3 do not comply with the upper spec limit. On the other hand, batches processed on Extractor Machine 1 are well below the limit. Hence, we may use the control variable extraction time to compensate the caffeine percentages in Extractor Machine 1 (lower duration can be used) and Extractor Machine 3 (higher duration is needed).

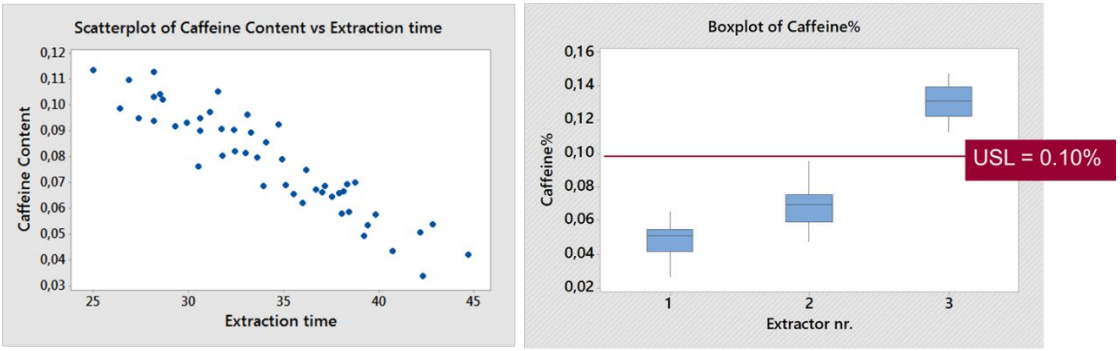


Figure 6.6 Relation of caffeine percentage and extraction time and its relationship with the categorical variable 'extractor'

6.4.3 Feedforward control

Sometimes noise variables that influence a characteristic of interest are not known, or they cannot be controlled (e.g., variation in raw material, shift-to-shift and machine-to-machine variations). It may be possible to identify a controllable factor that is related to the noise variable, and the SE may be able to develop a model relating the two. In this case, it may be

possible to neutralize the effect of noise variables on the characteristic of interest with a *feedforward* control system. The anticipated effect of the specific "level" of the noise factor is compensated by changing the controllable factor.

Example
The three different machines of Figure 6.6 (for decaffeinated coffee) produce different caffeine percentages. Even without knowing why, the SE can change the settings of the extraction machines (or the instructions to operate them) in order that their averages will all be equal. The SE will use the relation of percent caffeine with extraction time determined by linear regression.

6.4.4 Feedback control

The influence of unknown noise factors can also cause a drift of the characteristic of interest. To avoid such drifts away from the target, the SE can use a *feedback* control system. When there is a signal that the characteristic of interest moves away from target, then the setting of a downstream control variable is adapted to adjust the characteristic of interest back to the target. Feedback control is re-active, whereas feedforward control is pro-active.

Example
Driving a car on a road is a prime example of a feedback control system. When the driver notices that the car has moved to the center or the edge of the road, she uses the steering wheel to return to the proper track. Many processes in industry vary continuously under the influence of weather conditions. If the SE can identify a control variable and develop a predictive model, the SE can adjust the process and therefore largely eliminate this variation.

6.4.5 Narrow the tolerance for noise variables

The model tells the SE how the noise variables influence the characteristic of interest. The variables with the biggest impact are the best candidates for improvement efforts. If an SE can narrow the tolerances of these noise variables, he might reduce the variation of the characteristic of interest enormously. Establishing the tolerance limits is also known as *Tolerance Design* and is especially challenging when the tolerances of several factors must be decided upon simultaneously.

Example
The throughput time for an investment request process in a financial services organization was largely influenced by the time it takes the Purchasing Department to get quotations from potential suppliers. In order to improve throughput time, the organization therefore set an upper bound tolerance limit of 3 weeks for Purchasing to get the quotations determination. The challenge for Purchasing was then to organize the work differently, to be able to meet the new requirements.

6.4.6 Reduce the effect of a noise variable

Narrowing tolerances is often cost prohibitive. Sometimes a cheaper solution is available: simply using a different setting of a controllable factor can reduce the effect of a noise variable on the characteristic of interest. This concept of dealing with noise through adjustment of a controllable factor is called *Parameter Design* and was introduced by Taguchi (1986) as part of his methods for *robust design*. Parameter design makes use of

interactions between control variables and noise variables (see the example below). Such interactions are not obvious and can only be identified when the SE executes an appropriate designed experiment (DOE) to investigate them.

Example
On-time delivery is critical for a transport company. For a certain customer the drivers may use two alternative routes. In a project to improve customer satisfaction the SE decided to run an experiment with the travel time as the characteristic of interest. A potential influencing factor was the amount of precipitation, which is an uncontrollable noise variable. The SE collected the data from the experiment, and the analysis revealed the interaction plot of Figure 6.7. The travel time for route A is on average smaller, but with high precipitation this route takes more time. From a planning point of view, however, route B has advantages: the average travel time may be longer, but the variation is also much smaller. It is therefore easier to predict when a delivery will arrive at the customers.

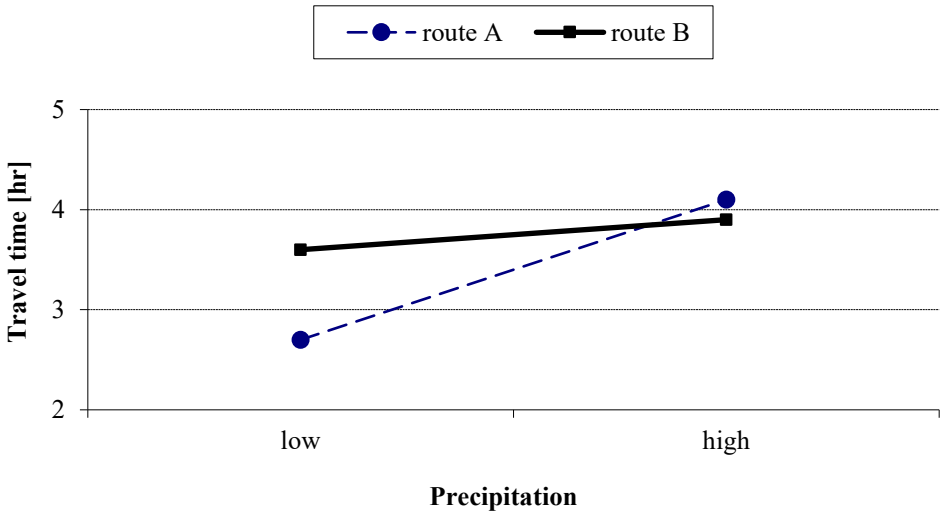


Figure 6.7 Interaction between route and precipitation

6.4.7 Make a list with improvement actions for disturbances

It is difficult to use statistical methods (testing hypotheses or design of experiments) to prioritize disturbances. In earlier sections (6.1.1) the FMEA (failure mode and effect analysis) was introduced as an instrument to identify disturbances. An important objective of this instrument is to determine the risk of a disturbance. Based on the so called RPN (risk priority number) priorities are assigned to potential disturbances, so that the SE can make improvement plans for these and eliminate the influence of, or prevent, these disturbances. The project team of the SE will generally be capable of doing this with brainstorming.

6.4.8 Conclusion

In the piloting solutions phase the following steps must be completed:

- Based on the model between the characteristic of interest and the influencing factors (noise and controllable factors), the optimal settings of control variables will be determined.

- If necessary, control systems have been designed, or tolerances for some influencing factors have been narrowed.
- Improvement actions have been defined for the most important disturbances, to eliminate their effects or prevent their occurrence.

Section 6.5 - Adjust the quality assurance system

6.5.1 Objectives

In Juran (1989) the different aspects of quality management are discussed. Juran is clear about the distinction between quality improvement and control. Quality improvement is usually a short term, intensive project-oriented search for process improvement. Quality control is a long term, less intensive monitoring activity aimed at detecting and responding to irregularities. In the final phase of a project, the SE adjusts the quality control system. On the one hand she probably has implemented changes in the process, and these changes have to be reflected in the control system. On the other hand, she probably has gained new knowledge about the process, and this knowledge can be used to improve the control system. In this section we describe some popular methods for quality control. The essence of these methods is that they systematize the day-to-day management of quality. We first discuss the implementation of quality control in the organization.

6.5.2 Quality control in the organization

An effective quality control system in an organization needs to be properly designed. The quality control system has many stakeholders, and their roles and responsibilities differ. These include the following:

- **Top management:** coordinate improvement projects.
- **Supervisors and process owners:** manage improvement projects to tackle chronic problems and organize quality control.
- **Operational employees (e.g., operators):** respond to sporadic problems.
- **Automatic process controls:** handle predictable day-to-day problems.

Quality control is the responsibility of everyone in the organization including most importantly the operational employees. Wherever possible, quality control tools should be automated and computerized. Operators should not be bothered with the predictable variation and disturbances, but they should focus on sporadic, unpredictable problems. Their task is to signal and respond to these disturbances as fast as possible. However, operators need to have appropriate training and skills to manage this task. The three key requirements are:

1. Clear instructions and training:
    a. Clear and complete operating procedures
    b. Characteristics of interest and performance standards
    c. Adequate training
2. Easy to use tools to monitor performance:
    a. Dashboards and control charts for real-time visual assessment and management
3. Ownership/authority to act:
    a. Authority to adjust and act
    b. Process knowledge, know-how and guidelines
    c. Capable process.

Operators cannot be made responsible for quality control when all three requirements are not fulfilled. Top management and the supervisors must ensure that operators are prepared to assume this responsibility.

The mini-company

One way to organize and empower operators is to form *autonomous groups*, working as *mini-companies* (Suzaki, 1993). A mini-company is a company within the organization, with its own name and mission statement, process description and operating procedures, objectives and performance indicators, report system and documentation. The mini-company gets a budget from the *banker* (depending on its objectives), there are *suppliers* and *customers*, and there are *personnel*. The dashboard of the mini-company contains all relevant characteristics of interest, allowing the mini-company to monitor its own *business*. A mini-company is typically organized around a key-process, or a physical location of a production line.

6.5.3 Acceptance sampling

Testing batches of products is one of the oldest quality control methods. Today it is often the supplier who checks whether the batch conforms to the requirements, but originally the customer tested to determine whether he would or would not accept the batch (hence the name). Supplier and customer have to negotiate to establish quality acceptance requirements. Usually, the customer is willing to tolerate a few defective items per batch, but he wants to be safeguarded against batches with a high number of defective items. If the customer requires no defective products at all, then a 100% inspection may be the solution. But this is cost prohibitive and is not viable. In practice it is nearly always the case that a small random sample of the complete batch is tested for lot acceptance. This is economically practical but not without risks.

Risks of sampling

The first risk of taking a sample is that it may not represent the batch very well. The best guarantee for a representative sample is to take a *random* sample, i.e., each item in the batch has an equal chance of being selected for testing. In reality, however, a random sample is an idealization. Most samples are not purely random but are rather systematic (e.g., one item from each filler head taken every hour) or stratified (e.g., a few zip codes first and then some addresses per zip code). The danger of any sampling scheme is that the sample may not be totally representative of the batch.

The second risk is sampling variability gives a too positive or too negative picture of the batch. With relatively few defective items in the sample, the batch might be falsely accepted; this is called the *consumer's risk*. With relatively many defective products in the sample, the batch might be falsely rejected, the *producer's risk*. A good sampling plan considers and balances both risks to the satisfaction of both customer and supplier.

Sampling by attributes

Acceptance sampling often has a very simple structure when the characteristic of interest is an attribute. If batches of a fixed size are tested for the number of defective products, then a *single sampling* plan is determined by only two numbers:

- The sample size n: the number of items from each batch that is tested
- The acceptance criterion c: i.e., the maximum number of defective items in the sample that is still allowed to accept the batch.

An illustration and a few characteristics of acceptance sampling are discussed in the following example.

Example
A supermarket receives a certain type of product in batches of 1000 vacuum bags. A few leaky bags are not a problem, but too many gives a lot of extra work for the personnel of the supermarket, and moreover there might not be adequate supply for all customers. To avoid batches with too many leaky bags, the supplier inspects every batch with the (n=100, c=1) sampling plan: thus, from each batch 100 random bags are sampled, and the batch is accepted when 0 or 1 leaky bags are found. A rejected batch is inspected completely (100%). A leaky bag is obviously replaced by a good bag. The quality of the sampling plan is expressed by its *operating characteristic* (OC) curve. See Figure 6.8, a graph of the acceptance probability versus the process quality. If, for example, 2% of the bags are leaky, then you may read from the graph that the acceptance probability is 0.4.

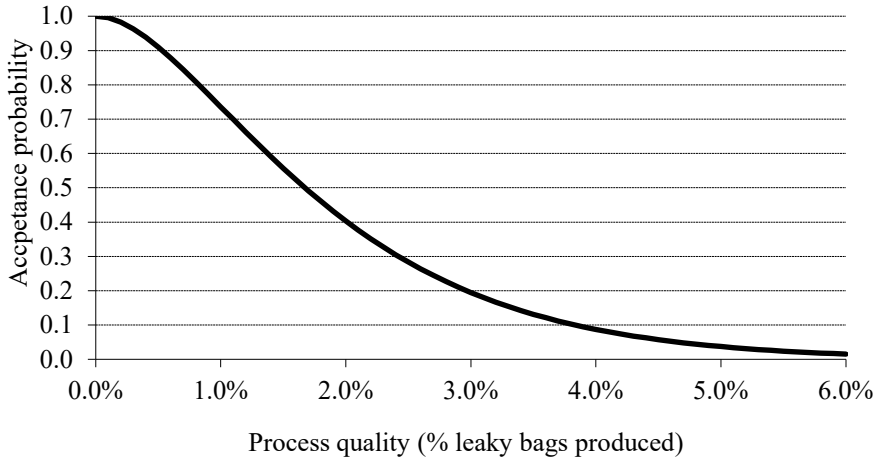


Figure 6.8 Operating characteristic curve

Acceptance sampling requires that the supplier and customer agree on quantities:

- **The Acceptable Quality Level (AQL):** The quality level, in terms of percent defective, which if attained will result in 95% of the product being accepted by the consumer. The producer's risk, the probability that an acceptable batch is rejected, is 0.05. In the example, the AQL is less than 0.4%.
- **The Limiting Quality Level (LQL):** The percentage of defective products that is not acceptable for the customer. Batches from a process at LQL level will be rejected with probability 0.90. The consumer's risk, the probability that unacceptable batches are nonetheless accepted, is 0.10. In this example the LQL is 3.8%.
- **The Average Outgoing Quality Limit (AOQL):** The worst possible average quality level that the customer receives. When there are few leaky bags, then the average outgoing quality is good, of course. But with many leaky bags, the sampling plan will reject the batch with high probability, and all leaky bags are replaced. Thus, the supermarket is spared the trouble of the supplier. Somewhere in between these two extremes the average outgoing quality is determined.

Because of the discrete nature of sampling plan definitions the probabilities mentioned are not exact.

6.5.4 Sampling plans for attributes

Supplier and customer have to agree on AQL and LQL, and then the matching sampling plan can be selected. The cost of sampling will also be important, and the AOQL probably as well. With only single sampling plans to choose from, it might be impossible to satisfy all criteria.

A single sampling plan is simple but not flexible: the sample size of each batch is fixed, irrespective of the quality. The disadvantage is that a lot of unnecessary work is done, when in fact the quality of the batch is either very high or very low. Sampling plans consisting of more stages rectify this problem: if the first sample is not clear enough, then a second sample (and possibly more samples) is taken. For the same performance a double sampling plan requires on average fewer products than a single sampling plan.

For further details see Duncan (1974) and Schilling (1982).

6.5.5 Sampling by variables

Variable measurements are more informative and therefore more efficient than attribute based measurements. This means that for similar statistical performance, as represented by an OC curve, we can make decisions with fewer variable measurements compared to attribute measurements. Only one variable measurement from a batch can already be enough to decide whether the whole batch is conforming. Variation due to sampling and measurement has to be considered (see Montgomery (2012) for further details).

Example
Decaffeinated coffee extract may contain at most 0.1% caffeine. The test for each batch consists of one caffeine measurement from an operator. The batch is rejected when the measurement is larger than 0.08%. A batch of coffee extract is homogenous, so there is no sampling variation. The measurement variation has standard deviation 0.0083%. Assuming that measurement errors are normally distributed, the SE can simply compute the acceptance probabilities for all possible caffeine percentages, to get an OC-curve as in Figure 6.8.

6.5.6 Conclusion

Acceptance sampling is a method to prevent exchange of sub-standard batches. It has been noted however that acceptance sampling is not effective against isolated, sporadic deviations in quality: the odds of finding a few defective items in a batch of a thousand items in a sample are low. 100% Inspection is necessary when not even a single defect is allowed. It may be clear that acceptance sampling is a reactive strategy that seeks to rectify or remediate poor quality. A fundamentally better solution is, of course, to prevent these isolated disruptions from happening in the first place.

Section 6.6 - Statistical Process Control (SPC)

Among quality practitioners, a distinction is made between quality improvement and quality control. Quality improvement refers to the systematic and focused pursuit of substantial improvement of a process ("breakthrough") as advocated by Juran (1989). Such improvement is usually achieved through focused projects that solve chronic problems once and for all.

Once the quality issues have been addressed, the focus shifts to control where the goal is to monitor the process for any changes or deviations from the expected level of performance. If a change or deviation is detected, corrective actions are taken immediately to return the process back to normal performance. Hence, a key advantage of the use of SPC is the accumulation of knowledge concerning unwanted process interventions. Control is about establishing a system for responding adequately to changes in process performance. In the final phase of a project, the SE sustains the gains of the improvements. To achieve this, guidelines for the detection of and reaction to changes and disturbances in the corresponding process are developed. In practice, there should be in-line inspections and a control plan which specifies appropriate interventions to mitigate these irregularities. This approach is reactive in nature and suitable for dealing with incidental ("sporadic") problems.

6.6.1 Control systems

In most organizations there exists a quality control and assurance system (based on ISO 9000 or ISO/TS 16949). When the SE has found solutions for the problem and implemented changes in the process, she must adjust the control system. In the course of improving the process, she probably has also gained new process knowledge and can use this knowledge to improve the control system. In this section we discuss the implementation of statistical quality control in an organization. The aim of Statistical Process Control (SPC) is to respond adequately to disturbances and irregularities in processes.

Disturbances acting on the process do not necessarily mean that defects are being produced, but rather cause the process to deviate from normal performance. A process operating in this manner is said to be out-of-control. The person who is responsible for the process takes immediate action when there is enough statistical evidence for this. The action to be taken is prescribed by the Out-of-Control Action Plan (OCAP), a document containing known problems and appropriate control actions. SPC is therefore a systematic approach to determine when action must be taken and in what way.

Statistical Process Control (SPC) started around 100 years ago. Walter A. Shewhart proposed in 1924 the so-called control chart, "a form which might be modified from time to time, in order to give at a glance, the greatest amount of accurate information." It is a trend chart with reference lines which represent the limits of normal process inherent variation (Shewhart, 1931). This control chart is a statistical tool used to monitor process control.

The control chart was introduced when management realized that to manufacture good products one should monitor the underlying process that generates these products. SPC was introduced as an important part of quality control activities. Within statistical engineering, SPC (also referred to as Statistical Process Monitoring: SPM) plays an important role in sustaining results.

Statistical Thinking, as defined in Hoerl (1996) and introduced in Chapter 1, posits that all work can be defined as a series of interconnected processes. It assumes all processes exhibit variation, and the key to success is to identify and reduce this variation. The process of implementing of SPC has been described by Does et al. (1999). We will describe SPC as a hands-on methodology supported by a control plan to analyze, improve and monitor processes. SPC can be applied for all kind of processes (e.g., in manufacturing, services and healthcare), and its implementation starts usually in operations. We assume a manufacturing context in the following discussion of SPC implementation.

6.6.2 Phases in the implementation of SPC

In this handbook various causes of SE solution deployment issues have been discussed. In a similar manner, there may be difficulties in implementing SPC procedures. There are several obstacles mentioned in the literature, such as lack of management and employee commitment, lack of knowledge, lack of training of SPC techniques, poor project support, and fading attention after the first introduction of SPC (Lockyer et al., 1984; Dale and Shaw, 1991; Gaafar and Keats, 1992; Mann, 1995).

The implementation of SPC requires an organizational culture change. The path to this change is populated with obstacles.

Based on our own experience with implementations of SPC (see Does et al., 1999), the following organizational obstacles are listed:

- It takes years to implement SPC in an organization.
- Time and money must be invested before you will receive the benefits of SPC.
- Constant attention and support of top management is necessary.
- SPC demands delegation of tasks, responsibilities and authority to the lowest possible level.
- Implementation of SPC must be guided by an expert with thorough knowledge and practical experience in statistics. Often an external consultant will be hired for a short period.
- The organization must have an evidence-based culture and data analytics mindset.
- Teamwork and project management skills are essential.

These methodological issues related to SPC can be avoided by carefully planning the implementation stages as described in this section. In the next section, an organizational structure for SPC deployment is described. Like Lean, SPC is a methodology that is usually implemented with shop floor operator teams. Before pilot teams are formed, top management should be convinced that it will be beneficial for the organization. Sometimes the motivation to start SPC comes from outside when customers require their suppliers to use evidence-based quality control methods. For example, Philips started implementing SPC because Ford mandated use of SPC by its suppliers. Similarly, Hollandse Signaalapparaten (Thales) and Fokker Aerostructures felt strong pressure from their customer, Lockheed.

After top management has been convinced that SPC will be beneficial, the following four stages may be implemented:

- **Stage 1:** Awareness

- **Stage 2:** Running some pilot projects
- **Stage 3:** Integral implementation in operations
- **Stage 4:** Setting the stage for SPC.

Stage 1: Awareness

A meeting to raise awareness of SPC for management and staff of the organization is a good start. The objective of this meeting is to let management and staffs become familiar with the fundamentals of the SPC approach and its impact for the organization. In Does et al. (1997), the agenda of the awareness meeting is described. The agenda includes the following topics: the shift from detection to prevention, tasks and responsibilities, establishing the capabilities of a process, dealing with variation, teamwork and project management, financial and non-financial benefits of SPC.

To get the best results from the awareness meeting, thorough preparation is necessary. The support of an (external) expert can be very useful. As an alternative, management and staff could visit an organization which has been using SPC for a while. After the awareness meeting, the next step is for top management to form a steering committee to plan for the implementation.

Stage 2: Running some pilot projects

To be successful, the value of SPC must be demonstrated. To this end, pilot SPC projects should be selected by the steering committee. Ideally, well-known process problems or issues should be identified for the pilots. Carefully trained and motivated teams are then assigned the task of bringing the selected processes under statistical control using the ten-step activity plan described in the following sections. This should be a cross-disciplinary team, which we will call “process action teams” or PATs.

Stage 3: Integral implementation in manufacturing

In this stage it is necessary that one or more employees within the organization will be appointed as the SPC coordinator(s). They will take over the task of the external SPC consultant as needed. The SPC coordinator plays a key role in the further implementation of SPC and will have a close contact with the steering committee. After the pilot phase, more PATs will be formed by the SPC coordinator and steering committee. The PATs get the assignment to implement an operational SPC functionality based on the ten-step activity plan. All operational processes must be controlled in this way. It is likely that a lot of processes must be handled by cross-disciplinary PATs. This implies that it will take several years to finalize this stage.

Stage 4: Setting the stage for SPC

If the PATs have brought the process in statistical control, their follow up will be continuous improvement. In this stage, the approach should be actively extended to other departments. One may expect that the PATs influence departments like product development, purchasing, marketing, finance, human resources, and other supporting departments. In these departments the work is also organized in processes. The in-house experience acquired practicing SPC in manufacturing can be used to stimulate other departments and even suppliers in the usefulness of SPC. The SPC coordinator and other colleagues may assist with the expansion.

6.6.3 Organizational structure for SPC implementation

As we have mentioned before there is a need for an organizational structure for SPC implementation and the PATs. Figure 6.9 illustrates the fact that both top management and steering committee play a supporting role for the PATs.

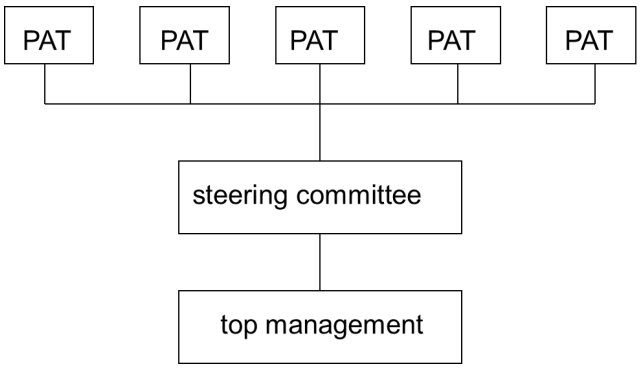


Figure 6.9 Organizational structure for SPC implementation

Top management

Top management is leading the implementation and establishes a steering committee. Key responsibility is to monitor the progress based on information from the steering committee.

Steering committee

The steering committee directs and controls the implementation process. It is quite common that the operations manager is the chairman of the committee. Typically, the SPC coordinator and managers of purchasing, development, quality and maintenance are also members of the steering committee. Managers from other departments may step in as needed. The main tasks of the steering committee are to (Does et al., 1997): initiate and promote SPC; provide methods, resources, and guidance for decision making; monitor of the progress; report to the top management.

Process actions teams

SPC is implemented by PAT teams which consist of employees from all departments involved. In manufacturing you primarily need operators. Their knowledge and hands-on experience are crucial to make SPC successful. SPC can also be applied to service processes, but some changes have to be made (Roes and Dorr, 1997). In service organizations the PAT team will consist of the account managers or relation managers (i.e., the employees who have direct contact with the customers). The goal of the PATs is to bring the process under control using the ten-step activity plan, which is described in the next section.

6.6.4 Methodological part of the framework: the ten-step activity plan

Implementing the SPC method follows a ten-step activity plan (Does et al., 1999):
1. Process Description
2. Cause-and-Effect Analysis
3. Risk Analysis
4. Improvement Actions
5. Define Measurements
6. Data Quality and Measurement System Analysis
7. Control Charts
8. Out-of-Control Action Plan (OCAP)
9. Process Capability Study
10. Annual audit and Certification

Steps 1 through 5 resemble a process FMEA (Failure Mode and Effect Analyses). Steps 6, 7 and 8 define the measurements and control loops. Finally, steps 9 and 10 involve assessment of the capability of the process and a certification process.

Step 1: Process description

This step starts with a description of the process on a macro level. Usually, the SIPOC flowchart is used, where SIPOC stands for Supplier (provider of the process's input), Input (the materials, resources or information required to execute the process), Process (set of activities that transforms input into output), Output (the product or service resulting from the process) and Customer (receiver of the output). The next step is to describe the process by defining the process steps in terms of only one distinct transformation. It is natural to use verbs to describe these steps. In Figure 6.10 an example of a SIPOC is given.

