Exploratory Data Analysis in Quality-Improvement Projects

JEROEN DE MAST and ALBERT TRIP

Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA)
Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands

The aim of this paper is to provide a prescriptive framework for exploratory data analysis (EDA) in quality-improvement projects. The framework is developed on the basis of a large number of real-life applications. The three steps of EDA are described: display the data, identify salient features, and interpret salient features. Graphical display of data, Shewhart's assignable causes, the maximum entropy principle, abduction, and explanatory coherence all are part of the resulting framework. Furthermore, the roles of probabilistic reasoning and automatic statistical procedures in EDA are discussed.

Key Words: Discovery; Entropy; Exploratory Data Analysis; Graphical Data Analysis; Hypothesis Generation; Pattern Discovery.

THE DISTINCTION between *exploratory* and *confirmatory* data analysis is mostly attributed to Tukey (1977). As Tukey pointed out, confirmatory data analysis (CDA) is concerned with testing a prespecified hypothesis. For instance, if an inquirer suspects that certain factors have an effect on a characteristic, he could collect experimental or observational data, estimate the parameters in a regression or other model, and calculate p-values, thus establishing which factors have an effect and modeling the relationship. But before a CDA can get off the ground, the inquirer must know which data to collect, and for that he must know which hypothesis he is willing to investigate. For instance, before the inquirer can design an experiment to investigate the effects of certain factors, he must have identified these factors in the first place.

Hypothesis generation is a different functionality than hypothesis testing and estimation. There are a number of approaches to generate hypotheses, such as brainstorming, making an inventory of process know-how, and exploiting suggestions from analogous problems (De Mast and Bergman (2006) give an overview). Exploratory data analysis (EDA), during which data are screened for clues, is one of these approaches. Where estimation, modeling, and hypothesis testing could be said to be the purpose of CDA, hypothesis generation is the purpose of EDA.

Having contrasted EDA with CDA, to delineate the subject more clearly, we contrast both CDA and EDA with descriptive data analysis (DDA), the summary of a dataset in a number of descriptive statistics. DDA is concerned with the presentation of data to reveal salient features. This is done by suppressing uninformative features of the data so as to make the important features stand out more clearly. The summary of a dataset in a number of summary statistics, such as average and standard deviation, is a matter of pruning simply because dealing with the full complexity of the dataset is far beyond human cognitive abilities. Also, the use of tables and graphs and other descriptive statistics is designed to match the salient features of a dataset to human cognitive abilities (Good (1983)). EDA goes somewhat further than descriptive statistics in that its aim is not merely to present salient features of a dataset, but in addition to speculate and formulate hypotheses that have the potential to explain these salient features. CDA goes further than DDA because it goes beyond summary statistics (which give information about the sample) to estimates, models, or predictions for a target population.

Dr. de Mast is a Senior Consultant at IBIS UvA and Associate Professor at the University of Amsterdam. He is a member of ASQ. His email address is jdemast@science.uva.nl.

Dr. Trip is Senior Consultant at IBIS UvA. He is a member of ASQ. His email address is albert.trip@saralee.com.

The literature on EDA is less elaborate than the literature on CDA, both in pure volume of texts devoted to the subject and in precision and depth of its theoretical development. Good (1983) speculates that "... perhaps EDA is more an art, or even a bag of tricks, than a science." But yet, we have to teach this art to practitioners. It is the purpose of this article to develop a number of explicated principles for EDA that can be taught to practitioners and statisticians to help them master this art faster (or at least to provide them with a better ordered bag of tricks). The empirical basis for our theories is formed by a review of a large number of applications of EDA from our consulting experience. Our study design is not very formal, but rather exploratory. The majority of EDA examples were part of Six Sigma quality-improvement projects and were conducted by Black Belts supported by experienced statisticians. Furthermore, we have drawn from examples of EDA published in the literature. Studying how EDA is performed in practice and by reconstructing the line of reasoning and the goals, we identified a number of principles, A–D, which form a prescriptive framework for EDA. We discuss these principles in the next sections. They are presented on the basis of a representative sample of the real-life EDA applications that we studied. In presenting these case studies, we simplified in some cases the account to some extent for reasons of clarity and brevity.

Throughout the paper, the subject of study is limited to EDA applications in quality-improvement projects (see De Mast (2003) for a framework of quality improvement based on statistical methods). We have in mind datasets of a relatively small scale (say, some 25 to a couple of thousand observations), where EDA is done by the inquirer with the help of elementary statistics software (as opposed to the large volumes of data that make the use of data-mining techniques a necessity).