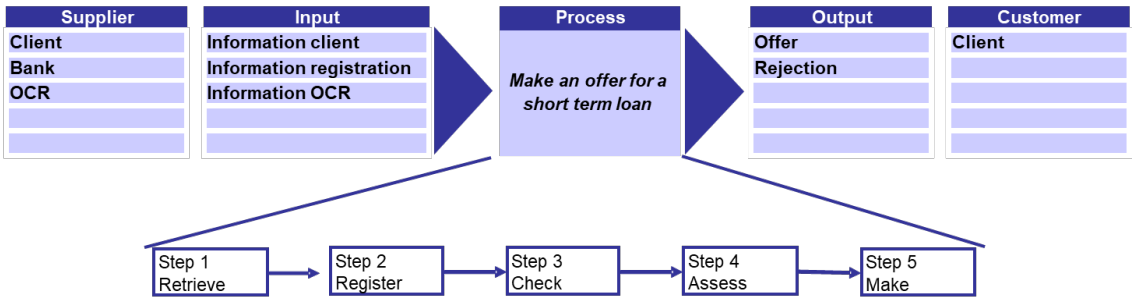


Figure 6.10 SIPOC of processing requests for loans

The flowchart provides a detailed description of the process. It can be used effectively to expose and clarify more granular sub-tasks in the process. The flowchart should be created by the entire PAT to ensure all members are aligned.

Step 2: Cause and effect analyses

A tool that can be used for steps 2 and 3 is a Failure Mode and Effect Analyses (FMEA), (Stamatis, 1995). A failure mode is a malfunction of a process step often caused by a disturbance. In this context, a disturbance could mean that a process step or task was not executed at all, or that it was not executed properly. In addition, it may be the case that usual practice created undesired side effects ("unintentional consequences"). The team should discuss, for each process step, the following:

- Which disturbances occur in this process step?
- What are causes of each disturbance? What is the effect of each disturbance on the process?

An alternative tool for step 2 is using Ishikawa diagrams (Wadsworth et al., 1986). The importance of each disturbance is determined in step 3: risk analysis.

Step 3: Risk analysis

In this step, the risk for each cause-and-effect relation is calculated (see Figure 6.11). The Risk Priority Number (RPN) of each combination is calculated by multiplying scores for:

- the frequency of occurrence (O) of the cause;
- the severity of the effect (S) of the cause;
- the detectability (D) which indicates how easily we can detect and react to the cause of a failure mode.

Frequency (of the cause)		Severity (of the effect)	
1	Hardly ever (did not yet occur this year).	1	Effect hardly noticed; does not cost time or money.
3	About once a month.	3	Effect concerns rework by the operator; costs time (> 10 min.).
5	About once a week.	5	Rework by others within the department; operator is irritated.
8	About once a day.	8	Rework by other department.
10	Cause occurs each hour.	10	Rejection of product by customer; danger for operator.

Detectability	
1	Cause readily detected and tackled.
3	Cause detected immediately after its occurrence and mostly tackled
5	Cause detected at next process step; can be tackled partly.
8	Cause detected just before shipping of product; or hard to tackle.
10	Cause detected too late or cannot be tackled before it is shipped to the customer.

Figure 6.11 An example for assigning scores in an FMEA

The priority of a cause-and-effect relation is given by its Risk Priority Number: RPN=O×S×D. Scores are rated on a scale from 1 to 10. An example for assigning these scores is given in Figure 6.12. High risk numbers should be analyzed for possible improvements in step 4.

Failure mode	Effect	S	Cause	F	Detection	D	RPN
Disturbance 1	*Effect*	*3*	*Cause*	*2*	*Inspection*	*4*	*24*
Disturbance 2	*Effect*	*8*	*Cause*	*5*	*Inspection*	*9*	*360*

Figure 6.12 An illustration of an FMEA form

Step 4: Improvements

The PAT can use the Pareto principle to prioritize improvements efforts. Cause-and-effect combinations with the highest risk scores should be addressed first. A high RPN score can be reduced by lowering the occurrence of the cause, improving the detectability, and reducing the severity of the effect. A useful technique is poka-yoke, also called mistake proofing (Shingo, 1986). Poka-yoke is a work process strategy designed to prevent inadvertent errors made by workers performing a process. Manufacturing examples include use of visual aids, color coding, and error-proof design. An alternative is to build in redundancy which means having back-up systems in case of failures. Also, preventative maintenance can reduce risks from equipment malfunction.

Step 5: Define measurements

The objective of this step is to select the parameters for controlling the process. The goal is to unravel the problem into clear characteristics and find metrics that are related to these characteristics. For this purpose, the PAT should develop a control plan which is a description of the measurements that will be collected, monitored and analyzed (see Figure 6.13). It is a survey of all measurements in the total process. When in doubt, all measurements should be considered. Subsequent data analysis can indicate which measurements are most relevant to the problem. This resembles the way in which the critical-to-quality characteristics are selected in a Lean Six Sigma project (De Koning and De Mast, 2007). Also, the use of generic project definitions in industry and services may be helpful in

the selection of the parameters because these generic project definitions include the relevant measurements (Lameijer et al., 2019b).

CONTROL PLAN						
Process:	Counting of bank notes				*Version:*	
Proc. owner:						
Measurement	***Who***	***How***	***Where***	***When***	***Norm / spec.***	***Which OCAP***
Check for forgeries	Forgeries team	Visual	Desk	For all rejected notes	Authenticity stds.	Forgeries OCAP
Discrepancies check	Automatically	Compare machine count to client's count		Each deposit	Difference should be zero	Discrepancies handling OCAP
Reject rate	Operator	Control chart	Workfloor	1 / hour	LCL = 1.4 UCL = 2.6	Machine OCAP

Figure 6.13 An example of a control plan

The control plan defines the control loops and associated documentation and responsibilities. The FMEA and the cause-and-effect relations from step 4 form a good starting point for creating a control plan. The control charts and OCAP's mentioned in the control plan of Figure 6.13 will be discussed in steps 7 and 8.

Step 6: Data Quality and Measurement System Analysis

The first step is to establish the validity of the measurements. Validity is the extent to which a concept, conclusion or measurement is well-founded and corresponds accurately to the real world. Questions that should be considered are:

- Are definitions and calculations of the measurement characteristic correct?
- Do we measure the right aspect of the characteristic?
- Are there perturbing influences that make measurement results invalid?

Before the measurement collection starts, the PAT plays devil's advocate and has a brainstorming session, in which possible validity problems are identified and potential problems are addressed. Also, the process of how a measurement value comes about are checked (especially if the measurement procedure is automated). Furthermore, face validity of the dataset is assessed by doing a Sanity Check of the data. For example, answers are given to the following questions:

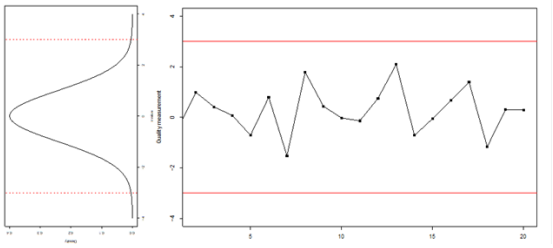
- Is the dataset complete?
- Are expected correlations present?
- Does the dataset contain impossible or improbable values?
- Do observations sum to expected totals?
- Are there strange or unexpected data patterns?

If one uses a measurement device to obtain a measurement, then one must check the measurement error when the measurement outcome does not correspond to the "true value" (i.e., does not correspond to the real world). Two components are distinguished: systematic error and precision of the measurement method. The systematic error is usually assessed by doing a calibration study. For precision, a repeatability and reproducibility study (Gage R&R study) is carried out (e.g., Kane, 1989). Gage R&R studies are designed for quantitative measurements to quantify the repeatability (variation of the measurement device) and reproducibility (variation in using the measurement device). Alternative analytical tools are

available for attribute data or when the measurement process is destructive (see e.g., Futrell, 1995). In most projects in service industries, one uses time stamps which can be assumed to have negligible measurement error.

Step 7: Control charts

In SPC the most popular statistical tool is the control chart, which was introduced almost 100 years ago by Shewhart (Montgomery, 2013 and Roes and Does, 1995). The control chart is a trend chart that plots data in chronological time order. The chart includes reference lines called control limits which define the level of natural variation in the process. Control charts are used to discriminate between common and special causes of variation. Common causes of variation are the collective effect of many minor independent influences. This variation is considered to be the natural "white noise" variation exhibited by the process. It is futile to act in response to this natural variation. In contrast are special causes of variation that act on the process. The objective of SPC monitoring is to detect and mitigate these special causes of variation. In Figure 6.14 an illustration of a Shewhart control chart of a quality characteristic which follows a normal distribution with mean μ_0 and standard deviation σ_0 is given.



$$UCL = \mu_0 + K\frac{\sigma_0}{\sqrt{n}}$$

$$LCL = \mu_0 - K\frac{\sigma_0}{\sqrt{n}}$$

Figure 6.14 Shewhart control chart when the quality characteristic follows a normal distribution

The Upper (and Lower) Control Limits (abbreviated by UCL and LCL) are the limits of the process inherent natural variation. When the parameters μ_0 and σ_0 are known, a typical value of K according Shewhart is 3 (Shewhart, 1931). The probability of exceeding the control limits with K=3 is a very low value of 0.27%. If an observation is outside the control limits, the process is said to be out-of-control and the operator will look for a special cause.

In practice, the parameters μ_0 and σ_0 in the control limits must be estimated using a reference sample. This takes place during Phase I. Because of sampling variation, different Phase I samples will provide different parameter estimates which will lead to different estimated control limits. The performance of the control chart is then conditional on these obtained estimates. The effect of Phase I estimation has received much attention in recent literature (Jensen et al, 2006, for an overview).

Two additional commonly used types of control charts are the Cumulative Sum (CUSUM) and the Exponentially Weighted Moving Average (EWMA) control charts. Each of the three types of charts has its own characteristics. This makes each applicable to detect specific types of out-of-control situations. For example, the Shewhart control chart is better suited to detect large shifts, while the CUSUM and EWMA yield better detection capabilities against small sustained shifts. These differences motivate comparative studies, where the control chart capabilities are evaluated under different disturbance scenarios. Zwetsloot and Woodall (2017) perform a comparative study on the conditional performance of the Shewhart, CUSUM, and EWMA control charts, where they compare the effect of estimation error across these charts. For each of these charts, the first step (Phase I) is to estimate the in-control behavior of the underlying process, before one can start the monitoring stage (Phase II).

Step 8: Out-of-Control Action Plan (OCAP)

To improve control there are two types of incident handling. As mentioned in step 4 some disturbances with high risks are fixed using poka-yoke devices, but other incidents still can occur and should be dealt with if they occur. All out-of-control behavior of the important (process control) characteristics should be addressed. Control loops have two elements: a trigger for intervention if an out-of-control signal or a disturbance occurs and an intervention to solve the problem. The central tool is again the control plan. An example is given in Figure 6.15.

The control chart effectiveness is improved when there are guidelines on which action must be taken when an out-of-control situation occurs. The OCAP is applied when the control chart signals an out-of-control situation and then guides the employee's intervention (Sandorf and Bassett, 1993). Guidelines for OCAP include:

- documenting process knowledge and lessons learned
- continuous updating as the OCAP is a "living document"
- providing a log of detailed problem description and solution.

Note that the application of control charts without OCAP, is like driving a car without steering wheel. Past learnings and experienced must be documented and leveraged whenever possible.

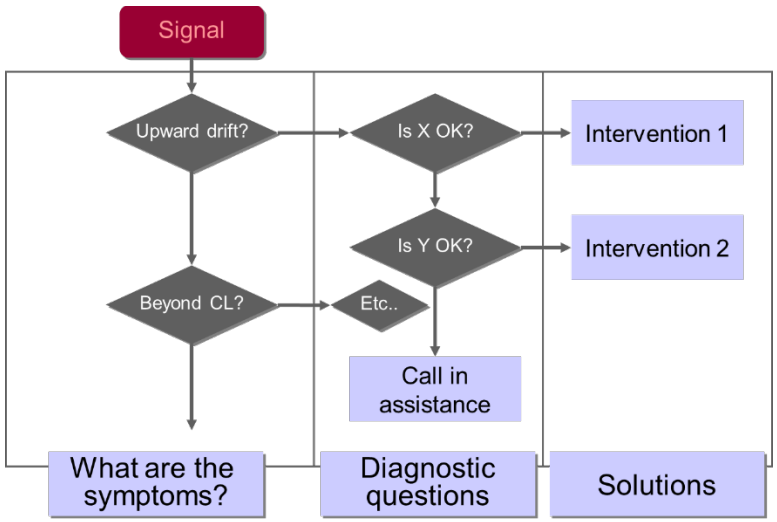


Figure 6.15 Typical form of an OCAP

Step 9: Process Capability Study (PCS)

The Process Capability Study (PCS) provides a means to relate the performance of a process to requirements, standards and other relevant benchmarks. A PCS is about analyzing the current process performance and making a diagnosis. The analysis in this step may entail a range of techniques as illustrated in Figure 6.16. The PAT selects the appropriate techniques for the project at hand. A clear diagnosis helps to redefine the project objectives that were initially established at the start of the project.

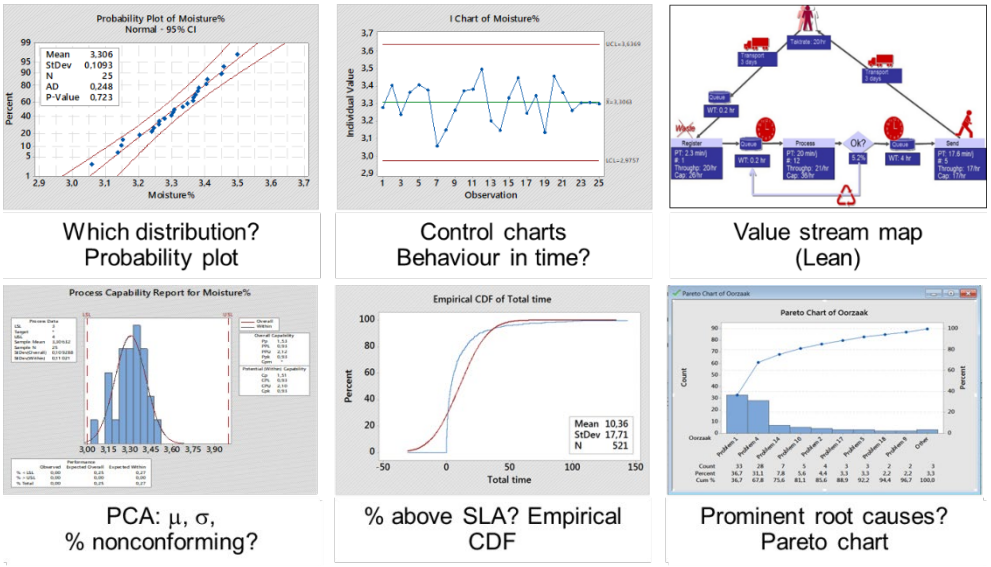


Figure 6.16 Process Capability Study

Step 10: Annual audit and certification

In the last step the activities of the PAT and the performance of the process will be audited by the steering committee. A checklist can be used to make sure that the PAT knows what is expected (Section 12.2 in Does et al., 1999). The audit of the steering committee includes the activities on the shop floor and a check on the follow-up activities. When the audit is positive the PAT members receive a certificate as an official acknowledgement of their effort. An example of a certificate is given in Figure 6.17. The certificate is valid for one year only and should then be re-audited. This will also stimulate the continuing attention of the PAT.

Figure 6.17 An example of a certificate of an SPC point

6.6.5 Conclusion

The theory of SPC, as described in this chapter, is applicable in most organizations without large modifications. Using the four stages and the ten-step activity plan ensures management and employee commitment, teamwork and a goal-oriented project approach.

Section 6.7 - Finish the project

The end of a project is a time for celebration. But it is also a moment to document all the results of the project. The SE has put a lot of effort in investigating the process, and the process knowledge that has been gained has permanent value for the organization. *Knowledge management* is a very important topic in organizations where people change positions regularly. If the organization truly values the learnings from the project and aspires to leverage this knowledge in the future, then much care should be taken to:

- Document the results of the project completely.
- Make the documentation easily accessible to others. The structured DMAIC approach of Lean Six Sigma is a great example to achieve this. A review of "lessons learned" should be a compulsory practice.

6.7.1 Follow-up activities

Another important element of this final step is to assure the continuation of the work, which requires a continuous, active involvement of the SE. First, she has to explain the changes in practices and organization to everyone involved, in order to get their approval. In the period immediately after the project, the SE still has the responsibility to ensure that the changes are implemented properly. She has to realize that almost always 'growing pains' are involved. Often improvement actions do not work perfectly on the first try – iterative fine tuning may be required. The difference between a successful project and one that is ultimately a failure is often the attention to the final implementation.

6.7.2 Conclusion

In the control stage the SE adjusts existing control loops or designs new ones. Adjustment of the quality assurance system comprises:

- Acceptance sampling
- Statistical Process Control
- Feedback and feedforward control loops
- Mistake proofing (poka-yoke)
- Maintenance control.

Determine the new process performance: The SE determines the new process capability to judge the realized improvements against the aims of the project.

Finish the project: The SE documents the results and conclusions, secures the improvements and makes a future plan, if necessary.

Section 6.8 - Evaluation and future plans

To monitor the achievements, the SE organizes an evaluation meeting in one or two months. The agenda is to discuss the new experiences and learnings from the improved process, and to bring to light potential new issues with the process. It may be useful to organize periodic evaluation meetings, especially when the SE normally works at another location.

A continuation of the project might be beneficial. *Continuous improvement* is a promising way to reap more benefits from the work that has been done: together with operators and other people very close to the day-to-day running of the process the SE elaborates on the results and suggestions from the project. Another possible next step might be to check if the results of the project are relevant to other processes, factories/plants, departments, etc.

6.8.1 Implementing Statistical Engineering in organizations

The results of the first SE projects are, amongst other success factors, determined by the quality of the applied tools and techniques. For instance, measuring the problem at hand and discovering root causes demands robust and proven statistical methods. Let us assume the SE has successfully executed the first project and achieved demonstrable improvement. As the most pressing area of improvement has been addressed, new problems come to attention. Instead of executing just one more SE project, the organization can decide to deploy a collection of SE projects and decide to implement SE as a strategic organizational change initiative. As more SE projects are executed and the SE methodology is applied more broadly throughout the organization, questions about implementing SE in organizations arise. These questions transcend the area of SE's quantitative tools and project management techniques.

Reasons for implementing Statistical Engineering

Improvement and optimization of processes, products and services have gained increased recognition as sound management practice. A common label for the organizational capability to improve and optimize processes, products and services is Continuous Improvement (CI), for which the dominant philosophy and methodology of SE has much to offer. To achieve the organizational capability to continuously improve processes, services and products, organizations face management challenges such as motivating the organizational adoption of SE, setting adequate goals and performance metrics for all involved in the SE implementation process and subsequently managing the SE implementation process. The rationale for SE implementation lays usually in one or more of the five performance dimensions: quality; dependability; speed; flexibility and costs (Ferdows and De Meyer, 1990).

Success factors for implementing Statistical Engineering

CI literature and adjacent CI methodologies (such as SPC, Lean, Six Sigma, Total Quality Management, Lean Six Sigma) proposes a variety of guidance for the implementation process. Topics covered in these guidance materials include learnings from case studies, deployment strategies and implementation maturity models (see Lameijer et al., 2017 for a review). Key success factors for implementing SE methodology are:

- empowerment and communication with the workforce
- the management of Statistical Engineers and SE projects in a systematic manner
- direct reporting to business executives, and
- an environment of psychological safety.

The first steps in implementing Statistical Engineering

When the organization decides to implement Statistical Engineering as a dominant technique for continuous improvement, there are a few preliminary tasks that need to be arranged prior to a formal start of the implementation, often in a top-down way (De Mast et al., 2012). These preliminary tasks include:

- **Clear vision and top management commitment:** SE is a big commitment and is not something that can be done on the side. Its implementation should be based on a conscious and well thought through decision. This presumes that management has a clear vision of what they want to achieve by implementing SE. First, the top management team should make a clear and deliberate choice to implement Statistical Engineering. Implementation requires a substantial investment in time and resources. A large number of personnel may be asked to take on supplemental work besides their regular work to run projects. Also, significant training investment may be required. The SE initiative should be part of the company's strategy, and management should be able to explain what strategic objectives it desires to achieve by implementing SE (see "reasons for implementing SE"). Further, management should be able to communicate why it has chosen the SE methodology. Failing to produce a brief, clear and inspiring vision, makes it hard to mobilize enough energy in the organization to overcome resistance and invest time and effort in the initiative. In some situations, there may be some reluctance to disturb the status quo and embark on something new like SE. A shared sense of urgency must be created. Top management should communicate across the organization that SE is a priority, why this is so, and what benefits and efficiencies it creates. Top management can demonstrate its commitment by including a financial target for the SE initiative in the annual report: "We expect SE to deliver X million euros savings by 20YY from the combined impact of revenue growth, cost reduction and efficiency improvement."
- **Statistical Engineering organization -- program management:** To start the implementation a program management organization should be established. In particular, the program director (member of the organization's senior management) should be assigned and one or more program managers (or SE coordinators, or a SE steering committee). In the first one or two years, senior SEs will likely be hired from outside the organization (external consultants, or experienced SEs from other companies).
- **Align Human Resource policies:** The organization should think through and clearly communicate how it intends to integrate SE in its HR policies. How many days per week can SEs work on their project? How is this integrated in people's personal targets? Should SEs get a new job assignment? If yes, for how long will they have this assignment, and what happens after that period? Will promotions be tied to SE involvement? What are the guidelines for certification of SEs? Should there be performance awards for best SEs?
- **Establish analytics culture:** Top management and SE program management have a role in making SE the usual way of doing many things. This means that SE vocabulary is used ("Show me the data", "Build the model", "What's the hypothesis?"), and that the SE steps are seen as the logical sequence of steps to take when solving a complex problem. Moreover, top management and program management have an ongoing responsibility to integrate SE activities in the company's strategy, programs, and other initiatives. With SE a company develops a highly effective organization for getting problems analyzed and dealt with – make sure that it is used.
- **Training and project support:** Once SE has established itself in the organization (after one or two years) the organization's senior SEs will deliver the SE training, and junior level SEs will give the introductory SE training. For the first projects the organization

probably needs to call in external consultants to give the training. Project support by a senior SE should be arranged for (at least 1 hour, biweekly per SE project). Additionally, champion reviews must be scheduled (30 minutes per phase, 4 or 5 reviews in total). Furthermore, executive trainings and workshops for champions must be organized.

- **Information technology facilities:** SEs need software to do statistical analyses. Excel is not enough. Minitab and JMP are used in many statistically oriented companies. Nowadays also programs like R and Python have become popular to use. Other implications for IT support comprise the need to have access to open source software tools and the need to increase accessibility to data. If the SE initiative becomes mature (after one or two years) it may be wise to set-up a project database accessible on the intranet so knowledge can be shared.
- **Start the first series of Statistical Engineering projects:** The first set of projects should be seen as a pilot. Although the issues above should be carefully planned before starting the first SE projects, it is only after the first training and projects have started that the organization finds out what works and what does not. The first group of SEs have to do a lot of pioneering work: parts of the organization will not yet have gotten used to the idea of SE, and acceptance will not yet be organization-wide. For this reason, one should select bright, persistent and motivated candidates for the first wave. Although in general SE project selection should be focused on strategically important issues, especially for the first wave one should include at least some projects that have the potential to bring quick wins. SE and the organization have to get accustomed to each other – use the series of projects to learn and adjust, and not worry if matters are not yet perfect.

6.8.2 Managing the Statistical Engineering implementation process

The implementation of Statistical Engineering in an organization requires that the organization goes through a process of implementing a new way of working and solving problems. In addition to training and the other activities as mentioned in previous sections, the organization needs to make several other adjustments to establish an organizational culture and infrastructure to support SE-based continuous improvement. CI literature and methodologies offer a large volume of implementation guidance for this process (see Lameijer et al. 2017 for a review).

A model for implementing Statistical Engineering

Useful guidance for the SE implementation process is illustrated in Figure 6.18 – 6.22 (based on Lameijer et al. 2019b). This holistic framework separates implementation into five phases and further clarifies roles and responsibilities of different organizational functions. The framework structures the SE implementation process in several dimensions that are relevant for the SE implementation leader to understand:

- The first differentiator is the phase or maturity of the implementation. A scale ranging from phase 1 (preparing for Statistical Engineering) to phase 5 (systemic Statistical Engineering) for maturity is designed.
- The second differentiator is the organizational dimension (Structure, Strategy, Systems, Style, Staff, Skills, Values). The strengths of this compartmentalization lay in the collectively exhaustive and mutually exclusive nature of the seven organizational dimensions, as well as the recognition of interrelationship between dimensions.
- Research has shown that various organizational Readiness Factors (RF) are strong predictors of project success. Important RFs for each phase and function are provided. Secondly, critical key Activities (AC) that should be executed by each function are

provided. Lastly, sustainability is about routinizing the SE mindset and practices by the organizational staff. To this end important Sustainability Factors (SF) that can help further advance the SE program are provided.

Phase 1: Preparing for Statistical Engineering implementation

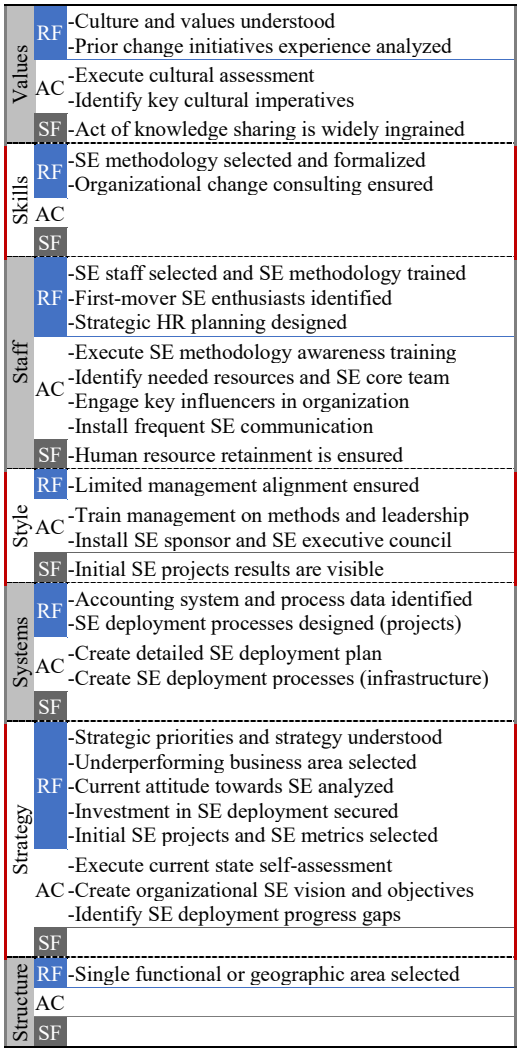


Figure 6.18 Phase 1
Preparing for Statistical Engineering

In the first phase an organization prepares for implementation. Relevant topics are formulation of the justification for SE implementation, creating an understanding of the current attitudes towards SE and creating the foundation for an organization specific implementation plan. Clarity of the business value of SE implementation is the cornerstone for a vision and the intended contributions towards the organization. A SE core-team must be set in place and SE implementation planning and project selection processes must become operational. This first phase produces the first tangible SE project results and is therefore very important. Promising early results support the credibility of the SE leadership team and provide impetus to further development of the SE trained workforce and builds an awareness and willingness of organizational staff to develop and share knowledge on SE application.

Phase 2: Foundational Statistical Engineering

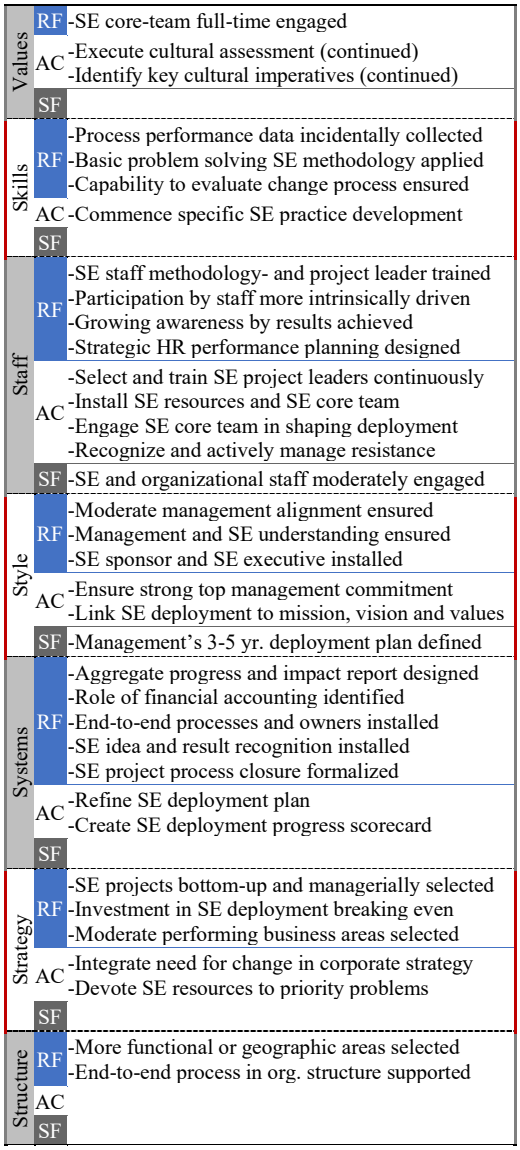


Figure 6.19 Phase 2
Foundational Statistical Engineering

The second phase is characterized by increased interest and participation in SE implementation. SE projects are still chosen opportunistically, and aggregated progress and impact reporting is installed. Leadership is more aligned, demonstrated by, for instance, incidental selection and reviewing of SE projects and structural focus on SE implementation in leadership meeting agendas. The first full-time SE project leaders return to their regular organizational position, and implementation in more than one organizational unit or geographical location is considered. The SE implementation core team develops the capability to evaluate and manage the organizational change process.

Integration of the SE implementation into the organization's existing strategy and strategic objectives is of pivotal importance to ensure that SE resources are devoted to priority problems. SE leadership is further strengthened by continued training efforts and installation of SE leadership teams that safeguard the contribution of SE projects to strategic objectives. The selection, training, support and retainment of SE project leads are further professionalized, and the installed base of active proponents is growing. The organization starts developing an idiosyncratic SE methodology based upon experience. The SE implementation plan is finalized. Further refined, cultural developments are continuously monitored, and cultural imperatives are identified and acted upon. The deliverable for the second phase is a leadership team defined three to five-year SE implementation plan, containing sections on budgets, resource planning, progress ambitions and monitoring, leadership and staff training and retention. In this phase all organization staff has directly or indirectly been involved in SE implementation.

Phase 3: Cross functional Statistical Engineering

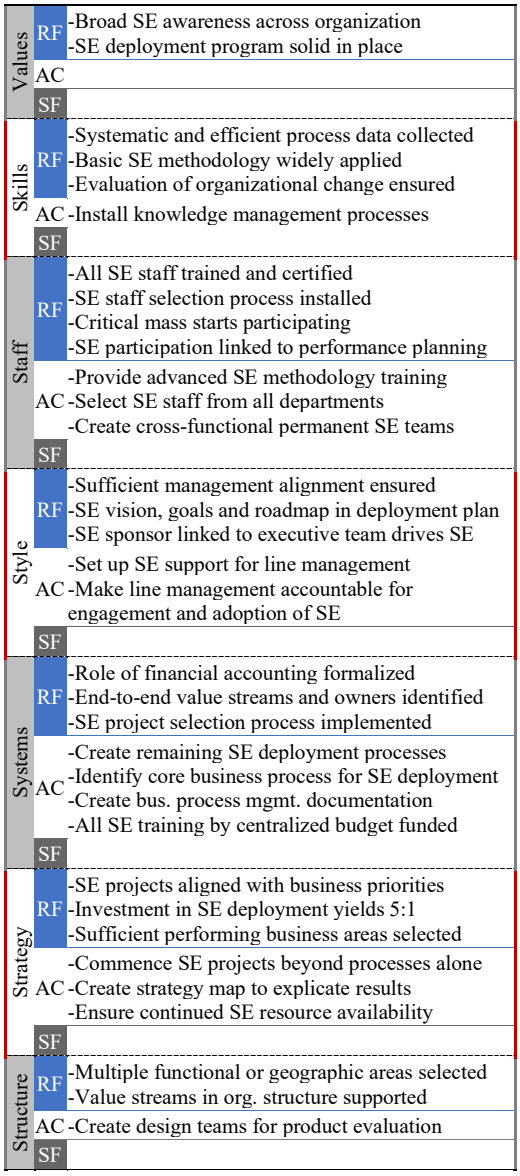


Figure 6.20 Phase 3
Cross functional Statistical Engineering

In the third phase an organization typically targets strategic goal realization with SE efforts. In this phase SE projects are aligned with business priorities and corporate strategy execution. The investment in SE starts to yield significant benefits and the vision, goals and roadmap are integrated in the implementation plan. Leadership takes an active role in SE project selection and reviews and leads the implementation. SE projects are focusing on increasingly more complex problems. SE staff selection processes should consider all different departments in the organization, In this phase autonomous development of cross-functional SE teams are emerging for specific problem solving.

SEs are trained and certifications are granted. A formal selection process for SE project leads is in place, and organizational staff starts engaging in SE activities by means of cross-functional problem-solving teams. The organization is more comfortable with data-based decision making, and the range of SE methods applied becomes more comprehensive. Financial control is engaged in every project. SE implementation progress and results are accurately measured. SE implementation processes become more mature, and more geographical locations and business units become involved. There is a broad SE awareness throughout the organization, and the driving core-team is solidly in place.

The contributions of the SE implementation are made visual and concrete in, for example, a strategy map, and continued resource availability is ensured. The organization's line management becomes more involved in SE deployment and is supported by an infrastructure of dedicated SE resources. The next step is to make line management accountable for SE adoption in their respective areas. More specialized SE training modules aimed at internalization and replication are designed. Compensation of SE staff is tied to project results, and knowledge management processes must be installed to ensure practice sharing.

In this phase core business processes in scope of SE implementation are defined. The development of process management documentation and training is commenced, and clarity over business as usual and SE responsibilities is further refined. Finally, all SE training activities should be funded by a centralized training budget.

Phase 4: Integrated Statistical Engineering

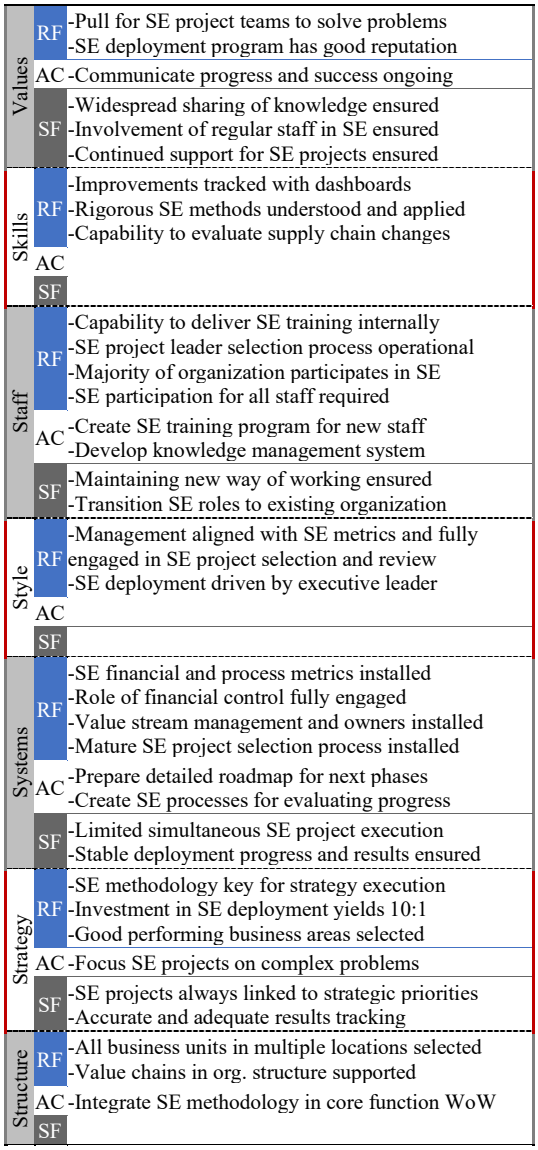


Figure 6.21 Phase 4
Integrated Statistical Engineering

In the fourth phase SE methodology is further ingrained in the organization. Not only are problems being solved, but future business opportunities also emerge from the execution of SE projects. In the fourth phase, SE methodology is considered key for corporate strategy execution. The investments in the implementation (training, projects, etc.) yield significant results. With strategy maturity maps, the contribution and progress of the SE implementation is visualized.

Leadership across the entire organization is aware of the SE implementation and adopts SE methods. The organization can autonomously deliver SE methodology training, and SE staff selection processes are formally in place. Organizational staff teams form temporary SE teams, and most of the organization is involved in SE.

Metrics that measure the SE implementation progress and impact are widely available, and bottom-line financial impacts are visible. A sound SE project selection process is in place. Performance data collection is mature, and rigorous SE methods are applied. This leads to an organization wide pull for SE project teams. SE projects are focusing on more complex problems and use tailored methodology.

The fourth phase ensures that SE project contributions remain focused on the strategic agenda by an accurate tracking of progress and results. The new way of working is ensured by transitioning SE roles and responsibilities into the organization. It is important that SE project momentum is sustained, by limiting the number of projects that are simultaneously executed. Furthermore, widespread sharing of knowledge and practices, communication and involvement (also to staff not involved in the deployment) are pivotal.

Phase 5: Systemic Statistical Engineering

Values	RF	-Strong SE culture and zero-defect mentality -SE deployment integral to culture of business
	AC	-Perform period cultural assessments and act
	SF	-Ensure ability to articulate basic values of SE
Skills	RF	-SE projects take advantage of all SE methodology
	AC	-Create progression to learning organization
	SF	
Staff	RF	-Entire organization participates in SE -SE methodology and system adoption linked to performance planning for all staff
	AC	-Identify and train (new) SE staff continuously -Connect SE involvement to intrinsic motivation
	SF	-Sustained involvement in SE ensured -SE across organizational boundaries ensured -Learning and sharing at all levels enabled
Style	RF	-Management understanding and faith in SE -SE deployment led by CEO with C-level reporting
	AC	-Develop managers dedicated to SE -Create ongoing clarity of SE ownership
	SF	-Creation and sustaining of SE behavior ensured -Statistical Engineering of SE system ensured
Systems	RF	-SE metrics in corporate dashboard integrated -SE project benefits linked to budgeting process -Value stream management has strategic targets -SE project selection process linked to strategy
	AC	-Review SE performance and impact at all levels -Create scorecard cascade at department level -Create core and supporting process maps
	SF	-Consistency in behavior and values ensured
Strategy	RF	-Investment in SE deployment yields 20:1 -Excellent performing business areas selected -Strategy- and product development data-driven
	AC	-Update strategy map for all core processes
	SF	-Link SE activities to strategic goals ensured -SE of Statistical Engineering ensured
Structure	RF	-All business units in all locations selected -SE extends to full supply chain deployment
	AC	-Create working cells (waste and variability) -Create SE methodology integration plans -Extend value chains to suppliers and customers
	SF	

Figure 6.22 Phase 5
Systemic Statistical Engineering

The fifth phase is when SE implementation results in a SE system; organizational routines and a way of working whereby all organizational staff and management are involved in continuous improvement with SE. In this phase SE methodology is fully aligned with corporate strategy execution as SE project metrics are linked to strategic metrics. Leadership visibly demonstrates SE support and active participation in SE implementation. The capability to develop resources by means of training and coaching is fully internalized, and SE staff remains fully trained. Commitment to SE projects for a period is seen as good for career advancement, and all organizational staff commit at least 5% of their time to SE support. SE implementation metrics are fully integrated with common reporting processes and dashboards. SE projects apply all relevant methodologies. there is a strong continuous improvement mentality, and the implementation expands to all functional areas and geographical locations.

Activities in this phase are focused on continuation of the SE implementation and SE methodology adoption. Strategy maps are updated with the latest measurements. SE minded managers are continuously developed, and SE involvement is continuously connected to the intrinsic motivation of junior and senior SE staff. SE implementation performance and impact are frequently reviewed and amended where needed and cultural assessments are periodically performed and acted upon. A learning organization must be created by facilitation of knowledge sharing and benchmarking both internally and externally.

Sustainability is ensured by persistent linking of SE activities to strategic objectives. SE behavior throughout the organization is sustained, and the SE system that comes into existence must also be subject to continuous improvement. For the organizational staff, sustained involvement in SE and ongoing learning between both people and groups about their SE experiences must be ensured. Consistency between the developed SE values and the existing organization must be ensured by ongoing reviews. To do so, the ability to articulate these basic values must be supported.

Learning processes in Statistical Engineering implementations

Implementing SE and its corresponding behavioral routines take time to institutionalize before they collectively provide a strategic advantage. In the implementation process, the organization goes through cycles of trial-and-error and discovers how to adopt the outside CI practices (SE methods) as useful instruments for the organization. This process is a long-term effort that consists of various organizational learning activities. Therefore, the SE implementation process is partially an organizational learning process, but it is also a programmatic adoption of outside SE methodology practices. The SE needs to be aware of this and to plan for organizational learning to take place. The SE should be aware that conflicts about next steps and future directions in the implementation are to be expected and are just manifestations of sense-making processes. Such manifestations are often needed to ensure organization-wide commitment and meaningful change.

6.8.3 Conclusion

After the initial SE projects are finalized and have demonstrated value, organizations start a process of implementing SE as a way to continuously improve products, processes and services. This process must be managed, and therefore, guidance as outlined in this handbook can be used. In addition, SE implementation leaders need to be aware that such an implementation process consists partly of organization learning and adaption as well as adoption of outside SE practices. In such organizational learning process, the SE implementation leader and the organizational management team go through cycles of trial, error and learning.

Chapter 6 - References

Box, G.E.P., Hunter, J.S. & Hunter W.G. *Statistics for Experimenters: Design, Innovation, and Discovery* (2nd ed.). Wiley, New York, 2005.

Chrysler, Ford, General Motors. *QS 9000 Quality Manuals*. West Thurrock: Garin Continuous Ltd. 1994.

Colley, R.H. *Defining advertising goals: For measured advertising results.* New York: The Association of National Advertisers, 1961,

Dale, B.G. & Shaw, P. Statistical process control: an examination of some common queries. *International Journal of Production Economics, 22*(1), (1991), 33-41.

De Koning, H. & De Mast, J. The CTQ Flowdown as a Conceptual Model of Project Objectives. *Quality Management Journal, 14*(2), (2007), 19-28.

De Mast, J., Does, R.J.M.M., De Koning, H. & Lokkerbol, J. *Lean Six Sigma for Service and Healthcare.* Alphen aan den Rijn (NL): Beaumont Quality Publications, 2012.

Does, R.J.M.M., Schippers, W.A.J. & Trip, A. A framework for implementation of statistical process control. *International Journal of Quality Science 2*, (1997), 181-198.

Does, R.J.M.M., Roes, K.C.B. & Trip, A. *Statistical Process Control in Industry*. Dordrecht (NL): Kluwer Academic, 1999.

Duncan, A.J. *Quality Control and Industrial Statistics* (4th ed.). Irwin, Homewood, 1974.

Ferdows, K. & De Meyer, A. Lasting improvements in manufacturing performance: in search of a new theory. *Journal of Operations Management 9*(2), (1990), 168–184.

Futrell, D. When quality is a matter of taste, use reliability indexes. *Quality Progress 28*(5), (1995), 81-86.

Gaafar, L.K. & Keats, J.B. Statistical process control: a guide for implementation. *International Journal of Quality and Reliability Management 9*(4), (1992), 9-20.

Goodman, M. *The Iceberg Model.* Hopkinton, MA: Innovation Associates Organizational Learning, 2002.

Hoerl, R.W. Enhancing the bottom-line impact of statistical methods. *ASQC Statistics Division Newsletter 15*(2), (1996), 6-18.

Jani, A. An experimental investigation of factors influencing perceived control over a failing IT project. *International Journal of Project Management 26*(7), (2008), 726–732.

Jensen, M.C. *Foundations of Organizational Strategy*. Harvard, Cambridge, MA 1998.

Jensen, W.A., Jones-Farmer, L.A., Champ, C.W. & Woodall, W.H. Effects of parameter estimation on control chart properties: a literature review. *Journal of Quality Technology 38*(4), (2006), 349-364.

Joiner, B.L. *Fourth Generation Management.* New York: McGraw-Hill, 1994.

Juran, J.M. *Leadership for Quality – An Executive Handbook*. New York: Free Press. 1989.

Kane, V.E. *Defect Prevention*. New York, NY: Dekker, 1989.

Lameijer, B.A., Antony, J., Borgman, H. & Linderman, K. Process improvement project failure: a systematic literature review and future research directions. *Submitted for Publication*, (2019a).

Lameijer, B.A., De Mast, J., & Does, R.J.M.M. Lean Six Sigma deployment and maturity models: a critical review. *Quality Management Journal 24*(4), (2017), 6-20.

Lameijer, B.A., Does, R.J.M.M., Antony, J. & Boer, H. Continuous improvement deployment models: A reconciliation and holistic metamodel. *Submitted for Publication* (2019b).

Lockyer, K.G., Oakland, J.S., Clive, H.D. & Followell, R.F. The barriers of statistical methods of quality control in UK manufacturing industry. *International Journal of Production Research 22*(4), (1984), 647-60.

Mann, R.S. Factors influencing the implementation success of TQM. *International Journal of Quality & Reliability Management 12*(1), (1995), 11-23.

McKinsey & Company. *The Science of Organizational Transformations*. McKinsey.com, (2015).

Montgomery, D.C. *Design and Analysis of Experiments* (8th ed.). New York, NY: Wiley, 2012.

Montgomery, D.C. *Introduction to Statistical Quality Control* (7th ed.). New York, NY: Wiley, 2013.

Roes, K.C.B. & Does, R.J.M.M. Shewhart-type charts in nonstandard situations, with discussion. *Technometrics 37*(1), (1995), 15-40.

Roes, K.C.B. & Dorr, D.C. Implementing statistical process control in service processes. *International Journal of Quality Science 2*(3), (1997), 149-166.

Sandorf, J.P. and Bassett, A.T. III. The OCAP: predetermined responses to out-of-control conditions. *Quality Progress 26*(3), (1993), 91-96.

Schilling, E.G. *Acceptance Sampling in Quality Control.* New York: Dekker, 1982.

Seddon, J. *Freedom from Command and Control: Rethinking Management for Lean Service.* Boca Raton, FL: Productivity Press, 2019.

Senge, P. M. et al. *The Fifth Discipline Fieldbook: Strategies and Tools for Building a Learning Organization.* New York: Crown Publishing, 2014.

Shewhart, W.A. *Economic Control of Quality of Manufactured Product*. New York: Van Nostrand, 2931,

Shingo, S. *Zero Quality Control: Source Inspection and the Poka-Yoke System*. Boca Raton, FL: Productivity Press, 1986.

Snee, R.D. & Hoerl, R.W. *Six Sigma Beyond the Factory Floor*. Upper Saddle River, NJ: Pearson, 2005.

Stamatis, D.H. *Failure Mode and Effect Analyses: FMEA from Theory to Execution*. Milwaukee, WI: ASQC Quality Press, 1995.

Suzaki, K. *The New Shop Floor Management*. New York: Free Press, 1993.

Taguchi, G. *Introduction to quality engineering: designing quality into products and processes*. White Plains, NY: UNIPUB/Kraus International, 1986.

Wadsworth, H.M., Stephens, K.S. & Godfrey, A.B.*Modern Methods for Quality Control and Improvement.* New York, NY: Wiley, 1986.

Case Studies in Statistical Engineering

The North America Competitive Product Laundry Initiative – A Case study on thorough understanding and quantitation of Laundry product performance through Statistical Engineering.

Authors: Sol Escobar, Cindy Rodenberg, Alex Varbanov (The Procter & Gamble Company)

0. Introduction

The Procter & Gamble Company (P&G) is one of the top 10 largest consumer packaged goods (CPG) companies and is considered one of the leading companies contributing to growth and innovation in an evolving market (reference BizVibe Top 10 Largest CPG Companies by Revenue in the World 2020 – CPG Industry Factsheet https://www.bizvibe.com/blog/largest-cpg-companies/). P&G's mission is to "provide branded products and services of superior quality and value that improve the lives of the world's consumers, now and for generations to come." (https://www.pg.com). Throughout its ten-category portfolio of products, P&G leverages deep consumer knowledge and category-changing innovation to identify consumer unmet needs and develop new technologies to address these needs. In order to stay competitive, it is essential for P&G to play at the leading edge of product superiority by providing consumers with high-performing options. This involves not only leading the market on key benefit spaces but also communicating those benefits to consumers around the world.

Case Study – North America Competitive Product Laundry Initiative

Fabric and Home Care (F&HC) is one of P&G's six industry-based Sector Business Units. The Laundry Cleaning and Care business is a highly competitive environment with multiple CPG players such as Henkel, Unilever, and numerous Private Label (store-brand) products. As such the Laundry Research & Development (R&D) division's Senior Leadership was interested in landscaping the product performance of North America (NA) P&G's laundry and fabric care products relative to the other competitive products (the NA Competitive Product Laundry Initiative) across the myriad of consumer benefit spaces: Stain Removal, Odor Removal, Whitening, Color Care, Freshness, Feel, and other attributes.

At first glance this may not seem like a complex problem. However, when we begin to think about the number of laundry and fabric care products, 1000+, and the number of consumer relevant benefit spaces, 10+, the time and cost (let alone the scheduling) for conducting the necessary tests to enable product comparisons, explodes dramatically. Additionally, the competitive laundry environment is very dynamic; including changes of 20 or more new product launches in a year, as well as, the potential for multiple blind formulation changes within pre-existing marketed products. Not only is it essential to

identify statistically appropriate designs and analysis for enabling reliable, unbiased product comparison, additionally, the scope of the initiative needs to be identified, the joint effort of multiple individuals needs to be coordinated, and a system that allows for ongoing updates and communication of results needs to be developed. The initiative meets the criteria of a large, complex, unstructured problem laid out in Hoerl and Snee (2017) that would benefit from the strategies of Statistical Engineering. In this paper, we discuss how each of the elements of Statistical Engineering, 1) Identify the high impact problems, 2) Providing structure, 3) Understanding context, 4) Develop Strategy, 5) Develop and execute tactics, and 6) Identify and deploy a final solution, were leveraged in initiative success. For simplicity, we will discuss each of these in a linear fashion, but as with all complex programs, iteration occurred throughout the program.

1. Identifying the High Impact Problem

As noted by Prof. Geoff Vining (Department of Statistics, Virginia Tech) during a P&G Statistical Engineering workshop (2019), rarely does senior management clearly or completely define the opportunity. Success requires systematic, systems thinking across the organization and therefore relies heavily on the scientists and researchers to scope out the full problem and potential opportunity, as well as, define the solution. As such, a critical first step is identification of the multi-disciplinary team.

P&G is comprised of Organization Units, some housed in the Business Units having deep category-specific knowledge of the business, customers, and/or needed technical capability, while others housed in organizations, such as Corporate Functions, with responsibilities to deliver and scale technical solutions across BUs. The multi-disciplinary team comprised of three individuals in the BU and two in Corporate Functions (identified in Table 1) ensured that the necessary expertise was included and enabled communication with the critical stakeholders.

Table 1: Team Members and Roles

Organization Unit	**Role**	**Responsibility**	**Link to Critical Stakeholders**
Fabric Care: Franchise	Product Researcher	Identifying: competitive product landscape, test methods for comparing product benefits	Senior Leadership to ensure that critical questions addressed and raise awareness of opportunity
Fabric Care: Franchise	Lab Researcher	Expert in test method execution and management of testing labs	Report to Product Research Leader for lab testing execution

Corporate Functions: Data & Modeling Sciences (D&MS)	Statistician	Experimental Study Design, Method Validation, Data Processing, Database Creation and Maintenance, Statistical Analysis, Results Interpretation	Senior D&MS Leadership for work accountability, work priority decisions, and availability of statistical resources
Corporate Functions: D&MS	Informaticist	Develop tools to access database of results	Project Statistician to ensure tools meet user requirements
Fabric Care/D&MS	Director Management	Communicate work/effort; Provide additional resources and funding as needed	Senior Leadership to ensure priority of initiative

Leveraging a small team of individuals with the necessary, unique domain expertise and a knowledgeable project leader, enabled efficient decision making and effective collaboration. Additionally, regular and direct communication between team members and with Senior Leadership ensured the work was going smoothly and in the direction of achieving the team objectives.

Additionally, to ensure the developed system met the needs of Senior Leadership, it was essential to understand the questions that were important to be addressed by this work. The project leader worked closely with Senior Management and the team, throughout the program, to identify the relevant questions, as well as, always keeping in consideration what was possible. Examples included:

- Status Quo
    - Where does my current product stand relative to competition and benefit spaces?
    - As the environment changes, is my product maintaining superiority or is the gap closing?
- Investment
    - How have other products changed and can they extract consumer and monetary value from it?
    - When is the optimal time to re-invest in my product formulation or innovate on new formulation?
- Consumer
    - How can I leverage my superiority gap to communicate the benefit I provide to consumers?

    - How can I confidently state my performance value in business-building claims for TV and Media advertisement?
- Expansion
    - Where can I launch a new product (performance whitespace) and what currently exists there?
    - What product performance trends exist and what is my product's potential in them?

Besides having the right team in place and defining the questions of interest, implementing the project strategy relied on other critical tactical decisions such as i) determining the scope, ii) establishing common product annotation hierarchy with unique codes, ii) using validated technical methods for data generations, iii) performing studies using statistical designs to get robust and reproducible data, iv) leveraging Network Meta-Analysis (NMA) as the statistical tool to integrate information across studies, and v) developing an online tool for easy access to results by different P&G users.

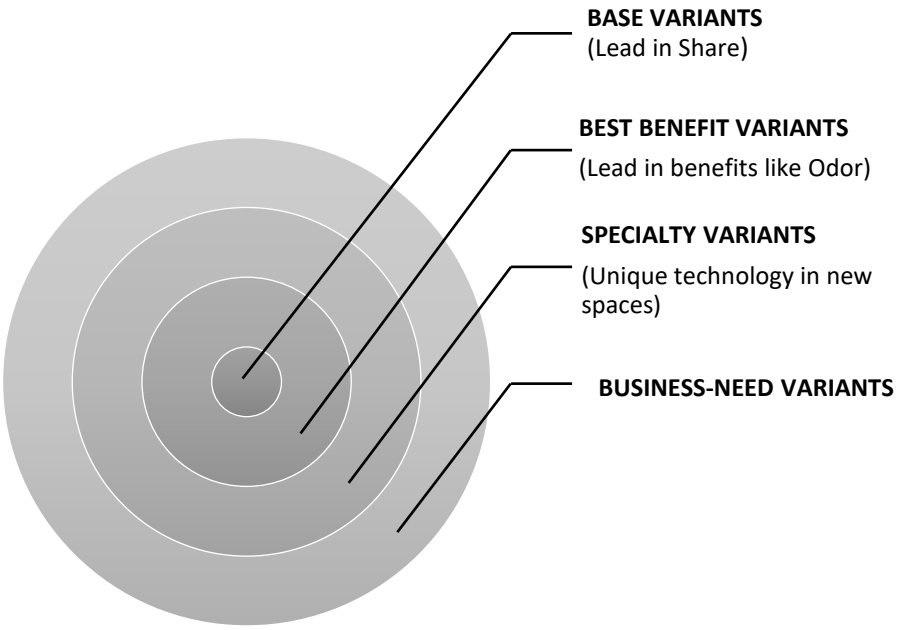
2. Provide Structure

Defining the scope and availability of actionable and robust data are two key areas necessary for determining the design strategy as well ensuring project success.