The Purpose and Process of EDA

De Mast and Bergman (2006) place EDA in the wider context of hypothesis generation from the viewpoints of philosophy of science (discovery), artificial intelligence (problem solving), and the medical sciences (diagnosis). They conclude that guidelines for hypothesis generation should not consist of algorithms, but should have the form of heuristics. This implies that EDA is informal (that is, reasoning does not methodically follow codified rules), flexible (there is no preconceived plan; instead, the path of investigation emerges in an interaction between inquirer and data), and speculative (pursuing hypotheses that have potential, not hypotheses that are true).

Before discussing the process of EDA, we address its purpose. The aims of EDA are described in the literature in phrases such as "to generate hypotheses", "to generate clues", "to discover influence factors", and "to build understanding of the nature of the problem". Problem solving in the paradigms of statistical thinking and statistical improvement strategies (such as Six Sigma; De Mast (2003)) requires that the problem be parameterized. This means that problems are framed in terms of variables (called CTQs, Ys, or KPIs in Six Sigma and other approaches) and likewise are the causes (Xs, sources of variation, influence factors). The study centers around potential relationships among these variables. Our first principle of EDA formulates this Purpose.

A. *The purpose of EDA is the identification of dependent (Y-) and independent (X-)variables that may prove to be of interest for understanding or solving the problem under study.*

Note that, thus defined, EDA is understood to be only a part of the Data-analysis movement as promoted by Tukey (see Mallows (2006) for a recent discussion). Data analysis seems to cover all of applied statistics in general and is much wider than EDA.

One could symbolize EDA's pursuit as follows. The data are taken to be the sum of a number of components,

$$Y = Y_1 + Y_2 + Y_3 + \cdots + Y_k, \qquad (1)$$

or the aggregate of the effects of a large number of causes,

$$Y = E_1 + E_2 + E_3 + \cdots. \qquad (2)$$

In the second situation, the data are seen as the result of a number of causal effects, where each E_i is the effect of a single or several causal factors, $E_i = f(X_{j1}, X_{j2}, \ldots, X_{jn})$, and the inquirer wishes to discover these X's.

In the first situation (Equation (1)), the data are taken to be the sum of a number of components. For example, the total number of scrap parts per week (Y) could be decomposed as a sum of weekly scrap parts per product type ($Y_1, Y_2, \ldots, Y_k$) or as a sum of scrap parts per production line. Likewise, if the data are throughput times, these could be decomposed into a sum of throughput times per process step. Especially when the inquirer has not yet a clear and focused parameterization of the problem he studies, he could apply EDA to find such a decomposition,

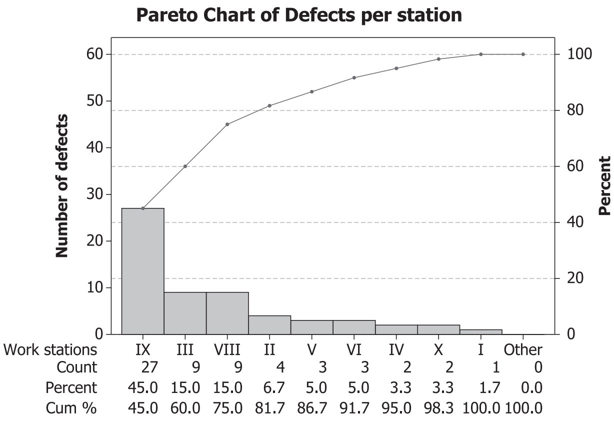

FIGURE 1. Defects Per Work Station.

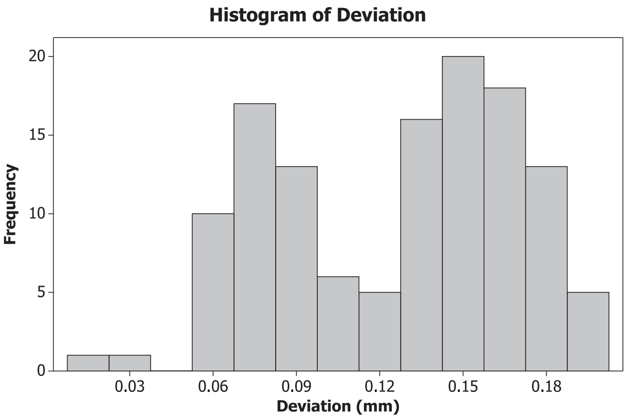

FIGURE 2. Distribution of Eccentricity of Pins on Cellphone Components.

which could help him better understand and focus the problem. The first example below illustrates the use of EDA to identify interesting Y variables, while the second example shows how EDA is used to discover X variables.