Scope of initiative: Products in the laundry space span large brand name products as well as small, niche category products. Identifying the scope of products and benefit spaces was essential. First, we started by ensuring we included the Market Share leaders (i.e. the highest selling variants) for each of the Market Core Brands which usually happen to be the Base or Basic variants for each (i.e. Base Tide, Base Arm & Hammer, Base Purex, etc.). This set the core baseline both from a sales and a consumer perspective, so we knew we were working with in a Relevant set. Second, we expanded the product line to the 5 Core Benefits in the Category (Stain Removal, Whiteness, Odor Removal, Color, Freshness). In doing so, we identified the best performing variants in each of the Core market laundry brands and assessed them for overall clean (stain removal) but also their respective benefit measure (i.e. Odor removal testing). At this point we were able to understand not only the 'cleaning core' of each Leading Base product, but also the cleaning and specific-benefit performance of the different portfolio variants from each of the core North American Laundry Brands. Third, we included Specialty Variants within the national leading brands – these included those small/niche spaces with potential to grow in the market (e.g. Wellness/Natural variants, Softening variants) but that might not be dominant leaders in the market

yet. Lastly, as needed by the Business, we included disruptors to the Category which includes anything from up-and-coming Private Label (Store Brand) products, to new forms to deliver detergency (e.g. Detergent sheets). Prioritizing our product inclusion criteria (Figure 1) enabled us to not only align R&D testing to the priorities of the overall Business, but it also enabled us to understand the relevant consumer market and to allow us to understand Competitive portfolio formulation and differentiation strategies.

Figure 1: Product Inclusion Criteria



Annotation Hierarchy: The team understood early on that without having a common annotation system across more than 150 products being tested across multiple types of laundry tests every year (stain, whiteness, color, etc.), there was no way to efficiently process the information. Linking the same product across all the different studies it was tested under was critical for the analysis. This was even more important to have for comparing Laundry products undergoing changes within the same year – often blind to consumers – where it was critical for P&G to understand and compare the different versions of that same marketed product. As a result, the team established a 6-layer hierarchy to annotate products. It included the Company, Brand, Form, Product Variant, Scent Variant, and Year for each individual product and/or product evolution and was cataloged in a Centralized Library. Figure 2, illustrates the annotation hierarchy and how corresponding products may differ across the fields.

Figure 2: 6-layer Annotation Hierarchy

With a unique code assigned to each product, it could then be traced across studies under a single identifier, and therefore allow for proper merging of the information across benefits in the deployment stage.

Validated Technical Methods: P&G relies strongly on using validated technical methods for assessing product performance. That requires understanding the major sources of data noise, as well as establishing reproducibility of results across different studies. Strict guidelines exist for a test method to be considered valid and reliable and usable for external credentialing. All the data collected by the Competitive Team followed that tactical rule, allowing the Company and the legal team to be confident in making product superiority claims, or in understanding how to improve current products. Additionally, having reproducible test methods was a cornerstone of the program enabling comparison of products run in separate tests.

3. Understand Context

Being part of a large organization like the P&G North America Fabric Care Business can sometimes bring some challenges to enabling a big undertaking and breakthrough initiative like this one. Some of these challenges include (1) working in highly focused (on current initiatives) organizations, (2) uncertainty in committing to future value initiatives that involve high initial activation energy/cost investment, and (3) pressure for delivering large and growing business needs while utilizing even less resources and time.

Due to growing priorities and threats in the market for such a large Business, an organization like North America Fabric Care can work more efficiently by having focused functions and projects (e.g. more focus on projects that deliver in the present, each function delivers according to their expertise/craft).

However, P&G has identified that there is also a large opportunity for innovation in having employees think beyond the bounds of a Function or Craft and looking past what the Business can see in the near term. This growing shift was a huge enabler for this Laundry Initiative. Allowing R&D and Corporate Functions create something that can bolster the Sales, Marketing, Finance and External Relations functions – versus only focusing these functions' capacities to their respective crafts (i.e. product making, models for product design) is a great example.

Second, undertaking the large task of measuring, categorizing, assessing, and modeling a category of more than 1000 products, which has a high-activation cost, time, and resource investment upfront but that could bring new value and capability to the Business is not an easy task – especially when the output and benefits have never been seen before. On top of this, trying to continue to drive more efficiency with less resources and time, would make it seem like an initiative like this one would carry more Risk than Reward to the Business. But this is where the innovative power of Statistics, R&D, and Core Business understanding helped.

Balancing (1) the knowledge of threats to the Business and their negative impact (i.e. how many competitors challenge P&G on a daily basis on advertisement and retail, the number of publishers and Ranking organizations that misinterpret the power of P&G products, etc.), (2) the limitations of R&D driven by resource constraints (i.e. not being able to deliver '#1' or Superiority claims for our products due to our inability to test 130+ products while also delivering other breakthrough innovation, lack of testing methodologies for new benefits like Odor Removal), and (3) identifying how the proposed initiative will deliver on a win-win scenario for the Business is key. This is where the partnership between R&D and Quantitative sciences to not only create new testing methods that deliver higher overall efficiency from the start, but to also identify innovative ways to reduce the amount of testing via powerful statistical methods was key. The compliment to this, which was perhaps even more critical in overcoming the unknown Risk of the investment, was demonstrating how it would directly answer to several of the significant threats the Business faced while creating a capability that – by default – also allowed several commercial areas to do their work more efficiently (i.e. new data-driven Sales content, new claims & communication vehicles for Marketing, new External Relations content for Influencers, and new models for Financial forecasting).

4. Develop Solution Strategy

Many of the components necessary for ensuring a successful solution strategy have been discussed. For example, the standardization of product annotation across all data sources (discussed in

Section 2) was an important strategic decision. In this section we lay out the statistical solution that was leveraged to efficiently integrate information across studies and enable all pairwise product comparisons.

Network Meta-Analysis (NMA): Once the team established validated technical methods to be included in the Competitive program, each study used a corresponding statistical design to account for potential variability sources. It relied on statistical concepts such as randomization, blocking, replications, and treatment balancing. In addition, controls were used in each test to establish critical connectivity between studies to be able to apply NMA.

NMA (Jones et al., 2011) was the main statistical tool to integrate the information across studies for a given laundry benefit (e.g., stain removal). It allowed us not only to compare products that were not placed in the same test, but also to be conservative in our treatment assessment by accounting for variability between tests. As such, many products were able to be compared directly every year, even if two products were never included in the same experiment or evaluation. Even more impressive, NMA enabled products to be compared across years, and therefore allowing us to historically track Laundry product performance evolution over time. As a result, the NMA approach gave the opportunity to establish an overall understanding of the laundry category for each measured benefit and became the basis for designing TV and Digital advertisement claims about Tide being the "#1 stain remover" and "#1 odor remover." That is only legally possible when Tide is assessed against a substantial majority of the share of marketed laundry products for that Laundry benefit. NMA therefore allowed the business team to be efficient in determining which claims to pursue.

5. Execution of Tactics

The team had to make many tactical decisions to execute the strategy of this complex NA Competitive Product Laundry Initiative. Discussing all of them is beyond the scope of this case study. We focus here only on the main four ones that have most significant contribution for the project success: i) use validated technical methods, ii) apply statistical experimental design principles, iii) use SAS software for all data processing, analysis, and activation, and iv) provide access to product results using an Online Access Tool. The first three are discussed briefly next while the last one is covered in Section 6.

Good business decisions rely on good product performance data. Making sure that the technical methods used to generate such data are validated is key to ensure good data reproducibility and sensitivity. Each method that we use in the NA Competitive Laundry Product Initiative goes thru a formal method validation study to understand different variability source and quantify data reproducibility across days and operators. After that, the follow up studies used for product ranking data are based on statistical

experimental design and using principles such randomization, blocking, replication to ensure robustness of statistical results and hence business decisions. We use controls in each study to be able to not only monitor quality control over time but also allow connectivity between studies for NMA.

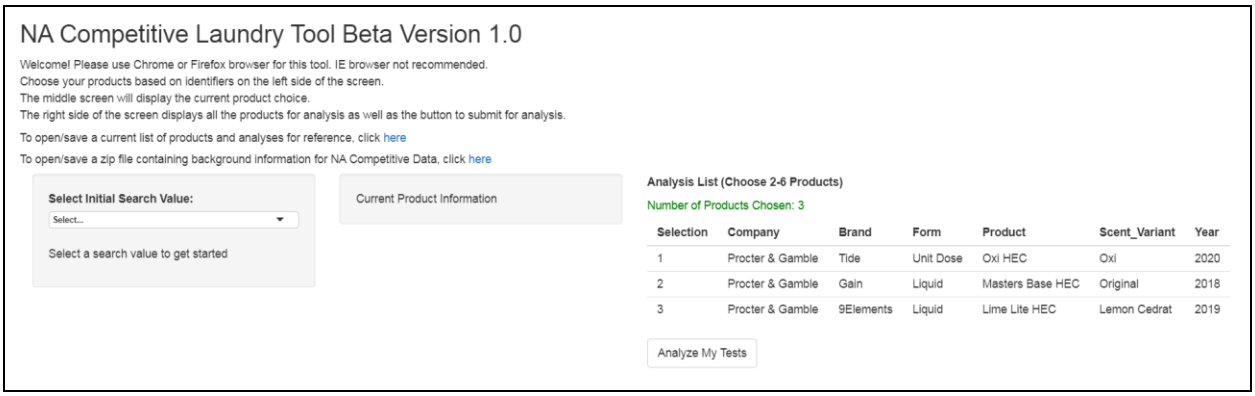
The generation of the product ranking for a given benefit requires a significant amount of data processing and using mixed effect models for statistical analysis. We decided to use SAS software for these tasks because of its ability to handle both. Alternative software (e.g. R) was considered as well in the beginning but SAS was better tool fit given the personnel resources supporting this project. In addition, it was easy to set an online access tool to the results following a successful route to that given by other tool examples.

6. Deployment of Final Solution

Online Access Tool: The last (but not least) critical tactical implementation step was to develop an online tool for easy access of the NMA results across benefits. To create it, the team engaged in a collaboration with the Informatics group, focusing on building a user-friendly interface with easy access to results. By enabling the user to easily specify the product(s) in question (Figure 3), the tool would output all corresponding performance data available across all benefits tested in a consistent structured format. The Tool element therefore became key as unexpected and urgent business questions requiring immediate comparative performance assessments could now be answered within seconds. In addition, even users with low expertise or familiarity with the Laundry category or competitive portfolio structures could access the multi-benefit data and apply the respective learnings to their business cases – which further elevated the level of expertise and capability across the Research & Development organization.

Because constant performance testing and data generation is an on-going process and the team desired minimal-to-no disruption to the business and users, the data tables used by the online tool were updated in the background by the statistician allowing for continuity to operations. In addition, given the team's desire to continuously improve, the deployed solution will be routinely assessed for upgrades such as adding product images and improving user interface.

Figure 3: Online Webtool Interface



7. **Conclusion**

The NA Competitive Product Laundry Initiative is complex covering multiple benefits, studies, and products. There are other everyday tactical decisions that the team makes to make sure objectives are met and information feeds other teams for best business decisions. However, the elements discussed above were *the most influential ones for the program success*. The final solution of the P&G Competitive Program is the ability to make informed data-driven business decisions about how to make superior products and/or allow for cost savings without sacrificing product performance. That was enabled by the tactical steps outlined in the previous section including the development of an online tool for easy access.

However, it also includes continuous communication between the project leader and P&G sales and other product research and development teams. This is critical for outlining correct use of the results and summarizing the multidimensional information in easy to comprehend way. It also covers the need to address follow up questions quickly and efficiently. As discussed above, the information is consistently updated each year as new products enter the laundry market and there is a need to understand their performance. The P&G Competitive program continues to evolve, enabling coverage of new benefits (e.g., measure scent liking of laundry touchpoints) and laundry categories (e.g., Liquid Fabric Enhancers).

The success of the NA Competitive Product Laundry Initiative - an integrative, fast-responding, business-building system of Laundry performance data would not be possible without use of comprehensive decisions to enable rapid testing, data collection, and analysis as well as deployment of information for driving business decisions. This case study illustrates the importance of utilizing all elements of Statistical Engineering when faced with a large, unstructured, complex problem. The work was done by a small efficient team with diverse complementary skills. It provided structure in solving the problem by defining the right scope of the initiative and product annotation hierarchy. Critical strategic

decisions (e.g., using validated technical methods or NMA) ensured success in the Statistical Engineering process. Using an online access tool helped with the deployment of the final solution. This initiative created significant business value for the company which made it a successful example for Statistical Engineering case study.

References

Hoerl, R.W. and Snee, R.D. (2017), "Statistical Engineering: An Idea Whose Time Has Come?" The American Statistician, 71(3), 209-219.

Jones, B., Roger, J., Lane, P.W., Lawton, A., Fletcher, C., Cappelleri, J.C., Tate, H. and Moneuse, P. (2011). "Statistical Approaches for Conducting Network Meta-Analysis in Drug Development", Pharmaceutical Statistics, 10, 523-531.

Journal of Quality Technology
A Quarterly Journal of Methods, Applications and Related Topics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ujqt20

Multilevel process monitoring: A case study to predict student success or failure

Leo C. E. Huberts, Marit Schoonhoven & Ronald J. M. M. Does

To cite this article: Leo C. E. Huberts, Marit Schoonhoven & Ronald J. M. M. Does (2020): Multilevel process monitoring: A case study to predict student success or failure, Journal of Quality Technology, DOI: 10.1080/00224065.2020.1828008

To link to this article: https://doi.org/10.1080/00224065.2020.1828008

JOURNAL OF QUALITY TECHNOLOGY
https://doi.org/10.1080/00224065.2020.1828008

CASE REPORT

OPEN ACCESS Check for updates

Multilevel process monitoring: A case study to predict student success or failure

Leo C. E. Huberts, Marit Schoonhoven, and Ronald J. M. M. Does

Department of Operations Management, Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

In this case study, we demonstrate the use of multilevel process monitoring in quality control. Using high school data, we answer three research questions related to high school student progress during an academic year. The questions are (1) What determines student performance? (2) How can statistical process monitoring be used in monitoring student progress? (3) What method can be used for predictive monitoring of student results? To answer these questions, we worked together with a Dutch high school and combined hierarchical Bayesian modeling with statistical and predictive monitoring procedures. The results give a clear blueprint for student progress monitoring.

KEYWORDS

hierarchical Bayesian; multilevel; predictive monitoring; statistical process monitoring; student performance

1. Motivation

"Early Warning Indicator Reports were invaluable to the success of our school" (high school principal, a quote from the Strategic Data Project Report by Becker et al. (2014)). These early warning indicator reports monitor students throughout their school career and warn teachers and staff of students with high dropout risks. According to Romero and Ventura (2019), such early identification of vulnerable students who are prone to fail or drop their courses is crucial for the success of any learning method. Also, monitoring allows for the identification of students who are insufficiently challenged and will benefit from more stimulating classroom material.

Navigating the large body of literature in statistical process monitoring, predictive monitoring and educational data mining is a daunting task when looking for answers as to what metrics should be monitored and which methods should be implemented.

Multilevel modeling is often a good method in educational settings and can be used for predictive monitoring in quality control. In this article, we demonstrate such a procedure and aim to guide researchers and practitioners in monitoring student performance, specifically in a high school setting. To achieve this, we work closely with a Dutch high school to answer the following questions 1) What determines student performance? 2) How can statistical process monitoring be used in monitoring student progress? 3) What method can be used for predictive monitoring of student results?

1.1. Statistical process monitoring

Statistical process monitoring (SPM) provides techniques to monitor a process real time. As the amount and complexity of available data are increasing, there is a need for SPM methods that utilize more of the inherent structure of the data. This need has driven SPM to evolve in recent years from univariate methods monitoring a single quality indicator, to monitoring methods for complex multivariate processes. A method that is used for multivariate processes are profile monitoring. Profile monitoring checks the stability of the modeled relationship between a response variable and one or more explanatory variables over time. Often profile monitoring uses regression control charts which were first introduced by Mandel (1969). The current body of regression control charting literature almost exclusively handles the monitoring of linear profiles using classical regression models. Weese et al. (2016) noted that large data sets often contain complex relationships and patterns over time, such as hierarchical structures and autocorrelation.

CONTACT Leo C. E. Huberts L.c.e.huberts@uva.nl Department of Operations Management, Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands.

Supplemental data for this article is available online at https://doi.org/10.1080/00224065.2020.1828008

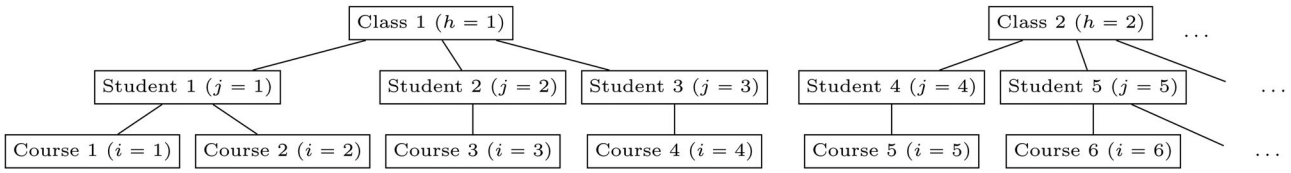


Figure 1. The hierarchical structure of the case study data with classes as the top level. Students within these classes are the middle level and courses followed by these students form the lower level.

The case study presented in this paper contains complex relationships and patterns, notably the hierarchical structure of courses, students and classes (see Figure 1). State-of-the-art multivariate control charting based on linear regression models ignores this structure. However, incorporating hierarchical structures into the models can improve the reliability of a monitoring system. Therefore, we will develop a control chart that can signal at three levels, the class, student and course level. Also, Woodall and Montgomery (2014) gave an overview of current directions in SPM and highlighted profile monitoring with multiple profiles per group as a topic for further research.

The advantage of using a hierarchical model is an improved estimation of process variability; according to Gelman (2006), hierarchical modeling is almost always an improvement compared to classical regression. The reason is that a hierarchical model includes the effects of both observed and unobserved variables, where unobserved variables are not explicitly measured but inherent to the group. Another advantage over classical regression is that a multilevel model provides a way to monitor new groups since the model generates some prior beliefs upon which to base the distribution and the prediction for the new groups. Furthermore, in contrast with classical regression, multilevel modeling is capable of prediction for groups with a small number of observations.

Multilevel models have been used in agricultural and educational applications for decades (Henderson et al. 1959; Aitkin and Longford 1986; Bock 1989; Aaronson 1998; Sellström and Bremberg 2006). Today, hierarchical models are used in spatial data modeling (Banerjee, Carlin, and Gelfand 2014), extreme value modeling (Sang and Gelfand 2009), quantum mechanics (Berendsen 2007) and even in the modeling of intimacy in marriage (Laurenceau, Barrett, and Rovine 2005). However, to the best of our knowledge, multilevel modeling has not found its way to SPM. Schirru, Pampuri, and De Nicolao (2010) modeled multistream processes in semiconductor manufacturing using a multilevel model, but it is only applicable to two levels. Qiu, Zou, and Wang (2010) considered nonparametric profile monitoring using mixed-effects modeling, although they did not consider hierarchical modeling.

This article will explore process monitoring for a school data set that contains the grades of students in different groups over time. The school is interested in monitoring deviations in student results from what is given by the model, which is a form of profile monitoring. Therefore, we will investigate SPM based on hierarchical Bayesian models. In the next section, we will discuss the use of a hierarchical model to predict outlying results on the student level.

1.2. Predictive monitoring

Becker et al. (2014) emphasized the need for actionable predictive analytics in high schools to keep students on track toward graduation and better prepare them for college and career success. The report discussed three examples of early warning indicator systems that help school teachers and management with early identification of students with a lower probability of passing, based on logistic regressions of student grade and attendance information.

Early prediction of learning performance has gained more traction in the literature, as showcased by a recent special issue of IEEE Transactions on learning technologies. Together with monitoring big and complex data, predictive monitoring is recently being considered in quality technology literature (for example Kang et al. 2018; Wang et al. 2019). Although our case study focuses on the use of predictive monitoring to improve the quality of education, the presented methods can be used in any setting where clear hierarchical data structures exist. Baghdadi et al. (2019) stated that the ability to estimate when the performance will deteriorate and what type of intervention optimizes recovery can improve the quality and productivity and reduce risk concerning worker fatigue. Our case study offers a very similar approach to improve the quality and productivity of high school education by monitoring student performance.

The hierarchical model will thus be applied in two ways. First, control charting is applied based on the

Table 1. Summary of determinants of student performance according to the literature and modeling approach.

Determinant	Effect on performance		
	Student level	Class level	Modeling approach
SES		+	Explanatory variable
Disabilities	–		Explanatory variable
Language	+/–		Explanatory variable
Non-native	+/–	–	Explanatory variable
Student effort	+	+	Student unobserved heterogeneity
Peer associations	+/–	+/–	Student/course unobserved heterogeneity
Parent involvement	+		Student unobserved heterogeneity
School climate	+/–	+/–	Course unobserved heterogeneity
Intelligence	+		Explanatory variable, student unobserved heterogeneity
Grades	+		Time varying explanatory/dependent variable
Absences	–	–	Time varying explanatory variable

multilevel model. Second, the multilevel model is used for predicting results on the student level. This results in a hierarchical early warning indicator system that can be applied in schools for predictive monitoring of student outcomes.

The outline of this paper is as follows. The next section describes the relevant educational literature, the practical problem we aim to solve and the data that was available. The hierarchical model and its performance are discussed in the section after this, followed by a section that investigates student performance monitoring. The last section summarizes the results.

2. Problem description

In this section, we describe related student performance literature, the goal of the method to be developed and the data set including the predictor variables.

2.1. Student performance literature

This section will shortly discuss a selection of determinants of student performance, whose selection has been based on a literature study. The determinants, their expected effects on performance and their modeling approach are summarized in Table 1. The important variables will be used in the modeling approaches of later sections. The "unobserved" variables represent variables that were not available in this study, but the hierarchical modeling specification incorporates many of these "unobserved differences" between students and students within courses.

Nichols (2003) found a significant relationship between poor performance at the beginning of students' educational careers and later on. Furthermore, students who struggle academically had increased school absences and students from lower-income families showed a higher probability of poor results. This suggests an important role for family income, absences and temporal effects in predicting individual high school performance.

Socioeconomic status (SES) has long been argued to significantly affect school performance, although the importance varies greatly among different analyses. Geiser and Santelices (2007) argued omission of socioeconomic background factors can lead to significant overestimation of the predictive power of academic variables, that are strongly correlated with socioeconomic advantage. They based this assumption on a study by Rothstein (2004), which argued the exclusion of student background characteristics from prediction models inflates college admission tests' apparent validity by over 150 percent.

Disabilities can be a determinant of student performance. Dyslexic children fail to achieve school grades at a level that is commensurate with their intelligence (Karande and Kulkarni 2005). Although they might not be directly linked to learning, disabilities like asthma, epilepsy, and autism can indirectly influence academic performance. Autistic children can face a lot of problems in school as their core features impair learning. Furthermore, medical problems like visual impairment, hearing impairment, malnutrition, and low birth weight can cause difficulties in school.

The language that children speak at home can influence their academic abilities both positively (Buriel et al. 1998) and negatively (Kennedy and Park 1994). Collier (1995) found that immigrants and language minority students need 4–12 years of second language development for the most advantaged students to reach deep academic proficiency and compete successfully with native speakers. It has been suggested that the presence of non-native speakers in schools harms the performance of native speakers, but this has been refuted by Geay, McNally, and Telhaj (2013). In contrast, children who interpret for their immigrant parents; "language brokers," often perform better academically (Buriel et al. 1998).

Some variables remain unobserved but can be incorporated in models by allowing for unobserved

heterogeneity. One is student effort, which is characterized by the level of school attachment, involvement, and commitment displayed by the student (Stewart 2008). Also, peer influence, i.e. the associations between high school students, matter a great deal to individual academic achievement and development (Nichols and White 2001). Besides, parent involvement is likely to influence academic achievement. Sui-Chu and Willms (1996) found that the most important dimension of parent involvement toward academic achievement is home discussion. They suggested facilitating home discussion by providing concrete information to the parents about parenting styles, teaching methods, and school curricula. Finally, school climate (a.o. Stewart 2008) and intelligence (Rohde and Thompson 2007; Laidra, Pullmann, and Allik 2007; Parker et al. 2006) are important for academic achievement.

Parent involvement, disciplinary climate, and individual intelligence are usually quite difficult to measure. This study aims to incorporate them nonetheless. Parent involvement is incorporated mostly in student unobserved heterogeneity. Limited observed information on the parents is included in the predictive model (i.e. education level and SES). Disciplinary climate and class disruptions are mostly covered by including absences that equate to dismissals from class and within unobserved course differences. Individual intelligence is approximated using primary school test scores.

Next, some time-varying variables are important. The first variable is the grade. For each course, specific tests are taken with varying weights. Anytime during the year, these tests determine a current weighted average grade for each student and course. The resulting end-of-year grade is the most important student performance indicator. Also, absences are important as attending class helps students understand the material and motivates their participation (Rothman 2001). The variables test grades and absences are generated over time. Finally, temporal effects on student performance encompass both inter-year changes and intra-year changes. Students will change the allocation of their effort and time according to their current average grade, their average grade for other courses, seasonal effects, within school changes and external factors. Ideally, modeling will allow for student and course-specific effects to vary over time. The next section will describe the Dutch high school system.

2.2. The Dutch high school system

The Dutch school system in general consists of eight years of primary school, followed by four, five or six years of high school. There is one level of primary school, but there are multiple levels of high school. Two criteria have been used in recent years to determine the level of high school a child is allowed to go to. Firstly, there is the teacher's advice. The teacher advises the level that fits the child in the final year of primary school. This advice is based on the performance of the child in a specific primary school.

Secondly, the National Institute for Test Development (in Dutch: Centraal Instituut voor Toets Ontwikkeling, abbreviated by CITO) test is a test that is developed by the CITO organization and is scientifically designed to test a child's academic abilities. It was initiated in the Netherlands by the famous psychologist professor A.D. de Groot in 1966 and every school is required to conduct the CITO or a similar test at the end of primary school as of 2014.

To pass any specific year of high school, conditions set by the school have to be met. These conditions usually consist of requirements on the end-of-year average grades for all the student's courses. The grades in most Dutch high schools are on a scale of 1 to 10. The end-of-year grades are usually rounded, and a course is failed or "insufficient" if the rounded grade is below 6. The amount of allowed "failpoints," i.e. the total points below six, can then be restricted. A school might, for example, have a student repeat the current year if he or she scores more than two failpoints, which could be a student with a grade of three for a single course, or a four and a five or three fives at the end of the year. The restrictions are not limited to the number of failpoints. There can be requirements on the total average grade and certain subtleties emerge once the students start splitting up into high school profiles, where different students do a different set of courses from their fourth year on. These school profiles can have special requirements, with usually more importance assigned to the profile courses.

When implementing a predictive monitoring scheme in a school, the specific rules a school employs define the passing probability that is estimated. When for example a student is failing a profile course, this can lead to failing the year directly. If the same student would obtain the same grade for a different course, this would not necessarily mean failing the year. Therefore, different courses have different levels of importance to the probability of success for individual students. The school that has kindly provided the data described in the next section has different passing conditions for each year. Although the implementation at the school incorporates all conditions, the predictive analyses in this paper reflect a

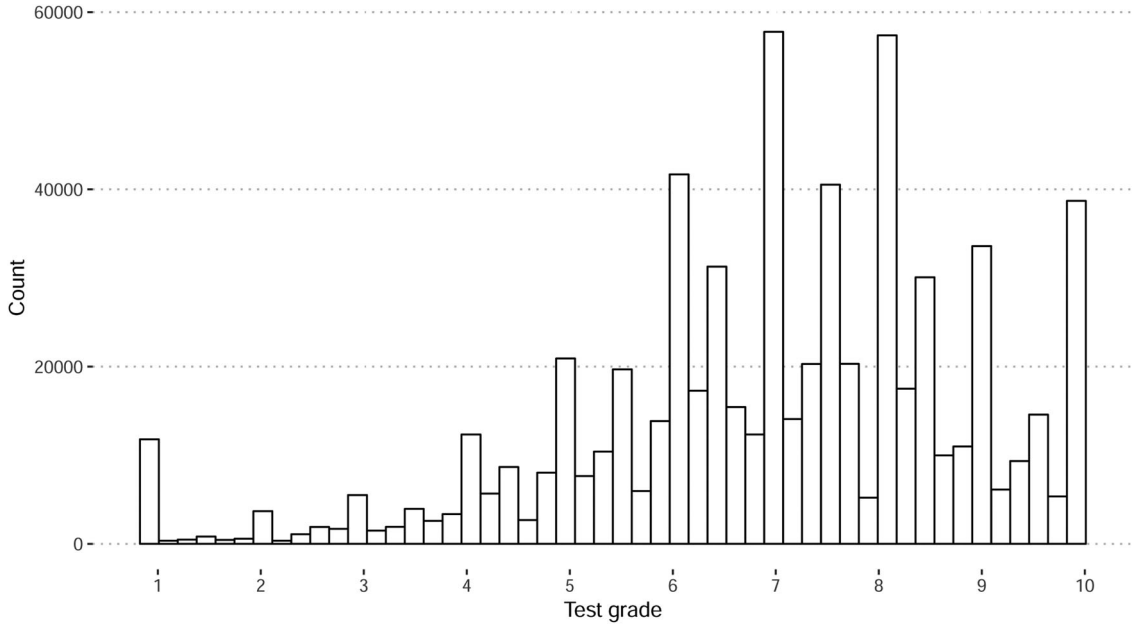


Figure 2. Histogram of the individual test grades in the data.

simplified version to demonstrate the detective capabilities of the methods.

2.3. Data set

A large, detailed data set was provided by a Dutch high school. In total there are eight years of data available, comprising of 36 different subjects followed by over 1,700 unique students (about 51% girls) and 711,653 individual tests. The students were born in 38 different countries, speak 18 different languages and were taught by 110 different teachers. Out of the unique students, 326 had some kind of disability while at school, 162 had a non-Dutch nationality and 51 students had a serious language barrier. The number of students with parents who have attended university or higher-level academics is 261 and 86% of students were residents of the large city that the school is located in during their time at the Dutch high school.

To incorporate socioeconomic status (SES) in this analysis, nation-wide social status data provided by the Dutch government was used. The relative SES score of a student using a country-wide ranking of his or her postal code was added to the data set.

Learning disabilities that have been confirmed by the school are included in the data set. The most common learning disabilities in the data are Attention-Deficit/Hyperactivity Disorder (ADHD) and dyslexia.

The data used in this paper contains grades that are on a 1-10 scale. Although easy to interpret, there arise some difficulties when using these grades for modeling. First, as Figure 2 shows, there are peaks at integer grades and grades on a.5 scale. This is due to teachers grading on an integer or.5 point scale instead of using continuous grades. This becomes less of a problem with average grades, as they are eventually rounded but fairly continuous during the year.

Second, when predicting the precise end-of-year grade, grades below 1 or above 10 should be impossible. However, both grades should have some positive probability, as some students do achieve average grades of 10 for specific courses during a year.

The following section describes the selected predictor variables in the data.

2.4. Determinants of student performance

We have discussed some of the literature on determinants of high school performance in Section 2.1. This section investigates these variables in the data.

The raw values for the most important categorical variables in the data are plotted in Figure 3. The first pair of boxplots in Figure 3 shows that girls seem to outperform boys in terms of final grades, which is consistent with the literature in different settings (see Rahafar et al. 2016; Deary et al. 2007; Battin-Pearson et al. 2000 for examples of gender gap findings in academic achievement). The second pair of boxplots in Figure 3 indicates that students with a disability achieve lower end-of-year grades, consistent with the findings of Karande and Kulkarni (2005). Children of highly educated parents seem to perform slightly

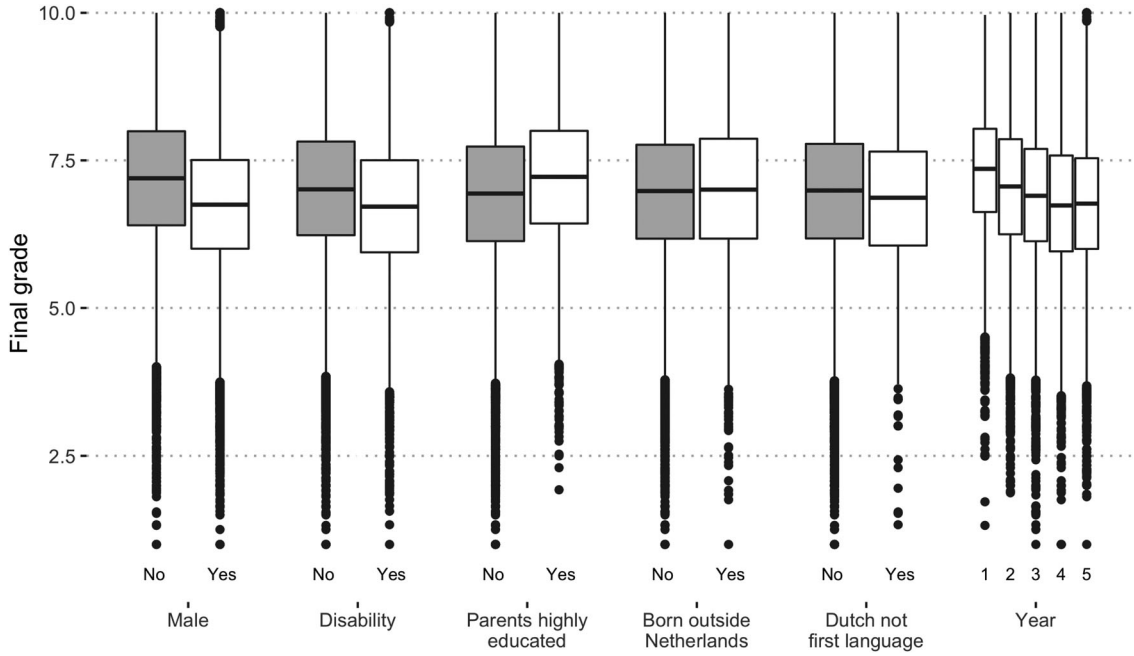


Figure 3. Boxplots of the final grades for the most important categorical predictor variables.

better at this school in terms of final grades, as depicted in the third pair of boxplots in Figure 3.

In line with Buriel et al. (1998), children born outside of the Netherlands do not underperform as shown by the fourth pair of boxplots in Figure 3. Students with a different native language do achieve slightly lower grades in the data, supporting conclusions by Collier (1995) and Kennedy and Park (1994). The end-of-year grades are lower toward the end of high school, as indicated in Figure 3.

Figure 4 shows the two most important numerical independent variables plotted against the final grades. The CITO score has a positive correlation with grades as shown by the positive linear trend in Figure 4a. This makes sense, as the CITO test is designed as a predictor of individual intelligence. Furthermore, in line with Rothman (2001), more absences mean lower final grades in the data, as indicated by the negative linear trend in Figure 4b.

3. Hierarchical model

The objective is to monitor student progress during the school year, where the school's main interest lies in signaling "exceptional" students. Exceptional students can be both underperforming and overperforming students. In this section, we introduce a three-level hierarchical model for student grades and compare its performance to simpler models in monitoring student performance.

3.1. The model

Throughout the year, students take tests for every course $i = 1, \ldots, n_0$. The grades for these tests are defined as $g_{ki} \in [1, 10]$ with $k = 1, .., K_i$, where K_i is the number of tests taken in course i. As these grades are obtained for individual tests, we have a set of cumulative weighted average grades $y_{i,j[i],h[j[i]]}$ for course i, student j and class h. For readability we drop subscripts j and h. The individual test results g_{ki} and the weights of the tests w_{ki} determine the average grade $y_i = \frac{\sum_{k=1}^{K_i} w_{ki} g_{ki}}{\sum_{k=1}^{K_i} w_{ki}}$, with $y_i \in [1, 10]$.

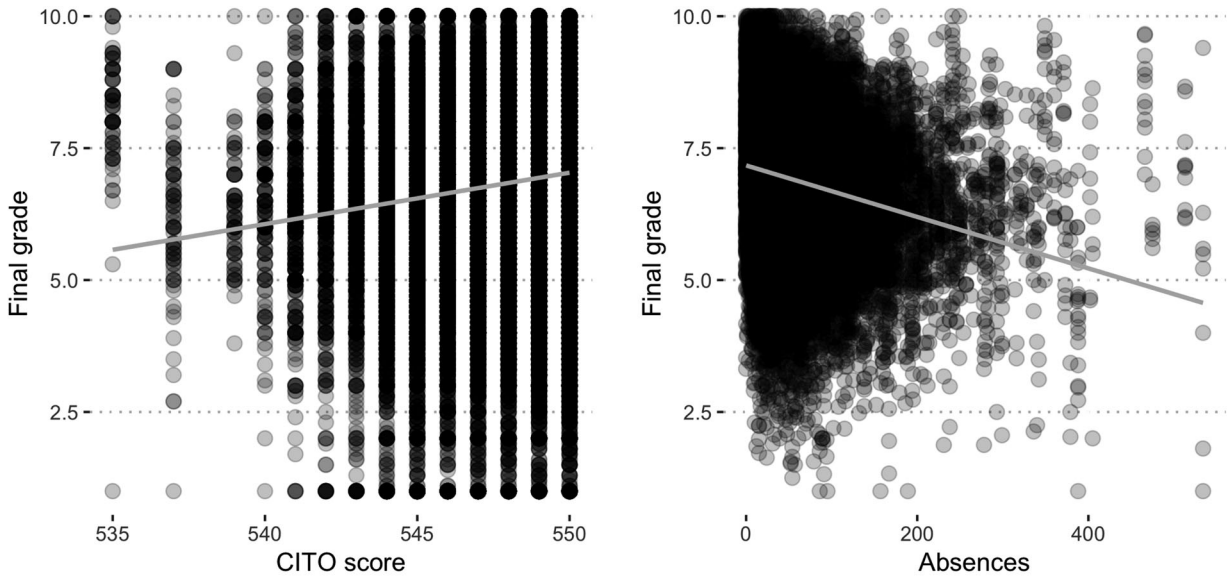
We consider a hierarchical model with three levels and use the index $i(i = 1, 2, \ldots, n_0)$ to denote the individual course level, $j(j = 1, 2, \ldots, n_1)$ to denote the individual student level and $h(h = 1, 2, \ldots, n_2)$ for the class level (see Figure 1). We have p_0 predictors for the course level, p_1 for the student level and p_2 for the class level. We define row vectors $\boldsymbol{X}_{\boldsymbol{i}}^{(L_0)}, \boldsymbol{X}_{\boldsymbol{j}}^{(L_1)}$ and $\boldsymbol{X}_{\boldsymbol{h}}^{(L_2)}$, which consist of the intercept and predictor values for the course, student and class levels respectively.

We model cumulative weighted average grade y_i for course i as

$$y_i \sim N(\boldsymbol{X}_{\boldsymbol{i}}^{(L_0)} \boldsymbol{\beta}_{\boldsymbol{j[i]}}^{(L_0)}, \sigma^2), \text{ for } i = 1, \ldots, n_0 \text{ (Course level)},$$

where the student levels are modeled as

$$\boldsymbol{\beta}_{\boldsymbol{j}}^{(L_0)} \sim N(\boldsymbol{\beta}_{\boldsymbol{h[j]}}^{(L_1)} \boldsymbol{X}_{\boldsymbol{j}}^{(L_1)'}, \boldsymbol{\Sigma}^{(L_1)}), \text{ for } j = 1, \ldots, n_1 \text{ (Student level)},$$



(a) Final grades versus CITO score (b) Final grades versus total absences

Figure 4. Scatterplots of the final grades and most important numerical variables with a linear trend-line.

and the class levels are specified by

$$vec(\boldsymbol{\beta}_{\boldsymbol{h}}^{(L_1)}) \sim N(\boldsymbol{\beta}^{(L_2)}\boldsymbol{X}_{\boldsymbol{h}}^{(L_2)'}, \boldsymbol{\Sigma}^{(L_2)}), \text{ for } h = 1, \ldots, n_2 \text{ (Class level)},$$

where $\boldsymbol{X}_{\boldsymbol{i}}^{(L_0)}$ is a $1 \times (p_0 + 1)$ row vector of subject specific variables such as course content and level; $\boldsymbol{\beta}_{\boldsymbol{j[i]}}^{(L_0)}$ is a $(p_0 + 1) \times 1$ vector of parameters for student j that follows course i; σ^2 is the variance for the course level; $\boldsymbol{\beta}_{\boldsymbol{h[j]}}^{(L_1)}$ is a $(p_0 + 1) \times (p_1 + 1)$ parameter matrix determined by the class h that student j is in; $\boldsymbol{X}_{\boldsymbol{j}}^{(L_1)}$ is a $1 \times (p_1 + 1)$ row vector of student specific variables such as age, absences and IQ; $\boldsymbol{\Sigma}^{(L_1)}$ is the covariance matrix for parameters $\boldsymbol{\beta}_{\boldsymbol{j}}^{(L_0)}$; $vec(\boldsymbol{\beta}_{\boldsymbol{h}}^{(L_1)})$ is the vectorized version of $\boldsymbol{\beta}_{\boldsymbol{h}}^{(L_1)}$ with dimensions $(p_0 + 1)(p_1 + 1) \times 1$; $\boldsymbol{\beta}^{(L_2)}$ is a $(p_0 + 1)(p_1 + 1) \times (p_2 + 1)$ parameter matrix at the class level; $\boldsymbol{X}_{\boldsymbol{h}}^{(L_2)}$ is a $1 \times (p_2 + 1)$ row vector of class specific variables such as class size; and $\boldsymbol{\Sigma}^{(L_2)}$ is the covariance matrix for parameters $\boldsymbol{\beta}_{\boldsymbol{h}}^{(L_1)}$.

3.2. Estimation

The parameters of a multilevel model can be estimated using, among other methods, maximum likelihood, generalized least squares and Bayesian theory (Hox, Moerbeek, and Van de Schoot 2017). A discussion of Bayesian and likelihood-based techniques for multilevel models is given by Browne and Draper (2006). These authors show that Bayesian estimation often provides an improvement over likelihood methods in terms of both point and interval estimates as well as the posterior distributions for the parameters. We use Bayesian estimation to estimate the parameters in this article.

The full parameter space $\{\boldsymbol{\beta}^{(L_0)}, \sigma^2, \boldsymbol{\beta}^{(L_1)}, \boldsymbol{\Sigma}^{(L_1),\beta^{(L_2)},\Sigma^{(L_2)}}\}$, where $\boldsymbol{\beta}^{(L_0)}$ and $\boldsymbol{\beta}^{(L_1)}$ are constructed by stacking the parameter matrices $\boldsymbol{\beta}_{\boldsymbol{j}}^{(L_0)}$ and $\boldsymbol{\beta}_{\boldsymbol{h}}^{(L_1)}$ for all groups j and h respectively, can be estimated based on data that are considered representative, i.e. in control. To estimate the parameters, we use the Bayesian method applying Markov Chain Monte Carlo (MCMC) methods which use the Gibbs sampling procedure. These methods are described in the appendix and are applied using the rJAGS package to link to JAGS (Plummer 2018).

As the number of parameters increases quickly with added group levels, estimation time increases greatly as well. Thus when defining a multilevel model, there is a tradeoff between added precision and the additional estimation time for a group level. In a two-level model, the number of parameters we need to estimate is 1 for σ^2, $(p_0 + 1)(p_1 + 1)$ for $\boldsymbol{\beta}^{(L_1)}$ and $\frac{1}{2}(p_0 + 1)(p_0 + 2)$ for $\boldsymbol{\Sigma}^{(L_1)}$ ($\boldsymbol{\beta}^{(L_0)}$ is constructed using the estimates for $\boldsymbol{\beta}^{(L_1)}$). For the three-level model this increases, with 1 for σ^2, $\frac{1}{2}(p_0 + 1)(p_0 + 2)$ for $\boldsymbol{\Sigma}^{(L_1)}$, $(p_0 + 1)(p_1 + 1)(p_2 + 1)$ for $\boldsymbol{\beta}^{(L_2)}$ and $\frac{1}{2}(p_0 + 1)(p_1 + 1)((p_0 + 1)(p_1 + 1) + 1)$ for $\boldsymbol{\Sigma}^{(L_2)}$ ($\boldsymbol{\beta}^{(L_0)}$ and $\boldsymbol{\beta}^{(L_1)}$ are constructed using the estimates for $\boldsymbol{\beta}^{(L_2)}$). For example, if there are three parameters per level, the number of parameters is 27 for a two-level model and 211 for a three-level model.

Table 2. RMSE and NN results for the predictions of the 2014/2015 end-of-year grades of 268 students using the average grade (y_i), the simple regression ($\hat{y}_{sr}$) and the hierarchical specification ($\hat{y}_H$).

Time	RMSE			NN		
t	y_i	$\hat{y}_{sr}$	$\hat{y}_H$	y_i	$\hat{y}_{sr}$	$\hat{y}_H$
0	–	1.152	0.860	–	0.802	0.902
0.1	1.526	1.069	0.835	0.699	0.830	0.908
0.3	1.037	0.831	0.741	0.856	0.917	0.940
0.5	0.773	0.668	0.648	0.931	0.956	0.957
0.7	0.511	0.478	0.474	0.980	0.983	0.984

Table 3. Confusion matrix of the predictions for the 2014/2015 end-of-year grades of 268 students based on the simple linear regression model at $t = 0$.

		Actual grades							
		3	4	5	6	7	8	9	10
Predicted	6	0	1	0	0	0	0	2	0
	7	9	53	208	722	962	747	283	33
	8	0	6	20	134	255	252	140	12

After applying the estimation procedure as described in the appendix, we obtain the estimations for the parameters in the three-level model, which we denote by $\{\hat{\boldsymbol{\beta}}^{(L_0)}, \hat{\sigma}^2, \hat{\boldsymbol{\beta}}^{(L_1)}, \hat{\Sigma}^{(L_1)}, \hat{\boldsymbol{\beta}}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$. Later on we can use this three-level model for monitoring the relationships given by the model as well as for predicting results.

3.3. Results

In this section, we consider the accuracy of the end-of-year average grade estimates for $N = 3,839$ courses and 268 students during the school year 2014/2015. This subset consists of the first-, second- and third-year students. In the fourth year students choose a profile, which changes the class compositions. The five school years from 2009 to 2014 are used to estimate the parameters.

As benchmarks, we consider using the weighted average grade (y_i) and a simple one-level linear regression model ($\hat{y}_{sr}$) to predict. The one-level linear regression fits $y_i = \boldsymbol{X_i}\boldsymbol{\beta} + \varepsilon_i$ using the same predictors as the multilevel specification.

As measures of accuracy, we report the Root Mean Squared Errors (RMSE) and the Nearest Neighbors proportions (NN). The RMSE is calculated as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i \in N}(y_i - \hat{y}_i)^2}, \tag{1}$$

with i identifying all the predicted grades and N the total number of grades. The RMSE score strongly punishes large errors. The second measure of performance is nearest neighbors percentage (NN)

$$\text{NN} = \frac{1}{N}\sum_{i \in N} I(\hat{y}_i - 1 \leq y_i \leq \hat{y}_i + 1). \tag{2}$$

Table 4. Confusion matrix of the predictions for the 2014/2015 end-of-year grades of 268 students based on the three-level model at $t = 0$.

		RMSE = 0.860 Actual grades							
		3	4	5	6	7	8	9	10
Predicted	3	0	1	1	0	0	0	0	0
	4	0	1	3	2	0	0	0	0
	5	3	10	19	27	11	2	0	0
	6	4	36	114	358	182	55	10	0
	7	2	10	83	425	749	434	79	3
	8	0	2	8	43	267	464	213	14
	9	0	0	0	1	8	44	118	22
	10	0	0	0	0	0	0	5	6

Note that an alternative criterion is the Mean Absolute Deviation (MAD). However, those results were comparable to the RMSE.

Table 2 reports the RMSE and NN for the hierarchical model ($\hat{y}_H$), the one-level linear regression fit ($\hat{y}_{sr}$) and the weighted average (y_i) at five points in time $t = 0, 0.1, 0.3, 0.5, 0.7$.

The two performance measures in Table 2 show the superiority of the hierarchical method $\hat{y}_H$ when predicting end-of-year grades at the beginning of the year ($t = 0$). As the year progresses, the relative advantage of the model decreases over time as more grades accumulate and the final grade is less uncertain. A comparison of Tables 3 and 4 clarifies the advantage of the hierarchical regression model compared to a one-level model. Both tables show the predicted and realized end-of-year grades before the start of the year. The difference in RMSE of 0.292 might not seem worth the trouble at first, but when we compare these two tables, Table 4 shows much more granularity in the results. The hierarchical model identifies much more structure in the data, which is especially valuable in predicting far above- and below-average grades.

4. Monitoring student performance

This section is about monitoring student performance using accumulated test grades. We will consider statistical process monitoring techniques and predictive monitoring.

4.1. Statistical process monitoring

To use a classical control chart technique (i.e. Shewhart, CUSUM or EWMA charts) we need a phase I data set that serves as a training set and a phase II data set that will be a test set (Vining 2009). Phase I is used to analyze the model and to estimate the parameters involved. The data used are assumed to be in control, and monitoring begins in phase II. In this case, and many other practical examples, there is no obvious phase I at hand. We could use student data from previous years as phase I. These are not available however, for first-year students, for new courses and in case of limited data. Furthermore, a second-year course is different from a first-year course and most students don't repeat a year. Identifying a clear phase I/phase II setup is thus difficult. These problems are amplified by the fact that y_i is not i.i.d., violating the assumptions of the basic use of charts.

By modeling y_i, we can correct for a lot of the problems we see for classical control charting techniques. We model y_i at time t using all test grades before time t, with $t \in \{t_I, T\}$ where t_I indicates the start of the school year and T the end of the school year. We then calculate an expected value $\hat{y}_i$. The difference between the expected value and the actual observed value y_i at time t can then be monitored in a phase II data set using a residual control chart setup.

4.1.1. Three-level control chart

In this case, we evaluate whether the relations given by the three-level model still hold. To this end, we monitor the residuals at the three levels. For existing groups, we have estimates of the full parameter space $\{\hat{\boldsymbol{\beta}}^{(L_0)}, \hat{\sigma}^2, \hat{\boldsymbol{\beta}}^{(L_1)}, \hat{\Sigma}^{(L_1)}, \hat{\boldsymbol{\beta}}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$. Then using these estimated parameters, we can calculate the residuals for the three levels for any new observation $\{y_i, \boldsymbol{X}_i^{(L_0)}, \boldsymbol{X}_j^{(L_1)}, \boldsymbol{X}_h^{(L_2)}\}$

$$
\begin{aligned}
r_i^{(L_0)} &= y_i - \boldsymbol{X}_i^{(L_0)} \hat{\boldsymbol{\beta}}_{j[i]}^{(L_0)} \\
\boldsymbol{r}_j^{(L_1)} &= \hat{\boldsymbol{\beta}}_j^{(L_0)} - \hat{\boldsymbol{\beta}}_{h[j]}^{(L_1)} \boldsymbol{X}_j^{(L_1)'}, \\
\boldsymbol{r}_h^{(L_2)} &= vec(\hat{\boldsymbol{\beta}}_h^{(L_1)}) - \hat{\boldsymbol{\beta}}^{(L_2)} \boldsymbol{X}_h^{(L_2)'},
\end{aligned}
$$

where $r_i^{(L_0)}, \boldsymbol{r}_j^{(L_1)}$ and $\boldsymbol{r}_h^{(L_2)}$ are the residual vectors at the three levels of size 1, $(p_0 + 1)$ and $(p_0 + 1)(p_1 + 1)$, respectively.

In line with traditional SPM techniques, we want to determine if a new observation stems from the in-control phase I distribution, which was obtained using estimation (i.e. phase I) data $\{\boldsymbol{X}_I^{(L_0)}, \boldsymbol{X}_I^{(L_1)}, \boldsymbol{X}_I^{(L_2)}, \boldsymbol{y}_I\}$ of size n_0, where $\boldsymbol{X}_I^{(L_0)}$ is the $n_0 \times (p_0 + 1)$ matrix with the ith row containing the intercept and predictor values for course i. The other matrices are constructed in a similar way. The residuals can be monitored using control charting techniques.

For example, we can use a Shewhart control chart taking the mean and variance estimates from phase I for $r_i^{(L_0)}$ with upper and lower control limits $\widehat{UCL}_y = 3\hat{\sigma}^2$ and $\widehat{LCL}_y = -3\hat{\sigma}^2$. The chart signals when the residual exceeds one of the control limits, after which the underlying cause can be investigated.

For $\boldsymbol{r}_j^{(L_1)}$ and $\boldsymbol{r}_h^{(L_2)}$, multivariate control charts are needed because these residuals are multidimensional. A multivariate Hotelling T^2 chart offers a solution with test statistics (cf. 11.23 in Montgomery 2007)

$$
T^2_{(L_1)} = n_0 \boldsymbol{r}_j^{(L_1)'} \hat{\Sigma}^{(L_1)} \boldsymbol{r}_j^{(L_1)}, \tag{3}
$$

$$
T^2_{(L_2)} = n_0 \boldsymbol{r}_h^{(L_2)'} \hat{\Sigma}^{(L_2)} \boldsymbol{r}_h^{(L_2)}, \tag{4}
$$

where n_0 is the number of observations used to estimate the covariance matrix. The lower control limit for these T^2 charts is $LCL = 0$, the upper control limit with false alarm percentage α is $UCL_{(L_1)} = \frac{p_1(n_0-1)}{n_0-p_1} F_{\alpha, p_1, n_0-p_1}$ for $T^2_{(L_1)}$ and $UCL_{(L_2)} = p_2(n_0 - 1)$ $n_0 - p_2 F_{\alpha, p_2, n_0-p_2}$ for $T^2_{(L_2)}$.