Example 1: Defect Reduction on an Assembly Line

This case is taken from Bisgaard (1996). An assembly line for motors produced an unacceptably high number of defective motors. Extensive records on scrap were available, but only after encouragement from an external consultant were these data studied. Presented in the form of tables, these data conveyed little useful information. Pareto charts categorizing the defects by type proved interesting. But it occurred to the improvement team that it would be even more interesting to categorize defects according to which of the 10 work stations on the assembly line the defects originated from. The team invested time in tracing defects to their sources, and thus managed to produce the Pareto chart in Figure 1. The large number of defects originating in station IX catches the eye. Discussion with the operator on station IX brought to light that major changes in the design of the motors had resulted in considerably more operations to be performed on station IX. Moreover, the layout of the workstation had not been adjusted to the new operations and was no longer efficient. Thereupon, workstation IX was redesigned, and several operations performed at that station were moved to station VIII.

The input for EDA were the data on defects. A categorization by defect type did not result in interesting discoveries, but a categorization by workstation helped focus the problem: the problem was reframed from “too many defects” to “too many defects from station IX”. Thus, EDA helps go from a broad problem description (aggregating many different issues and aspects) to a focused description of the problem (pinpointing the one or a few issues that dominate the others). Given this focused problem definition, the causes and solution were easily found (but by other means than EDA).

In this case, EDA helped identify potentially interesting Y-variables (as symbolized in Equation (1)). The second case study illustrates how EDA helps identify X-variables.

Example 2: Eccentricity of Pins on Cellphone Components

A quality-improvement project aimed to reduce the eccentricity of pins on components of cellphones. The histogram in Figure 2 shows the distribution of 125 measurements (where eccentricity is measured as an absolute deviation from the center). The data were collected during outgoing quality inspection. The project leader recognized the bimodality of the distribution, conjectured that the data could be interpreted as stemming from two populations, and consulted the operators whether they knew what caused the distinction in two populations. The operators conjectured that the two groups might correspond to the two different molds that were used in the injection molding process that produced the components. It was not possible to trace which data points corresponded to which mold, but a new data collection confirmed the hypothesis that the two molds gave deviating results. The result of the EDA in this case is the identification of potential causal influence factors, namely the influence of properties of the molds.

The identification of the variables into which the data decompose (the Y_i) or the variables that determine the causal structure underlying the data (the X_i) is the purpose of EDA. EDA seeks to do so by studying F_Y, the distribution of Y (or in practice: a realization of F_Y). Note that this implies that EDA can identify only variables that are associated to variance components of the distribution of the data Y. Put more simply: only factors that actually vary during data collection can be identified with EDA. Factors that remain constant during data collection (machine settings, for instance) leave no traces in the data and must be identified using procedures other than EDA (De Mast and Bergman (2006)).

In the process of EDA, three steps can be discerned:

1. Display the data.
2. Identify salient features.
3. Interpret salient features.

In Example 2, for instance, the inquirer made a histogram (display the data), recognized the bimodality (identify salient features), and conjectured with the operators that properties of the molds could explain the bimodality (interpret salient features). The next sections discuss the principles that apply to these steps.

Display the Data

The first step in EDA is to display the data in such a way that we can maximally exploit the pattern-recognition capacities of our brains. This pursuit is explicit in the work of Chernoff (1973), who uses drawings of faces to represent multidimensional data; Ehrenberg (1981), who gives guidelines for adapting tables to the perceptive capabilities of humans; Cleveland and McGill (1984); and others. Many authors (Good (1983), Hoaglin et al. (1983), Bisgaard (1996)) have claimed that graphical presentations are to be preferred in ED, because they have the power to reveal to the inquirer what he did not expect to see beforehand. The information in the data that is relevant for EDA is contained in their distribution. Graphical presentations tend to show the data's distribution in a way that human brains can handle. A table of the raw data, to the contrary, is too complex for human brains, whereas tables of aggregate statistics eliminate (components of) the data's distribution, thereby losing information that is potentially crucial for EDA. Note how, in both Figures 1 and 2, the relevant information in the data is captured in their distribution. Especially in Example 2, summary statistics (mean, standard deviation, quartiles, etc.) would not have brought across the salient feature as powerfully as the histogram. Alternatively, the inquirer could have studied a plot of the empirical cumulative distribution function (an ogive, or a probability plot; see Snee and Pfeifer (1988)). Our second EDA principle is as follows.