If the $T^2_{(L_2)}$ chart gives a signal, the root cause analysis can focus on the class level; if the $T^2_{(L_1)}$ chart gives a signal the root cause analysis can focus on the student level; and if the Shewhart chart gives a signal, the root cause analysis can focus on the course level.

Besides monitoring the residuals, there is the option of monitoring the parameter estimates. Similar to Kang and Albin (2000), a T^2 chart can be used to monitor the parameter estimates $\{\hat{\boldsymbol{\beta}}^{L_0}, \hat{\sigma}^2, vec(\hat{\boldsymbol{\beta}}^{(L_1)}), \hat{\Sigma}^{(L_1)}, \hat{\boldsymbol{\beta}}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$.

4.1.2. Example

To illustrate this three-level monitoring approach, we monitor the cumulative weighted average y_i at 15 times throughout the school year 2014/2015 using the same subset as in the previous. Phase I consists of the five school years from 2009 to 2014; phase II is the school year 2014/2015 for the 3,839 courses followed by 268 first-, second- and third-year students. We apply the hierarchical regression model and monitor the residuals using a Shewhart control chart.

The school aims to detect "exceptional" courses and students. It considers exceptional courses as final grades below 6 or above 8. Each point below 6 is counted as a "failpoint." A single course with an end-of-year grade 5 equals 1 failpoint; a single course with an end-of-year grade 3 equals 3 failpoints, and one course grade of 4 and one of 3 equals 5 failpoints, etc. On the other hand, each point above 8 is counted as an "excelpoint." Thus the maximum grade of 10 for a course equals 2 excelpoints. An exceptional student is a student with at least four failpoints, and/or at least four excelpoints.

The three-level model estimates have an overall RMSE of 1.172. Figure 5 displays an example of a Shewhart chart monitoring the residuals of the first level $r_i^{(L_0)}$. The chart signals four times near the end of the year. In total, the residuals charts signal 190 times (88 of which (46.32%) are exceptional courses), for 112 different students (36 of which (32.14%) are exceptional students).

As given by Eq. [3], we can also monitor the student level residuals using a Hotelling T^2 chart. Using the same data as in the previous, the T^2 chart signals at least once for 105 students (38 (36.19%) of which are exceptional students).

The charts signal exceptional cases throughout the year. However, we cannot retrospectively determine if at the time of a signal there was some unknown factor that influenced the performance of student j for course i. We are thus unable to distinguish false from true signals. It does, however, out-of-the-box, identify students whom we know have interesting performance during the monitoring phase.

The statistical monitoring approach identifies *incidental* anomalies in the weighted averages. However, the school's main focus is to identify students who need either support or more challenging coursework. This monitoring approach is insufficient for that goal. Therefore, in the next section, we use the hierarchical model to monitor student *expected* end-of-year results to identify under- or overperforming students.

4.2. Predictive monitoring

The high school in this case study aims to predict the end-of-year grades of its students. This enables the school to receive early warnings on exceptional students. In this section, we will thus consider predictive monitoring of student performance.

4.2.1. Multilevel predictive monitoring

As demonstrated in Section 3.3, the predictions of the three-level model are relatively accurate. Furthermore, the three-level model can be used for new students/classes and when there are a small number of courses per student or students per class. In this section, we will thus use the three-level model for predictive monitoring.

We want to monitor $P(E)_t$, defined as the probability of some event E at time t. $P(E)_t$ summarizes the outcome of the model into a single predictive probability at time t, with $t \in \{t_I, T\}$ where t_I indicates the start of the year and T the end of the year. The chart signals when $P(E)_t$ exceeds threshold C, which is defined as the maximum allowed probability of event E occurring ($0 < C < 1$). Event E concerns the values of y_i, which is context dependent and can take many forms ($y_i = e$, $y_i \geq e, y_i \leq e, e_1 \leq y_i \leq e_2$, $\sum_{i=a}^{b} y_i \geq e$ etc., where e, e_1 and e_2 are arbitrary constants and a and b are integers between 1 and n_0). Following the MCMC estimation of the posterior densities of the parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}^{(L_0)}, \sigma^2, \boldsymbol{\beta}^{(L_1)}, \Sigma^{(L_1)}, \boldsymbol{\beta}^{(L_2)}, \Sigma^{(L_2)}\}$ as described in the supplementary material, we can use the posterior densities to calculate $P(E)_t$.

The steps for predictive monitoring are

1. Define event E and threshold C
2. Specify the multilevel model for y_i

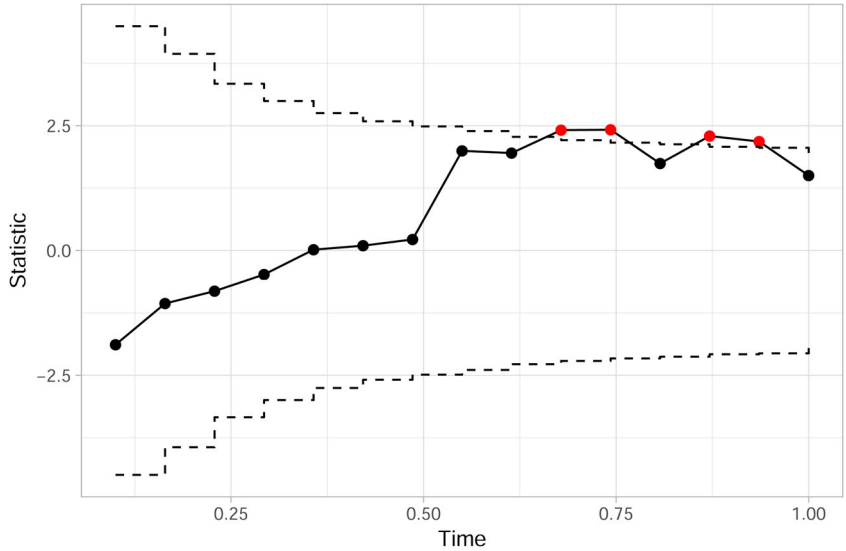


Figure 5. Residual Shewhart control chart monitoring $r_i^{(L_0)}$ based on a three-level regression (signals in red).

3. Estimate the parameters to obtain $\hat{\boldsymbol{\theta}}_{\boldsymbol{I}}$ using the phase I data at time t_I using MCMC, described in the appendix
4. Calculate $P(E)_t$ using the newly available observations at time $t > t_I$
5. Signal if $P(E)_t > C$
6. Re-estimate the parameters to obtain $\hat{}_t$ using all available data at time t and go back to step 4 for a new timepoint $t_{II} > t$.

Assume that we have a large in-control phase I data set $\{\boldsymbol{X}_{\boldsymbol{I}}^{(\boldsymbol{L_0})}, \boldsymbol{X}_{\boldsymbol{I}}^{(\boldsymbol{L_1})}, \boldsymbol{X}_{\boldsymbol{I}}^{(\boldsymbol{L_2})}, \boldsymbol{y_I}\}$ at time $t = t_I$. At time $t < t_I$ we obtain the estimates for the parameters $\{\hat{\boldsymbol{\beta}}^{(\boldsymbol{L_0})}, \hat{\sigma}^2, \hat{\boldsymbol{\beta}}^{(\boldsymbol{L_1})}, \hat{\boldsymbol{\Sigma}}^{(\boldsymbol{L_1})}, \hat{\boldsymbol{\beta}}^{(\boldsymbol{L_2})}, \hat{\boldsymbol{\Sigma}}^{(\boldsymbol{L_2})}\}$ based on observations in phase I. As described in the appendix for the three-level model, using the estimates of the parameters, at any time $t > t_I$ we have a predicted distribution for the outcome variable $\hat{y}_{i,t}$

$$\hat{y}_{i,t} \sim N((\boldsymbol{X}_{i,t}^{(L_0)} \otimes \boldsymbol{X}_{j[i,t]}^{(L_1)'})\hat{\boldsymbol{\beta}}^{(L_2)}\boldsymbol{X}_{h[j[i,t]]}^{(L_2)'},$$
$$(\boldsymbol{X}_{i,t}^{(L_0)} \otimes \boldsymbol{X}_{j[i,t]}^{(L_1)})\hat{\boldsymbol{\Sigma}}^{(L_2)}(\boldsymbol{X}_{i,t}^{(L_0)} \otimes \boldsymbol{X}_{j[i,t]}^{(L_1)'}) + \boldsymbol{X}_{i,t}^{(L_0)}\hat{\boldsymbol{\Sigma}}^{(L_1)}\boldsymbol{X}_{i,t}^{(L_0)'} + \hat{\sigma}^2),$$

where $\otimes$ is the Kronecker product. We can use this result to estimate the probability of the outcome $P(E)_t$. The event E can take several forms. Suppose we consider $y_i \leq e$, i.e. we study that the grade y_i is less than e. The monitoring scheme we propose uses the posterior distribution of $\hat{y}_{i,t}$ to calculate the probability $P(E)_t$. The chart signals when $P(E)_t > C$, with C the threshold that determines the maximum allowed probability of event E.

Monitoring $P(E)_t$ requires periodic re-estimation of the parameters to incorporate newly available information at time t. Around the time event E occurs, the probability $P(E)_t$ converges to 1 if $t \to T$. The major advantage of monitoring $P(E)_t$ instead of $y_{i,t}$ is that, depending on the predictive capability of the multilevel model, the monitoring scheme provides early warning and the opportunity to intervene before event E occurs. If intervention occurs, it is important to include this in the predictors $\{\boldsymbol{X}^{(\boldsymbol{L_0})}, \boldsymbol{X}^{(\boldsymbol{L_1})}, \boldsymbol{X}^{(\boldsymbol{L_2})}\}$ by including an additional variable, to extract the effect of the intervention on outcome E. Furthermore, there is no need for n_0 control charts. All that is required is a single control chart plotting values of $P(E)_t$ and signaling for observations or groups for which $P(E)_t$ exceeds C.

4.2.2. Example

Following the steps outlined before, we define two events: E^f as a student failing the year and E^e as a student excelling that year. E^f occurs if a student has four or more failpoints, as defined in the previous section (the number of points below 6 for all courses a student follows in a year). E^e occurs if a student has four or more excelpoints (the number of points above 8 for all courses a student follows in a year).

The end-of-year rounded grade of student j for course i is defined as y_{ij}. At time t, the probability of a student failing the year can thus be summarized by $P(E_j^f)_t = P(\sum_{i=1}^{n_j} \max(0, (6 - y_{ij})) \geq 4)_t$, where n_j is the number of courses for student j. The probability of a student excelling in the year can then be summarized by $P(E_j^e)_t = P(\sum_{i=1}^{n_j} \max(0, (y_{ij} - 8)) \geq 4)_t$ at time t.

Using the same data set as in the previous section, Figure 6 shows a control chart of $1 - P(E_j^f)_t$ for $J = 268$ students at 15 points in time. As an example, the threshold $C = 0.05$ is depicted as a dashed line. Note that $1 - P(E_j^f)_t$ equals the probability of passing the year. The $J_p = 238$ students who passed are depicted in blue and the probabilities of the $J_f = 30$ students who failed in red. Although there are some exceptions, overall the model consistently estimates the passing probabilities for the students who fail the year much lower than the students who pass the year. This can also be seen in the probabilities of failure in Table 5. This table reports the values of $\frac{1}{J_p}\sum_{j \in J_p} P(E_j^f)_t$ (the average estimated probability of failure for students that pass the year) in the top row and $\frac{1}{J_f}\sum_{j \in J_f} P(E_j^f)_t$ (the average estimated probability of failure for students that fail the year) in the bottom row. The model consistently assigns a higher average probability of failure to students that end up failing the year.

Figure 7 plots $P(E_j^e)_t$ for the same $J = 268$ students. The $J_n = 222$ students who did not excel are depicted in red and the probabilities of the $J_e = 46$ students who excelled are depicted in blue. As an example, threshold $C = 0.95$ is depicted as a dashed line. The model has impressive performance, shown also by the differences in average probabilities over time between students who excel, $\frac{1}{J_e}\sum_{j \in J_e} P(E_j^e)_t$, and those that do not, $\frac{1}{J_n}\sum_{j \in J_n} P(E_j^e)_t$, as depicted in Table 6.

Depending on the threshold C that determines if the monitoring scheme signals, the model correctly identifies several students who will fail/excel as well as some false positives. Tables 7 and 8 report the precision and recall values monitoring E^f and E^e, respectively, where the precision is defined as

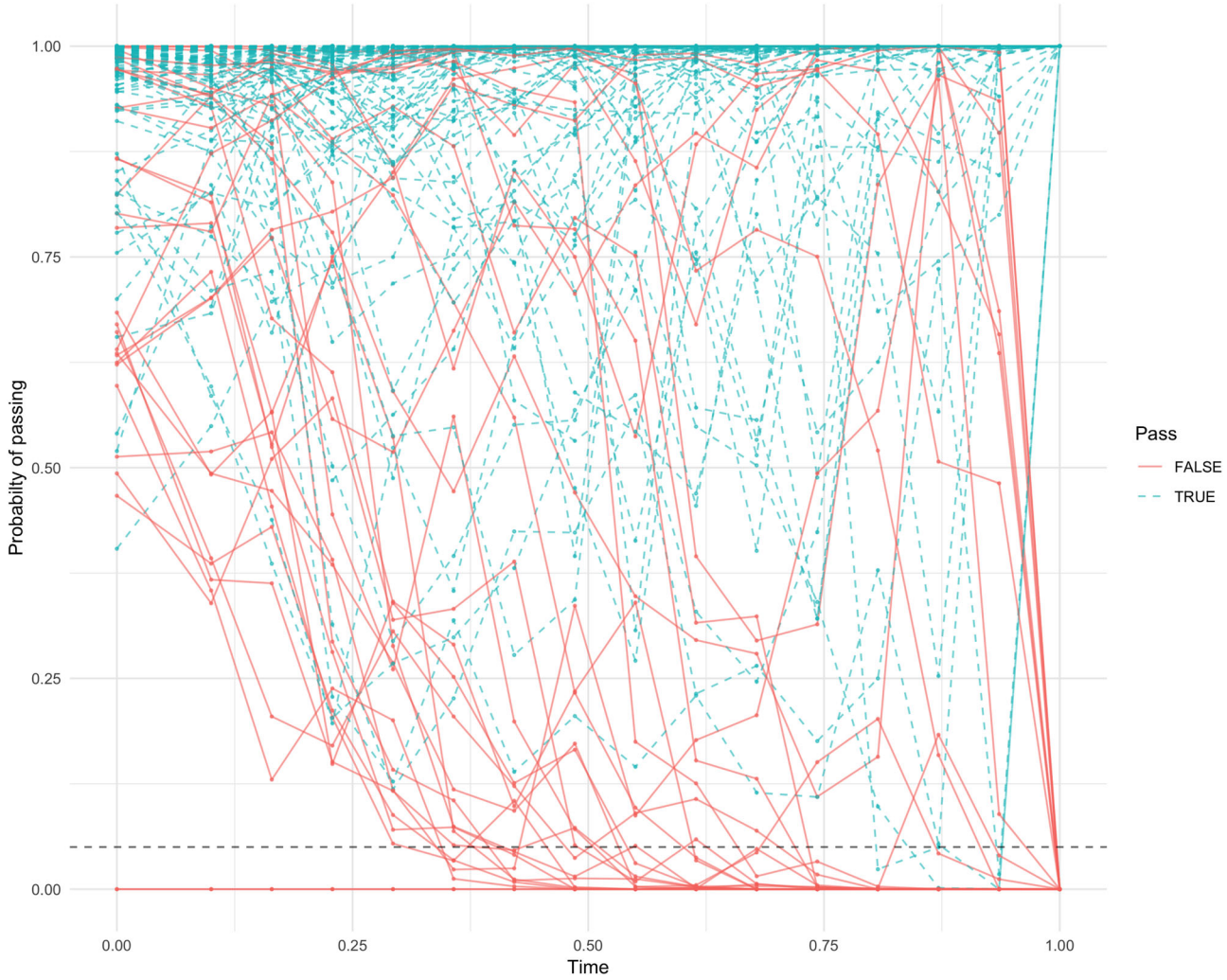


Figure 6. A control chart monitoring the estimated probabilities of passing $1 - P(E^f)_t$ for 268 students in 2014/2015, with dashed threshold $C = 0.05$ in black. The dashed blue lines represent students that passed, the red solid lines students that failed.

Table 5. Average estimated probabilities of failing $P(E^f)_t$ for 268 students in 2014/2015, split by observed outcome.

	Time						
Failed	0	0.1	0.3	0.5	0.7	0.9	1
No	0.02	0.02	0.04	0.03	0.03	0.01	0.00
Yes	0.27	0.28	0.52	0.61	0.75	0.79	1.00

$$\text{Precision}_t(C) = \frac{tp_t(C)}{tp_t(C) + fp_t(C)}$$

with $tp_t(C)$ equal to the number of true positives at time t for threshold C and $fp_t(C)$ the number of false positives at time t for threshold C. The recall is given by

$$\text{Recall}_t(C) = \frac{tp_t(C)}{tp_t(C) + fn_t(C)}$$

where $fn_t(C)$ equals the number of false negatives at time t for threshold C (Powers 2011).

Table 7 shows the procedure correctly identifies students who will fail the year early on. The performance is impressive, where, depending on the chosen level of C, multiple early warnings are generated aiding in the student support system. For example, setting C at 0.75, the procedure identifies almost half (14 out of 30) of the students who will fail before the start of the year with only 26% (5) false positives.

Table 8 shows the precision and recall values when predicting excelling students. Depending on the school's preferences, high precision or recall can be achieved early on in the year. For example, setting C at 0.50, the procedure identifies half (23 out of 46) of the students who will excel before the start of the year with only 15% (4) false positives.

The multilevel monitoring procedure has shown its value in a high school setting, as it adequately provides expected end-of-year grades for all students and subjects. This can aid in classifying at-risk students

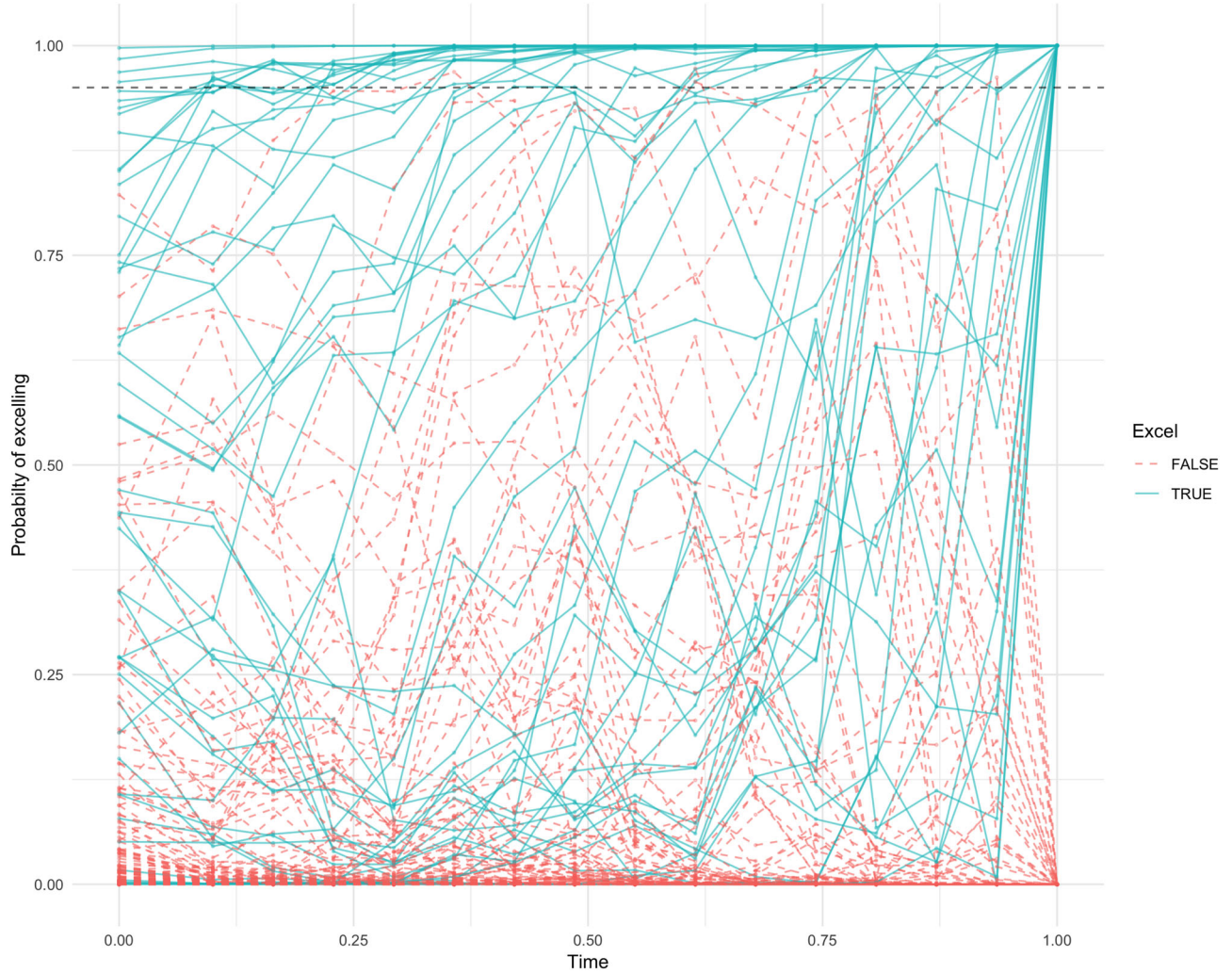


Figure 7. A control chart monitoring the estimated probabilities of excelling $P(E^e)_t$ for 268 students in 2014/2015, with dashed threshold $C = 0.95$ in black. The solid blue lines represent students that excelled, the red dashed lines students that did not excel.

Table 6. Average estimated probabilities of excelling $P(E^e)_t$ for 268 students in 2014/2015, split by observed outcome.

	Time						
Excelled	0	0.1	0.3	0.5	0.7	0.9	1
No	0.05	0.04	0.04	0.06	0.04	0.03	0.00
Yes	0.50	0.49	0.50	0.61	0.67	0.81	1.00

Table 7. Precision$_t(C)$ (Recall$_t(C)$) results when monitoring $P(E^f)_t$ with various values of C and t using the three-level model predictions of end-of-year grades for 268 students in 2014/2015.

C							
		0.05	0.1	0.25	0.5	0.75	0.999
Time	0	1 (0.07)	1 (0.07)	1 (0.07)	0.67 (0.13)	0.74 (0.47)	0.25 (0.93)
	0.1	1 (0.07)	1 (0.07)	1 (0.07)	1 (0.27)	0.71 (0.40)	0.25 (0.93)
	0.3	1 (0.10)	1 (0.20)	0.85 (0.37)	0.76 (0.53)	0.67 (0.67)	0.27 (1)
	0.5	1 (0.33)	1 (0.43)	0.94 (0.53)	0.79 (0.63)	0.67 (0.67)	0.34 (0.97)
	0.7	1 (0.57)	1 (0.63)	0.88 (0.73)	0.77 (0.70)	0.70 (0.77)	0.40 (0.97)
	0.9	0.90 (0.63)	0.86 (0.63)	0.88 (0.70)	0.81 (0.70)	0.81 (0.73)	0.59 (0.90)
	1	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)

who need support, as well as the areas in which they need help. On the other side of the spectrum, the model successfully identifies excelling students who can benefit from more challenging schoolwork. The model further provides easily interpretable results, as well as good explainability for the parameters.

Table 8. Precision$_t(C)$ (Recall$_t(C)$) results when monitoring $P(E^e)_t$ with various values of C and t using the three-level model predictions of end-of-year grades for 268 students in 2014/2015.

C							
		0.99	0.95	0.75	0.5	0.25	0.01
Time	0	1 (0.02)	1 (0.09)	0.93 (0.3)	0.85 (0.5)	0.69 (0.72)	0.38 (0.89)
	0.1	1 (0.04)	1 (0.2)	0.94 (0.35)	0.72 (0.46)	0.71 (0.65)	0.45 (0.87)
	0.3	1 (0.09)	1 (0.28)	0.89 (0.37)	0.83 (0.54)	0.68 (0.54)	0.45 (0.85)
	0.5	1 (0.37)	1 (0.41)	0.92 (0.52)	0.77 (0.59)	0.65 (0.67)	0.47 (0.96)
	0.7	1 (0.43)	1 (0.48)	0.89 (0.54)	0.88 (0.61)	0.72 (0.78)	0.57 (0.93)
	0.9	1 (0.57)	1 (0.63)	0.94 (0.74)	0.88 (0.83)	0.8 (0.87)	0.64 (1)
	1	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)

5. Conclusions

This case study has considered three research questions concerning high school students' performance. We worked together with a Dutch high school in attempting to answer the following questions (1) What determines student performance? (2) How can statistical process monitoring be used in monitoring student progress? (3) What method can be used for predictive monitoring of student results? This resulted in the use of a three-level model in a predictive monitoring scheme that can be applied when monitoring hierarchical data. We discuss our results in the following.

5.1. What determines student performance?

The detailed data set made available by a Dutch high school has shown interesting determinants of student performance. These are generally in line with the educational literature and are useful when monitoring student progress.

Female students were found to obtain higher final grades. In line with the literature, students with disabilities perform slightly worse. Children with highly educated parents outperform their peers with less-educated parents in this case study.

The nationality and language barrier variables represent an interesting case study of the discussed theory on immigrant and language barriers in academia. Consistent with work by Geay, McNally, and Telhaj (2013) and the "language broker" effect of Buriel et al. (1998), students born abroad achieve similar results to their locally born peers. A serious language barrier does seem to produce slightly lower grades. This, in turn, is consistent with findings by Kennedy and Park (1994) and Collier (1995).

Students show a decrease in performance through their high school career, with around half a point difference in grades between the first and fourth years of high school. Absences seem to have a strong negative correlation with grades, which justifies the penalization of these types of absences. On a policy level, the relationship between the primary school test scores (CITO) and student grades should be considered toward current discussion around the determinants of the high school level.

The main goal of the school was to monitor student performance as the process output throughout the year. Therefore, statistical and predictive monitoring techniques were considered.

5.2. Statistical process monitoring

Classical statistical process monitoring techniques are often insufficient when applied to complex processes, for which increasingly large data sets are available. When a hierarchical structure is present in the data set, multilevel modeling improves the reliability of process monitoring. Using multilevel models improve estimation accuracy and explainability over regular linear regression models. Furthermore, the method is essential for predictive modeling of new students/classes or students/classes with small sample sizes.

Univariate statistical process monitoring techniques proved insufficient in this case study and one-level linear regression models did not provide satisfactory results. We have discussed a three-level model together with the monitoring options. Residual control charting at the three levels was proposed as the multilevel statistical monitoring method for online monitoring of process output. The proposed multilevel monitoring framework did provide promising results.

5.3. Predictive monitoring

A predictive monitoring method has been developed to enable an early warning monitoring system. This method monitors the probability of an event, rather than a process output. The three-level model was used to continuously predict end-of-year individual grades. Using a Bayesian hierarchical model, probability distributions for the student outcomes are obtained. These can be used to monitor unwanted results in the form of under- and overperforming students using a

single predictive control chart setup. This predictive monitoring approach was shown to be very useful in practice, as the school obtains valuable early warnings on both under- and overperforming students.

The proposed multilevel process monitoring framework can be useful across many applications, including industrial processes (batch production, multiple factories), market monitoring, HR analytics, sports and more. Implementation of multilevel models can be challenging, however, especially in a Bayesian setting. Sampling procedures can be used to simplify the analysis. We have provided a full analysis of the three-level model and its estimation in the supplementary material, where we used Gibbs sampling to estimate the parameters. Using these parameters, predictions were made for the monitoring period, after which the parameters can be updated to improve the predictive power of the model. Predictive monitoring results in early warning systems, that can greatly aid in early detection and prevention of special cause variation.

We argue the importance of predictive monitoring in general. As more and more data are available, the use of more complex models can extract more information toward valuable predictions. Summarizing complex processes into simple and interpretable results is essential. Multilevel modeling is one method that achieves this, which is applicable in cases where a clear hierarchy is present. There are of course many more statistical and machine learning methods that can be applied. We encourage research that investigates the use of these methods in a predictive monitoring setting.

Concluding this paper, early warning indicator systems have the potential to improve the educational system at a low cost. These systems can add a layer of sophistication to school and teacher performance evaluation and work toward fulfilling individual student needs.

Acknowledgments

We want to thank the Dutch high school for participating in this project and sharing the valuable data. We are grateful to Dr. Reza Mohammadi and Dr. Maurice Bun (University of Amsterdam) and Dr. Inez Zwetsloot (City University of Hong Kong) for their helpful comments. We also thank the referees and the case studies department editor for their valuable suggestions to improve the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

About the authors

Leo C. E. Huberts is a PhD student and lecturer at the Department of Operations Management and consultant at the Institute for Business and Industrial Statistics of the University of Amsterdam, the Netherlands. His current research topic is statistical and predictive process monitoring.

Marit Schoonhoven is an associate professor at the Department of Operations Management and senior consultant at the Institute for Business and Industrial Statistics of the University of Amsterdam, the Netherlands. Her current research interests include control charting techniques and operations management methods.

Ronald J. M. M. Does is professor of Industrial Statistics at the University of Amsterdam and Head of the Department of Operations Management at the Amsterdam Business School. He is a Fellow of the ASQ and ASA, an elected member of the ISI, and an Academician of the International Academy for Quality. His current research activities include the design of control charts for nonstandard situations, healthcare engineering, and operations management methods.

References

Aaronson, D. 1998. Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *The Journal of Human Resources* 33 (4):915–46. doi: 10.2307/146403.

Aitkin, M., and N. Longford. 1986. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (General)* 149 (1):1–26. doi: 10.2307/2981882.

Baghdadi, A., L. A. Cavuoto, A. Jones-Farmer, S. E. Rigdon, E. T. Esfahani, and F. M. Megahed. 2019. Monitoring worker fatigue using wearable devices: A case study to detect changes in gait parameters. *Journal of Quality Technology* :1–25. doi: 10.1080/00224065.2019.1640097.

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2014. *Hierarchical modeling and analysis for spatial data*. London: Chapman and Hall/CRC.

Battin-Pearson, S., M. D. Newcomb, R. D. Abbott, K. G. Hill, R. F. Catalano, and J. D. Hawkins. 2000. Predictors of early high school dropout: A test of five theories. *Journal of Educational Psychology* 92 (3):568–82. doi: 10.1037/0022-0663.92.3.568.

Becker, J., L. S. Hall, B. Levinger, A. Sims, and A. Whittington. 2014. Student success and college readiness: Translating predictive analytics into action. Strategic Data Project, SDP Fellowship Capstone Report. http://sdp.cepr.harvard.edu/files/cepr-sdp/files/sdp-fellowship-capstone-student-success-college-readiness.pdf

Berendsen, H. J. 2007. *Simulating the physical world: Hierarchical modeling from quantum mechanics to fluid dynamics*. Cambridge: Cambridge University Press.

Bock, R. D. 1989. *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Browne, W. J., and D. Draper. 2006. A comparison of Bayesian and likelihood-based methods for fitting

multilevel models. *Bayesian Analysis* 1 (3):473–514. doi: 10.1214/06-BA117.

Buriel, R., W. Perez, T. L. de Ment, D. V. Chavez, and V. R. Moran. 1998. The relationship of language brokering to academic performance, biculturalism, and self-efficacy among Latino adolescents. *Hispanic Journal of Behavioral Sciences* 20 (3):283–97. doi: 10.1177/07399863980203001.

Casella, G., and E. I. George. 1992. Explaining the Gibbs sampler. *The American Statistician* 46 (3):167–74. doi: 10.2307/2685208.

Collier, V. P. 1995. Acquiring a second language for school. *Directions in Language and Education* 1 (4):3–13.

Deary, I. J., S. Strand, P. Smith, and C. Fernandes. 2007. Intelligence and educational achievement. *Intelligence* 35 (1):13–21. doi: 10.1016/j.intell.2006.02.001.

Geay, C., S. McNally, and S. Telhaj. 2013. Non-native speakers of English in the classroom: What are the effects on pupil performance? *The Economic Journal* 123 (570): F281–307. doi: 10.1111/ecoj.12054.

Geiser, S., and M. V. Santelices. 2007. Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. *UC Berkeley: Center for Studies in Higher Education* 6 (7):1–35.

Gelman, A. 2006. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* 48 (3):432–5. doi: 10.1198/004017005000000661.

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. M. von Krosigk. 1959. The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15 (2):192–218. doi: 10.2307/2527669.

Hox, J. J., M. Moerbeek, and R. Van de Schoot. 2017. *Multilevel analysis: Techniques and applications*. London: Routledge.

Kang, L., and S. L. Albin. 2000. On-line monitoring when the process yields a linear profile. *Journal of Quality Technology* 32 (4):418–26. doi: 10.1080/00224065.2000.11980027.

Kang, L., X. Kang, X. Deng, and R. Jin. 2018. A Bayesian hierarchical model for quantitative and qualitative responses. *Journal of Quality Technology* 50 (3):290–308. doi: 10.1080/00224065.2018.1489042.

Karande, S., and M. Kulkarni. 2005. Poor school performance. *Indian Journal of Pediatrics* 72 (11):961–7. doi: 10.1007/BF02731673.

Kennedy, E., and H.-S. Park. 1994. Home language as a predictor of academic achievement: A comparative study of Mexican- and Asian-American youth. *Journal of Research & Development in Education* 27 (3):188–94.

Laidra, K., H. Pullmann, and J. Allik. 2007. Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and Individual Differences* 42 (3): 441–51. doi: 10.1016/j.paid.2006.08.001.

Laurenceau, J.-P., L. F. Barrett, and M. J. Rovine. 2005. The interpersonal process model of intimacy in marriage: A daily-diary and multilevel modeling approach. *Journal of Family Psychology* 19 (2):314–23. doi: 10.1037/0893-3200.19.2.314.

Mandel, B. J. 1969. The regression control chart. *Journal of Quality Technology* 1 (1):1–9. doi: 10.1080/00224065.1969.11980341.

Montgomery, D. C. 2007. *Introduction to statistical quality control*. Hoboken, NJ: John Wiley & Sons.

Nichols, J. D. 2003. Prediction indicators for students failing the state of Indiana high school graduation exam. *Preventing School Failure: Alternative Education for Children and Youth* 47 (3):112–20. doi: 10.1080/10459880309604439.

Nichols, J. D., and J. White. 2001. Impact of peer networks on achievement of high school algebra students. *The Journal of Educational Research* 94 (5):267–73. doi: 10.1080/00220670109598762.

Parker, J. D., M. J. Hogan, J. M. Eastabrook, A. Oke, and L. M. Wood. 2006. Emotional intelligence and student retention: Predicting the successful transition from high school to university. *Personality and Individual Differences* 41 (7):1329–36. doi: 10.1016/j.paid.2006.04.022.

Plummer, M. 2018. *rjags: Bayesian graphical models using MCMC*. R package version 4-8. https://CRAN.R-project.org/package=rjags

Powers, D. M. W. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2 (1):37–63.

Qiu, P., C. Zou, and Z. Wang. 2010. Nonparametric profile monitoring by mixed effects modeling. *Technometrics* 52 (3):265–77. doi: 10.1198/TECH.2010.08188.

Rahafar, A., M. Maghsudloo, S. Farhangnia, C. Vollmer, and C. Randler. 2016. The role of chronotype, gender, test anxiety, and conscientiousness in academic achievement of high school students. *Chronobiology International* 33 (1): 1–9. doi: 10.3109/07420528.2015.1107084.

Rohde, T. E., and L. A. Thompson. 2007. Predicting academic achievement with cognitive ability. *Intelligence* 35 (1):83–92. doi: 10.1016/j.intell.2006.05.004.

Romero, C., and S. Ventura. 2019. Guest editorial: Special issue on early prediction and supporting of learning performance. *IEEE Transactions on Learning Technologies* 12 (2):145–7. doi: 10.1109/TLT.2019.2908106.

Rothman, S. 2001. School absence and student background factors: A multilevel analysis. *International Education Journal* 2 (1):59–68.

Rothstein, J. M. 2004. College performance predictions and the SAT. *Journal of Econometrics* 121 (1-2):297–317. doi: 10.1016/j.jeconom.2003.10.003.

Sang, H., and A. E. Gelfand. 2009. Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics* 16 (3):407–26. doi: 10.1007/s10651-007-0078-0.

Schirru, A., S. Pampuri, and G. De Nicolao. 2010. Multilevel statistical process control of asynchronous multi-stream processes in semiconductor manufacturing. In 2010 IEEE International Conference on Automation Science and Engineering, 57–62, Toronto, ON, Canada.

Sellström, E., and S. Bremberg. 2006. Is there a "school effect" on pupil outcomes? A review of multilevel studies. *Journal of Epidemiology & Community Health* 60 (2):149–55.

Stewart, E. B. 2008. School structural characteristics, student effort, peer associations, and parental involvement: The

influence of school-and individual-level factors on academic achievement. *Education and Urban Society* 40 (2): 179–204. doi: 10.1177/0013124507304167.

Sui-Chu, E. H., and J. D. Willms. 1996. Effects of parental involvement on eighth-grade achievement. *Sociology of Education* 69 (2):126–41. doi: 10.2307/2112802.

Vining, G. 2009. Technical advice: Phase I and phase II control charts. *Quality Engineering* 21 (4):478–9. doi: 10.1080/08982110903185736.

Wang, Y. F., S. T. Tseng, B. H. Lindqvist, and K. L. Tsui. 2019. End of performance prediction of lithium-ion batteries. *Journal of Quality Technology* 51 (2):198–213. doi: 10.1080/00224065.2018.1541388.

Weese, M., W. Martinez, F. M. Megahed, and L. A. Jones-Farmer. 2016. Statistical learning methods applied to process monitoring: An overview and perspective. *Journal of Quality Technology* 48 (1):4–24. doi: 10.1080/00224065.2016.11918148.

Woodall, W. H., and D. C. Montgomery. 2014. Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology* 46 (1): 78–94. doi: 10.1080/00224065.2014.11917955.

Appendix

A.1. Predictive distribution

If we represent the three-level model in the following way

$$\begin{aligned} y_i &= \boldsymbol{X}_i^{(L_0)}\boldsymbol{\beta}_{j[i]}^{(L_0)} + \varepsilon_i^{(L_0)}, \varepsilon^{(L_0)} \sim N(0,\sigma_y^2) \\ \boldsymbol{\beta}_j^{(L_0)} &= \boldsymbol{\beta}_{h[j]}^{(L_1)}\boldsymbol{X}_j^{(L_1)'} + \boldsymbol{\varepsilon}_j^{(L_1)}, \boldsymbol{\varepsilon}^{(L_1)} \sim N(0,\boldsymbol{\Sigma}^{(L_1)}) \\ vec(\boldsymbol{\beta}_h^{(L_1)}) &= \boldsymbol{\beta}^{(L_2)}\boldsymbol{X}_h^{(L_2)'} + \boldsymbol{\varepsilon}_h^{(L_2)}, \boldsymbol{\varepsilon}^{(L_2)} \sim N(0,\boldsymbol{\Sigma}^{(L_2)}) \end{aligned} \tag{A1}$$

we can summarize the model into

$$\begin{aligned} y_i = \boldsymbol{X}_i^{(L_0)} vec^{-1}(\boldsymbol{\beta}^{(L_2)}\boldsymbol{X}_{h[j[i]]}^{(L_2)'})\boldsymbol{X}_{j[i]}^{(L_1)'} + \boldsymbol{X}_i^{(L_0)} vec^{-1}(\boldsymbol{\varepsilon}_h^{(L_2)})\boldsymbol{X}_{j[i]}^{(L_1)'} \\ + \boldsymbol{X}_i^{(L_0)}\boldsymbol{\varepsilon}_{j[i]}^{(L_1)} + \varepsilon_i^{(L_0)}. \end{aligned}$$

We obtain parameter estimates $\{\hat{\boldsymbol{\beta}}^{(L_0)}, \hat{\sigma}^2, \hat{\boldsymbol{\beta}}^{(L_1)}, \hat{\Sigma}^{(L_1)}, \hat{\boldsymbol{\beta}}^{(L_2)}, \hat{\Sigma}^{(L_2)}\}$ using the observations during phase I time period $t < t_I$. At any time $t > t_I$ we have a predicted distribution for the outcome variable $\hat{y}_{i,t}$. Considering the distributions of the error terms $\hat{y}_{i,t}$ has a normal distribution

$$\begin{aligned} \hat{y}_{i,t} \sim N((\boldsymbol{X}_{j[i,t]}^{(L_1)} \otimes \boldsymbol{X}_{i,t}^{(L_0)})\hat{\boldsymbol{\beta}}^{(L_2)}\boldsymbol{X}_{h[j[i,t]]}^{(L_2)'}, \\ (\boldsymbol{X}_{j[i,t]}^{(L_1)} \otimes \boldsymbol{X}_{i,t}^{(L_0)})\hat{\boldsymbol{\Sigma}}^{(L_2)}(\boldsymbol{X}_{j[i,t]}^{(L_1)} \otimes \boldsymbol{X}_{i,t}^{(L_0)})' + \boldsymbol{X}_{i,t}^{(L_0)}\hat{\boldsymbol{\Sigma}}^{(L_1)}\boldsymbol{X}_{i,t}^{(L_0)'} + \hat{\sigma}^2), \end{aligned}$$

where $\otimes$ is the Kronecker product and we use the relationship $vec(ABC) = (C' \otimes A)vec(B)$.

A.2. Prior distributions

The full parameter space $\boldsymbol{\theta} = \{\boldsymbol{\beta}^{(L_0)}, \sigma^2, \boldsymbol{\beta}^{(L_1)}, \boldsymbol{\Sigma}^{(L_1)}, \boldsymbol{\beta}^{(L_2)}, \boldsymbol{\Sigma}^{(L_2)}\}$, where $\boldsymbol{\beta}^{(L_0)}$ and $\boldsymbol{\beta}^{(L_1)}$ are constructed by stacking the parameter matrices $\boldsymbol{\beta}_j^{(L_0)}$ and $\boldsymbol{\beta}_h^{(L_1)}$ for all groups j and h respectively, are estimated using the Gibbs sampler (Casella and George 1992). The Gibbs sampler approximates the posterior distribution by sampling from the full conditional distributions of the parameters. We use the rJAGS package in R to link to JAGS (Plummer 2018).

The estimation requires prior distributions for the unknown parameter space. Parameters $\boldsymbol{\beta}^{(L_0)}$ and $\boldsymbol{\beta}^{(L_1)}$ have priors given explicitly by the model. Proper diffuse priors are chosen for parameters $\{\sigma^2, \boldsymbol{\Sigma}^{(L_1)}, \boldsymbol{\beta}^{(L_2)}, \boldsymbol{\Sigma}^{(L_2)}\}$.

The vector $vec(\boldsymbol{\beta}^{(L_2)})$ has a multivariate normal prior $N(\boldsymbol{a}, \boldsymbol{B})$, with diagonal covariance matrix $\boldsymbol{B}$ and larger values of $\boldsymbol{B}$ reflecting greater uncertainty. Thus proper but diffuse priors were determined, with $\boldsymbol{a} = 0$ and $\boldsymbol{B} = 1000\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

The covariance matrix $\boldsymbol{\Sigma}^{(L_1)}$ associated with level 1 student unobserved differences and the covariance matrix $\boldsymbol{\Sigma}^{(L_2)}$ for unobserved group level 2 differences are both defined as positive definite matrices with Inverse Wishart priors $W^{-1}(\boldsymbol{C}, (p_0 + 1) + 1)$ for $\boldsymbol{\Sigma}^{(L_1)}$ and prior $W^{-1}(\boldsymbol{D}, (p_0 + 1)(p_1 + 1) + 1)$ for $\Sigma^{(L_2)}$. $\boldsymbol{C}$ and $\boldsymbol{D}$ are diagonal matrices, where smaller values correspond to more diffuse priors. Values for these inverse Wishart distributions are set at $\boldsymbol{C} = \boldsymbol{D} = diag(0.001)$.

For the variance parameter σ^2 of the error term in the model the inverse Gamma distribution, $IG(a, b)$, was chosen. We use an uniformative prior, with parameters $a = 0.001; b = 1$; $\sigma^2 \sim IG(0.001, 1)$.

Using Statistical Engineering in Solving Pharmaceutical and Biotech Problems

Ronald D. Snee

Statistical engineering can be applied in a wide diversity of areas. This case study shows how statistical engineering is used to solve a large, complex biopharmaceutical supply chain problem. This major undertaking has been discussed previously (McGurk 2004). This article shows how the project has all the characteristics of a statistical engineering project. The project story complete with steps and results is presented to illustrate how a statistical engineering project is conducted.

Statistical Engineering can help solve big problems. Since the emergence of Six Sigma around 1987 (Six Sigma, Wikipedia), there has been a growing awareness that some problems are too large, complex and unstructured to be solved with traditional problem-solving methods, including Lean Six Sigma. The case study to be discussed in the article is different than from those that can be solved through routine problem solving, or even through the use of Lean Six Sigma. Some attributes of these types of problems are:

- Large
- Complex
- Unstructured
- Data Challenges
- Lack of a single "correct" solution
- Need for a strategy

Statistical engineering (Hoerl and Snee 2017) was developed as an overall approach to developing a strategy to attack such problems. The International Statistical Engineering Association (ISEA) defines Statistical Engineering as: "The study of systematic integration of statistical concepts, methods, and tools, often with other relevant disciplines, to solve important problems sustainably" Note that statistical engineering is not a problem-solving methodology per se, such as Lean Six Sigma, but rather a discipline. A generic statistical engineering framework to attack large, complex unstructured problems is discussed in Snee and Hoerl (2018) and in the International Statistical Engineering Association (ISEA) website (www.isea-change.org).

The phases of a statistical engineering framework are shown in Figure 1 along with the purpose and critical work elements in each phase. We see that there are six phases, not "seven easy steps". It is not a "linear process"; recycles and iterations are required as the project progresses and learnings accumulate. Each phase needs to be tailored, depending on the problem structure and context. Several projects and data-based studies are often required.

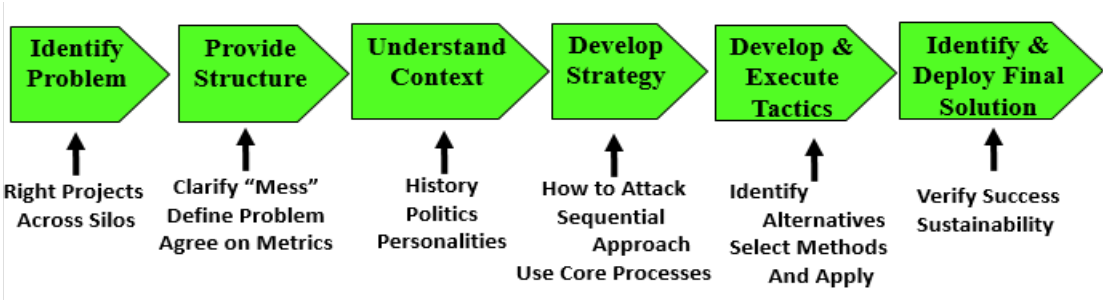


Figure 1. Phases of Statistical Engineering Projects

This framework will be used to describe a major biopharmaceutical supply chain improvement project including the problem solving process and results. As we will see the statistical engineering aspects of this project should come as no surprise. Talented problem solvers have been using the fundamentals of statistical engineering for a very long time. The problem has been that each project has been treated as a new event requiring that an approach be built from scratch. All the learnings from previous major projects were not recorded, lost or ignored. Statistical Engineering codifies the problem solving process giving the project team a head start including guidance on the work to be conducted, the sequence in which the various phases should be done and problems and issues that might be encountered along the way.

Phase 1: Identify the Problem

A major global pharmaceutical company faced supply challenges with two of its major biopharmaceutical products and decided to take systematic action to ensure a reliable supply of patient-critical products; a blockbuster drug and a monoclonal antibody. These two drugs were produced by different biopharmaceutical processes. This Company knew that one of the products was certain to be a blockbuster. A review of production and released product levels convinced the Vice President in charge of the two products that market demand for the blockbuster product could not be met. The monoclonal antibody product, had already encountered supply problems. The company was determined to establish predictable manufacturing capability for both products and meet the challenge of enormous market demands, all while ensuring sustainable Good Manufacturing Practices (GMP) compliance.

Management was also concerned that the organization may not have the experience and skills to identify and implement the needed changes. Clearly a major comprehensive review and improvement of these products and process that produced them was needed. The organization realized that this was a major problem that crossed organizational lines. A consulting firm was engaged to undertake this initiative as the firm had the needed capability and capacity to do the work.

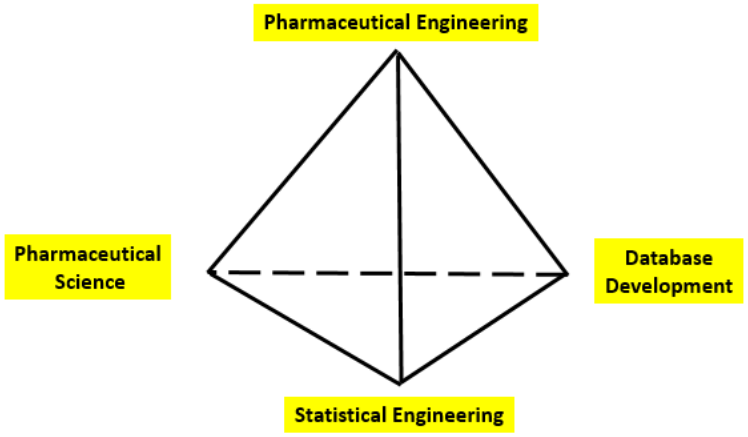


Figure 2. Required Project Team Skills

The consulting firm and the company worked collaboratively throughout the project. One of the first steps was to form a team that had the required experience, knowledge and skills. The personnel involved pharmaceutical science, process engineering, data base development and management and statistical engineering (Figure 2). Such a broad skilled team was needed to ensure that all the relevant knowledge was available to carefully collect the relevant data, perform the appropriate analyses and develop solutions that were workable and would be sustainable over time. Such diversity of skills is characteristic of statistical engineering projects. It is rare that the broad range of skills and knowledge required to successfully complete these projects reside one or two persons. Thus, a multi-skilled team is needed. Separate sub-teams were created for each product.

Phase 2: Provide Structure – Clean up the Mess

The scope and project goals were very clear at the beginning. The focus was on meeting the launch date for the new product and obtaining a major increase of the yield of the monoclonal antibody product. The two sub-teams were managed by a common project leader. The timing for the project was approximately one year. It was also clear that a considerable of data collection and analysis would be involved.

Such a major undertaking requires a well-defined structure to guide and prioritize the work. Large, complex and unstructured problems are typically very messy at the beginning. Creating structure for the problem and work helps, as we fondly say, “clean up the mess”. This was done by first conducting a process and organizational assessment. Conducting such assessments is a very effective tool to identify the needed information to properly structure the problem, identify opportunities for improvement and conduct the improvement work.

As shown in Figure 3 the assessment team looked at seven (7) focus areas using a mixture of tools including document review, interviews of critical personnel at all levels in the organization, surveys and process observation Four (4) areas of opportunity were identified: throughput and quality, leadership and management, process and equipment reliability and compliance and organization and behavior.

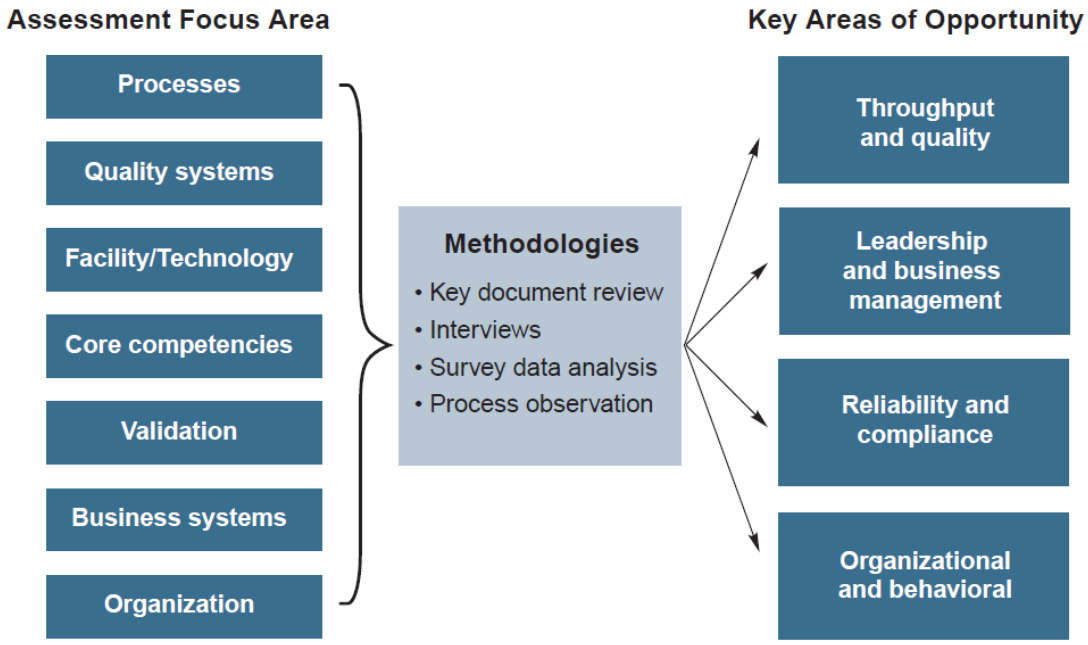


Figure 3. Elements of a Comprehensive Assessment (Source: McGurk 2004)

The supply capability of an organization depends on its ability to plan and execute not just the manufacture of product, but also wider operations including process reliability, Good Manufacturing Practices (GMP) compliance, effective leadership, clear communications, and operational metrics. To generate a targeted list of improvement opportunities in all these areas, the company first assessed a range of operational expertise including manufacturing and quality assurance (QA) knowledge, operational metric design, organizational development, and statistical process control (SPC).

The initial target list was broad, ensuring no significant opportunities were missed (Figure 3). As Linus Pauling points out, "The best way to have a good idea, is to have a lot of ideas". By uncovering causes common to the problems on the initial list, the project team was able to consolidate improvement opportunities into categories and further characterize them, aligning them with the overall goals and strategy of the organization and calculating the potential return on investment.

Phase 3: Understand Context - History, Politics and Personalities

Stock outages of pharmaceutical and biological products occur with disappointing regularity. Even newly launched products have not been immune, particularly biopharmaceuticals, which take longer to manufacture and require production operations that are more difficult to control. These shortages result in Industry and regulatory issues and bad perception by the public. As a result not meeting demand is both a financial and political issue.

The two products involved were of critical importance to the health of the company and were highly visible to senior management. The Vice President in charge of these products was concerned that the launch date and yield improvement goals would not be met. There was much risk involved.

The organizational and process assessment was very useful in understanding the context of the problem. In particular the interviews with leaders in all levels and functions provided critical information regarding the history, politics and personalities of the organization and the people working in it.

The interviews made it clear that the staff had little manufacturing experience as manufacturing was staffed by several employees who developed the drugs in R&D. Operating manufacturing requires a mindset change from developing new and innovative products and processes to operating processes to consistent and compliant manufacture. Development of skills for manufacturing excellence will have to be developed. Thus an enabling objective was to get the staff at ALL levels to a new level of performance. A significant amount of training and one-on-one mentoring would be required.

Phase 4: Develop Strategy – How to Attack the Problem

The assessment uncovered a number of gaps that impeded both the production ramp-up for the blockbuster drug and the consistent supply of the monoclonal antibody. Knowledge of these gaps greatly facilitated the development of strategies regarding how to attack the problem and the subsequent creation of tactics for implementing the strategies.

The gaps identified included:

- Need for 100% more manufacturing capacity to meet demand for the blockbuster drug
- Suboptimal manufacturing reliability for the monoclonal antibody, which stood at only 80% ‘right first time”.
- Inefficient processes for review of batch records, including review periods more than twice as long as necessary
- Inadequate measurement system for monitoring operational performance

- Insufficient leadership and interpersonal skills for cross-functional teamwork
- Inadequate training for production operators.

These gaps identified four critical focus area: Manufacturing Capability, Batch Record Review, Create Metrics and Training and Leadership. Tactics were developed to address each of the four (4) critical focus areas. The team compiled a list of specific projects and formed cross-functional teams to address them. As the projects crossed functional lines, the teams employed were also cross-functional. The plan called for the four focus areas to be worked on simultaneously. The project was of critical importance so management made the resources available. While these focus areas were part of the same project, improvement could be made in the areas independent of each other.

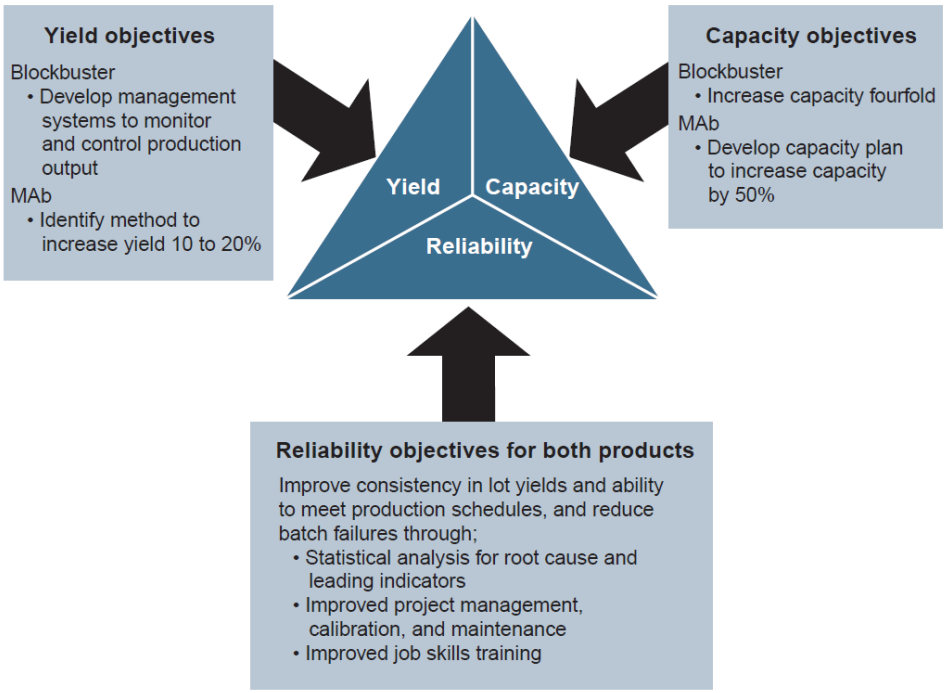


Figure 4. Manufacturing Capability Model (Source: McGurk 2004)

Phase 5: Develop and Execute Tactics

Manufacturing capability depends on the interaction of three components — capacity, yield, and reliability (Figure 4). Understanding their interaction was critical to prioritizing improvement activities. Two manufacturing product teams, one for the blockbuster drug and one for the monoclonal antibody, were formed. To address capacity, the teams conducted a bottleneck analysis of equipment, people, and training constraints.

To improve yield, the team applied statistical process control and multivariate analysis to historical production data. This method provides knowledge that cannot be derived from a few pilot, demonstration, or even validation runs (Hoerl and Snee 2020). For example, after the

monoclonal antibody production team identified process shifts in fermentation yields, they formed a cause-and-effect team that included fermentation experts and a statistician.

It is critical when analyzing production data to assess the quality of the data. The critical question is are the data "fit for use" in the problem solving venture being pursued. This is accomplished by understanding the "data pedigree". Do we know the origin of the data and the route it has taken prior to be considered for analysis? Data pedigree is defined as "documentation of the origins and history of a data set, including its technical meaning, background on the process that produced it, the original collection of samples, measurement processes used, and the subsequent handling of the data, including any modifications or deletions made, through the present." (Hoerl and Snee 2019)

Assessing data quality is particularly important in analyzing production data because the data are "observational data" collected without the aid of a planned protocol such as a statistically designed experiment (Montgomery 2019). Observational data are of lower quality for many reasons including missing variables, recording errors and poor measurement technique (Hoerl and Snee 2020). In this case the data were transcribed from batch records. Great care was taken to get the transcription done correctly and involving subject matter experts to check the validity of the data. It was concluded that the available production data were adequate for studying the production process behavior.

To improve reliability, the teams analyzed historical process variances. Variances are instances of "things that went wrong", which are referred to as "special causes" in the quality improvement world. These can include process control parameters that were not in control, changes to procedures, or any other variation from normal practice. Without process reliability, accurate and precise supply predictability remains impossible. Of course companies compensate for this unpredictability by adding manufacturing capacity— and incur additional costs. In the pharmaceutical industry, production variances and the resulting investigations pose a great threat to reliability and, in turn, supply capability.

Batch Record Review can greatly affect supply chain performance which depends as critically on the flow of required documentation as it does on the flow of product. A Lean Six Sigma approach was used to analyze the batch record review process (Snee and Hoerl 2005). A cross-functional team of production and quality personnel constructed a process map of the batch record review process, detailing bottlenecks such as excessive time spent in the queue, an overly complicated flow of records, and a lack of clarity in the company's expectations of reviewers. The team collected baseline batch release data for both products and used fishbone diagram analysis to organize the variables that would impact defects of the batch record review process. They then used Pareto analysis to prioritize the correction of defects and control charts to

measure the progress and impact of changes to the process. The use of a large collection of quality tools in the problem solving process is characteristic of statistical engineering projects.

Process Metrics are central to the effective monitoring, control and improvement of manufacturing processes. Complicated operations require clear operational definitions and accurate and timely flow of information among shop floor, planning, and operations management personnel. For example, inconsistency concerning when a particular operation is considered complete can cause enormous confusion. A batch could be deemed "done" in a number of ways: when an operator finishes making the batch, when the documentation is reviewed, when Quality releases the batch, or when it is in inventory, ready for shipment. A month or more could elapse from the time something is *believed* to be done to when it *is actually* done.

Training and Leadership Development are always required for major organizational change. This need was especially acute given the preparation for launching a blockbuster product. Approximately 70 members of the quality and operations functions underwent training in such interpersonal skills as understanding people, expressing oneself, and resolving conflict —all critical for the smooth functioning of any organization with extremely complicated and highly interdependent processes.

Members of the leadership group participated in an assessment of leadership knowledge, the results of which were compared to an extensive database and used to create ongoing leadership development plans. A significant amount of individualized and group coaching and ongoing assessment of the program's effectiveness supplemented the organizational development work.

Operator Training was needed to increase production levels with the launch of the blockbuster product, the project team undertook a detailed analysis of the manufacturing operation and its operators. A digital video camera recorded the actions of trained operators using both existing and revised standard operating procedures (SOPs) and batch record instructions for the operations. The videotapes provided both a model of appropriate behavior for operators and a forum for them to work together to develop best practices.

Phase 6: Identify and Deploy Final Solution

As discussed in the section on "tactics" several critical process changes were deployed in each of the critical focus areas: manufacturing capability, batch record review, process metrics and training and development.

Manufacturing Capability. As a result of the bottleneck analysis, the team broke the bottlenecks by purchasing needed equipment such as additional storage vessels and refrigeration capacity, identifying the personnel required for production ramp up, and cataloguing skill deficiencies — especially those in biopharmaceutical production, such as batch weighing, batch

charging, chromatography, and GMPs. After identifying the needed skills, the team set up a training program to teach them how to effectively employ these new skills in their daily work.

Analysis of the monoclonal antibody production data using a control chart identified 30% shift in yields (Figure 5). The root cause of the shifts was the media lots used in the antibody production. On investigation it was learned that there were no specifications for the media lots. The explanation was that ‘we take what the supplier sends us”. Specifications were developed for the media lots along with associated measurement methods. Yield was significantly improved when the new specifications were instituted.

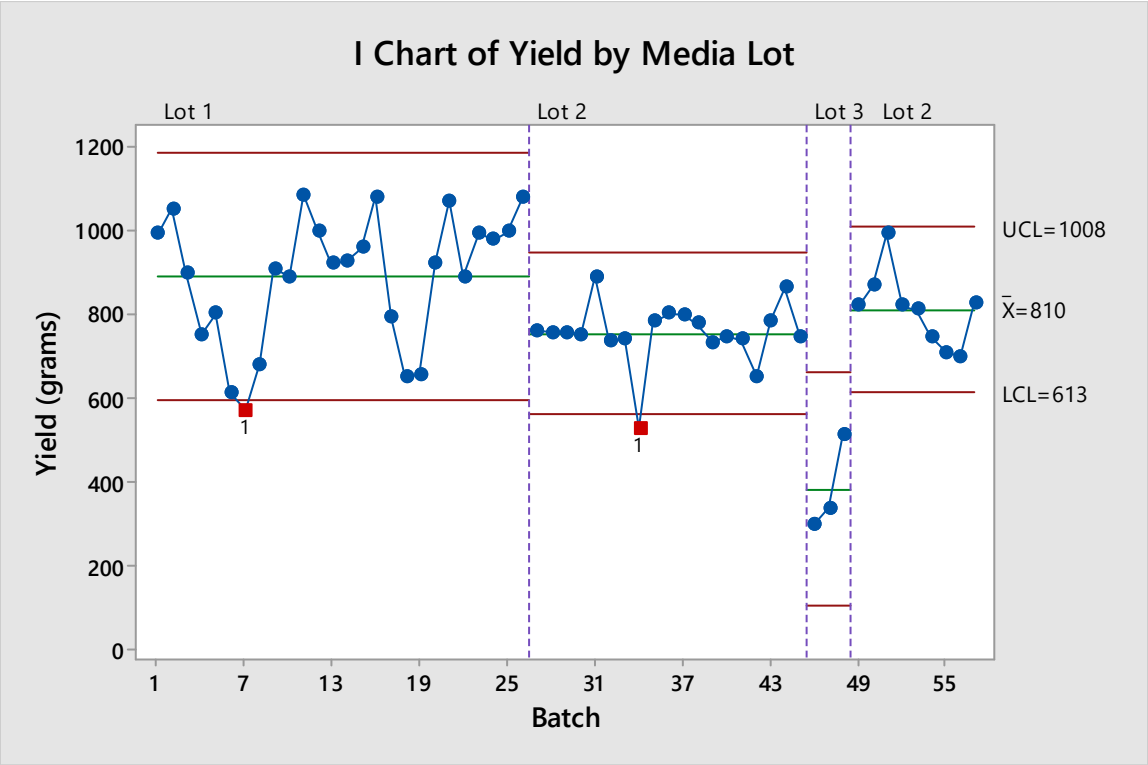


Figure 5. Control Chart of Process Yield for Production Batches

The application of similar statistical and analytical techniques to other areas, including column chromatography performance and optimum column loading, resulted in higher reliability and a 10% improvement in yield. By identifying causes and effects, these techniques focus sharply on particular problems, thereby saving time and maximizing return

The analysis of variances identified systems that were prone to problems — either mechanical problems, such as design and equipment suitability, or operational problems such as how the system was used. For example, repeated variances in the batch weighing process could reveal mechanical failures (such as inadequate scale design and installation suitability) as well as

operational failures (which can include unskilled or inadequately trained operators and poor operating documentation).

By uncovering root causes of variances, the team was able to strengthen both the mechanical and operational aspects of vulnerable systems. Most variance reduction programs fail because they do not uncover the root cause of variances, lack connection to a thorough system analysis, and poorly execute the corrective action. Given a poor definition of the problem and the lack of a system-level analysis, poor corrective action is inevitable.

Batch Record Review. As in most documentation processes, time for review is a major bottleneck. Through the use of a process map, the team established a framework for allocating the *who*, *what*, *when* and *where* of the review. The document review process which was embedded in the batch record process was divided into seven steps. The cycle time for each step was recorded for each process step associated with the review of 37 batches. The goal for cycle time was a 50% reduction which would be a major step forward and greatly increase the ability to release the product for distribution to customers

These cycle times were plotted on a Pareto chart. As we see in Figure 6 the ‘big bar” on the chart is for the review time of the manufacturing organization. Several improvements were made including the following changes:

- Target cycle times and associated monitoring procedures were developed.
- Periodic retraining of process operators was instituted and backup personnel were identified for critical positions.
- The process tracking meeting was redesigned to focus on problem identification and solution rather than data reporting and review.
- O*ne-unit flow*, a lean manufacturing concept, was instituted to manage the document review process. Individual records were submitted for review when complete rather than waiting for all the records for a batch to be complete before submitting the “batch of records” for a given production batch.

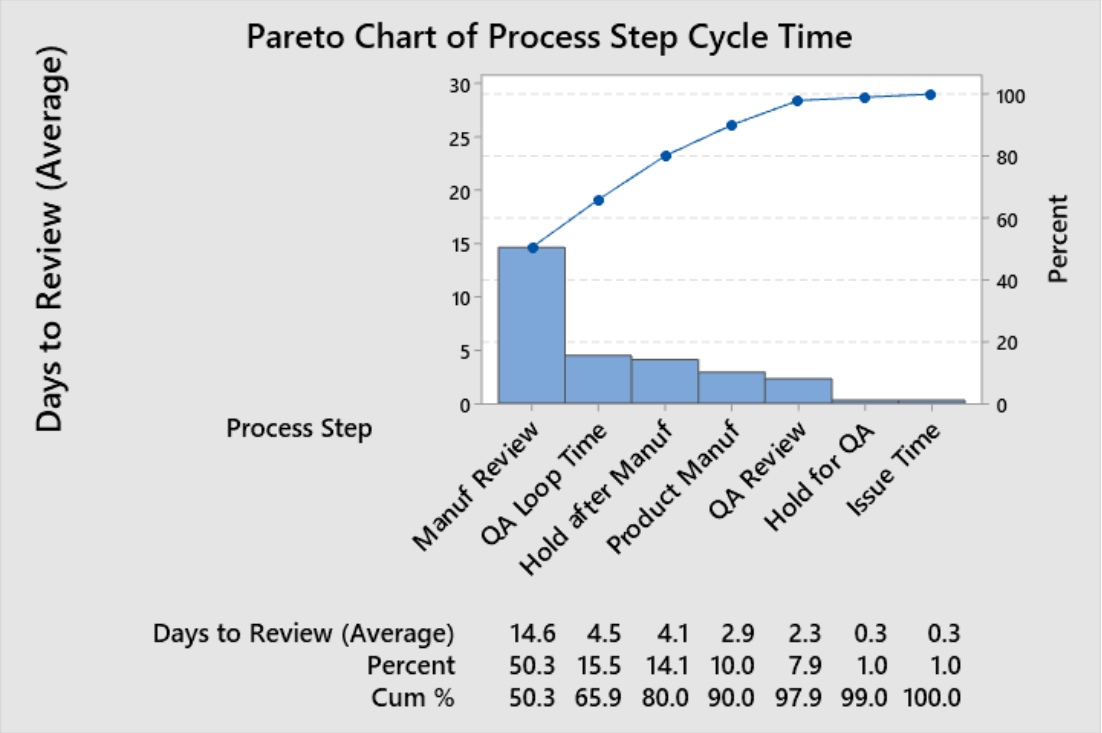


Figure 6. Pareto Chart of Process Step Cycle Time

In Figure 7 we see that when these and other improvements were made, a major drop review cycle time of 35% for one product and 55% for the other product. We also see in Figure 7 that the variation in cycle time also resulted as a result of the improvements.

Not surprisingly, this had a significantly favorable one-time impact of inventory levels and costs of approximately $5 million, especially for the monoclonal antibody with its longer cycle time. Annual operating costs were also decreased $200,000 per year.

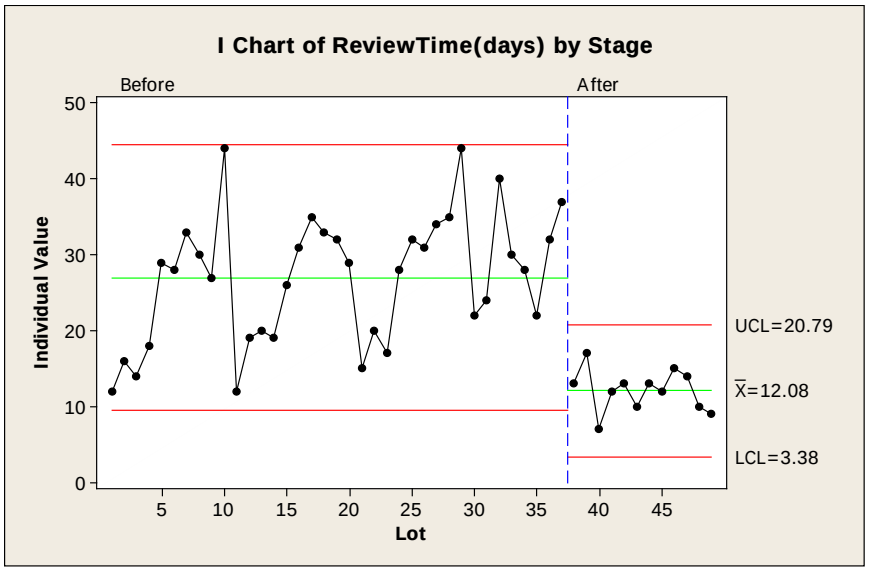


Figure 7. Cycle Time for Document Review for Product A

Better management of investigational reports for manufacturing variances and better design of the batch records themselves also improved cycle time. Further details on this document review process improvement project can be found in Snee and Hoerl (2005)

Process Metrics. To establish relevant metrics, the team first reviewed the management goals of the operational, quality, and compliance functions to align them with the goals of the entire organization. Each metric was defined to ensure clarity regarding what was being measured and to ensure that it contributed to the desired outcome (Table 1).

The measurement system consisted of both a broad set of metrics called the "dashboard" and a more detailed set called the "manager's metrics." The broader, summary-oriented dashboard serves site management. The more targeted manager's metrics enable functional managers to gauge improvements in their respective areas. For example, the dashboard metric for batch-record release time indicates release times for the entire operation. The manager's metric, however, encompasses only the release times for the batch records in that manager's area. The measurement system immediately established a common understanding and communication of performance. Moreover, it provided a platform for improvement in numerous areas.

Table 1. Critical Metrics for the Operational, Quality, and Compliance Functions
(Source: McGurk 2004)

Operational	Quality	Compliance and Documentation
Product cycle times	Batch record release cycle times	SOP revisions
Inventory levels	Root-cause tracking	Investigational reports
Product supply plans	Variance tracking	Commitment tracking
Costs: direct and assessed	Training schedules and execution	Revalidations
Overhead	Environmental monitoring actions	Training activities
Yields		

Regular Management Review of Metrics was instituted to ensure the solutions deployed were effective and sustainable over time. Executive and manager metrics dashboards were implemented and regularly reviewed monthly or weekly depending on the metric and the group (Executive or Manager) performing the review. More than 50 parameters including compliance, scheduling, training and costs were monitored at various levels in the organization. Regular management review is essential for a process to operate as desired over time.

Training and Leadership Development. Approximately 70 members of the quality and operations functions participated in the training and leadership development workshops. Approximately 95% of the participants reported they were "comfortable in applying the new skills." Forty people from that group received additional training in leadership, using case studies constructed from actual company experiences.

Participants reported "a common language and a shared understanding of concepts" that could be used in their day-to-day activities. Most importantly, this training ensured the thorough integration of the operational and organizational elements designed into the entire project at its inception.

Operator Training. In the operator training group meetings the experienced operators and process experts discussed and analyzed the operators' actions. The training resulted in clearer SOPs and greater consistency of action from all operators without the interference of shop floor noise, production gowns, or the fast pace of production. The operators themselves confirmed its effectiveness in their feedback.

Reaping the Benefits - Process and Business Results

The project team's efforts produced dramatic improvements for both the blockbuster product and the monoclonal antibody. For the blockbuster, the breaking of bottlenecks increased capacity by more than 100%. Batch record review cycle time was reduced 35% resulting in a one-time inventory decrease of $5 million. Meanwhile, reliability improved in five manufacturing systems, including the maintenance system. SPC ensured the continued and accurate monitoring of critical process variables through the dashboard and manager's metrics.

For the monoclonal antibody, the company attained 50% more capacity through optimized production scheduling: the proper sequencing and usage of equipment and utilities and the flexibility of operating personnel who were now trained to handle a variety of tasks. Statistical tools helped improve yield by 20%, and reliability improved in weighing systems and two other manufacturing systems. As with the blockbuster, SPC ensured continued monitoring.

The overall batch review cycle time of both products was reduced 35-55% depending on the product. The improvements included enhanced document flow, improved operator training, a redesigned batch record, and streamlined investigations. These benefits were sustained, the process and structure were monitored and reinforced, ensuring the changes are taking root, operational and organizational improvements became more integrated, and the improvement program's momentum is sustained. Taking into account yield increases, a reduction in safety stock of 10%, material savings, and cost avoidance, the improvement of so many areas and systems produced a *tenfold return* on investment in the project.

The real return, however, was even more significant. After the project, a much improved operating group exists with the confidence of the company to deliver on other challenging opportunities. The measure of that confidence? Following the successful launch of the blockbuster product, the biopharmaceutical group was cited for a global corporate award for their crucial role in the supply of the blockbuster.

So What Have We Learned?

This exercise has highlighted a number of things regarding the value and deployment of Statistical Engineering including:

- A general framework such as that shown in Figure 1 is useful in solving large, complex unstructured problems
- An organizational and process assessment is a very useful tool for providing the needed structure and context for the project
- One or more multi-shilled teams are needed to be successful. It is very rare that one or two persons have all the experience, skills and knowledge to solve such problems
- Assessment of the data pedigree is essential to ensure that the data are trustworthy and fit for use.
- A variety of tools, technical and non-technical are needed in such projects

This case study affirms that Statistical Engineering is a very useful approach to solving large, complex and unstructured problems. The methodology provides the philosophy, concepts, methods and tools needed to bring the project team up the learning curve quickly and develop useful and timely solutions. As the approach is used in an organization it can be customized to the culture of the organization enabling the approach to be used broadly across the organization.

Opportunities for Pharma and Biotech to Use Statistical Engineering

Pharmaceuticals and biotech have many opportunities for improvement that involve large, complex problems that can utilize statistical engineering in their solution. These opportunities include both the building of new processes as well as the improvement of existing processes.

Major Enhancement of Legacy Processes, processes that have been in operation for a long time, have been often ignored as opportunities for improvement unless a major problem occurs. Even though these processes have been in compliance and producing in-spec product, these processes frequently become wasteful and inefficient over the years. A process and organizational assessment like the one discussed in this case study is a good way to identify the source and process and financial value of improvement opportunities. Experience has shown that the financial payoff can be large having a return on investment of 4:1 to 10:1 and more in many cases.

New Initiatives such as Quality by Design (ICH 2009, Snee 2019), Continued Process Verification (FDA 2011) and Continuous Manufacturing are major undertakings that can be classified as large, complex unstructured problems that can be addresses by statistical engineering. These initiatives typically involve multiple sites, multiple functions and multiple processes and products. The structure provided by statistical engineering framework (Figure 1) speeds up implementation as teams get off to a good start developing and implementing strategy and getting the initiative producing useful results quickly.

Predictive Analytics and Big Data Problems is another major opportunity. The literature contains several mentions of the use of this methodology in the Pharmaceuticals and Biotech industries. These projects typically involve multiple data sets, from different organizations with different management agendas as well as multiple sites, functions, processes and products. The result is a large, complex problem. These problems are typically messy and involve several organizational and political issues beyond those of the data analytics considerations.

These opportunities and other major initiatives resulting from customer, environmental and regulatory issues can be profitably addressed using the concepts, methods and tools of statistical engineering. The result is better outcomes being created in a timely fashion.

Acknowledgement

The author is pleased to acknowledge the guidance and support of Thomas L. McGurk which was very helpful in preparing this case study.

References

Hoerl, R. W. and R.D, Snee (2017) "Statistical Engineering – An Idea Whose Time Has Come", *The American Statistician*, Vol. 71, No. 3, 209-219.

Hoerl, R. W. and R. D. Snee (2019) "Show Me the Pedigree: Part of Evaluating the Quality of Data Includes Analyzing Its Origin and History", *Quality Progress*, January 2019, 16-23.

Hoerl R. W. and R. D. Snee (2020) *Statistical thinking: improving business performance, 3rd Edition*, John Wiley and Sons, Hoboken, NJ.

ICH (2009) "Harmonized Tripartite Guideline: Pharmaceutical Development, Q8, Current Step 4 Version", International Conference on Harmonization, November 10, 2005

ISEA Free membership to the International Statistical Engineering Association (ISEA) is available at www.isea-change.org. Membership provides ISEA members access to the organization's *Statistical Engineering Handbook*, as well as publications and presentations on the subject.

McGurk, T. L. (2004) “Ramping Up and Ensuring Supply Capability for Biopharmaceuticals, *BioPharm International*, Volume 17, No. 1, January 2004.

Montgomery, D. C. (2019), *Design and Analysis of Experiments*, 9th Edition, John Wiley and Sons, New York, NY.

Snee, R. D. (2016) “Adjust, Adapt and Advance: An Enhanced Version of Quality by Design—a Risk-Based, Dynamic Approach to Improving Products and Processes”, *Quality Progress*, May 2016, 34-41.

Snee, R.D. and R. W. Hoerl (2002) *Leading Six Sigma: a Step-by Step Guide Based on Experience with GE and other Six Sigma Companies.* New York: FT Prentice Hall.

Snee, R. D. and R. W. Hoerl (2005) *Six Sigma Beyond the Factory Floor: Deployment Strategies for Financial Services, Health Care and the Rest of the Real Economy*, Pearson Prentice Hall, Upper Saddle River, NJ.

Snee, R. D. and R. W. Hoerl (2018) *Leading Holistic Improvement with Lean Six Sigma 2.0*, FT Prentice Hall, New York, NY.