B. *Display the data such that their distribution is revealed.*

Often, the data are not just Y-data, but have an additional structure, such as strata or a time order, as is illustrated in Example 3.

Example 3: Throughput Time of Invoicing Process

Figure 3 is from a project aiming to reduce the throughput time of an invoicing process. The project leader collected throughput times from each of the five sites where the process is executed and made the box plots in the figure. Noting the large differences between the distribution of the throughput times in the five sites, he compared the procedures used at the various sites. He noticed that the poorly performing sites processed invoices in batches (once per 15 days), whereas the better performing sites process invoices in smaller batches or even immediately.

Note that the results of EDA sometimes appear trivial: is it not obvious that processing invoices in batches every 15 days leads to long throughput times

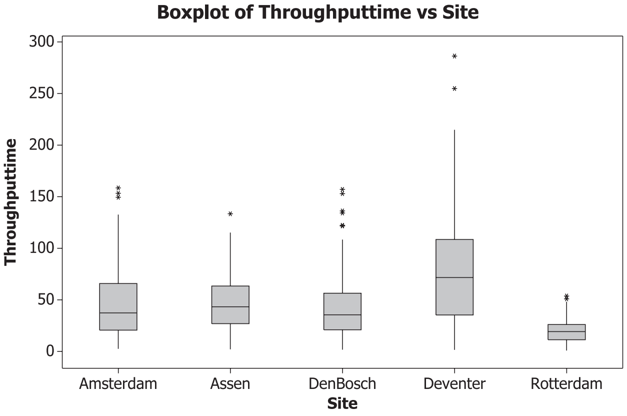

FIGURE 3. Throughput Times of Invoicing Process by Site.

(as follows from the principles of Lean, for instance)? But the fact that discoveries seem trivial in retrospect should not be a reason to dismiss such analyses as superfluous. Often, assumptions are completely taken for granted and the inquirer's mind has become conditioned. Blindness for the obvious is the result. The usefulness of simple tools consists in part of their power to force the obvious even onto a mind that is not open for it.

The case shows that the strata in the data set (*sites*) play an important role in the way the data are displayed. This can be understood by observing that these strata act as a sort of "container variable", which confounds the effect of many variables the values of which coincide with the stratum levels. As a result, the space of all variables is split into a class of variables that only vary between groups and a class of variables that also vary within groups. We formulate this as a refinement of principle B.

B1 (Stratified Data). *Display the data such that both their within-stratum and across-stratum distributions are revealed.*

Suitable options include box plots per stratum and individual value plots. A similar consideration shows us that also time order can play this role of container variable, confounding the effects of many variables.

B2 (Data Plus Time Order). *Display the data such that their distributions within time intervals and across time are revealed.*

If a single datum per time unit is available, a time-series plot (and Derivatives, such as the individuals control chart) is the main technique to use. If a sample is collected on each time unit, a box plot (Iglewicz and Hoaglin (1987)) or an individual value plot per time unit are preferred to the $(\overline{X}, R)$-control chart because the latter does not show the distribution within time units.

In case one deals with multivariate (especially high dimensional) data, one needs another refinement.

B3 (Multivariate Data). *Project the data onto a* 2- *or* 3*-dimensional subspace and display the data's distribution over this subspace.*

If the identification of salient features is to be done by the inquirer (and not by an automatic procedure), the projection of the data onto a low-dimensional subspace is a necessity in order that human perception can cope with them. The distribution of the data over the low-dimensional subspace can be visualized by a scatter plot or a contour plot of the empirical density. Several procedures are available to find suitable projections.

- Principal components analysis (and related procedures, such as factor analysis and cluster analysis applied to variables): can be used to find suitable 2- or 3-dimensional subspaces on which to project the data. These subspaces are spanned by linear combinations of the original axes of the data space. Principal components analysis selects a subspace onto which to project the data such that the proportion of total variation accounted for by the projected data is maximal.
- Projection pursuit: helps find 2- or 3-dimensional subspaces onto which to project the data, in such a way that salient features in the data are maximally preserved. Projection pursuit does so by selecting from the set of all possible projections the one that maximizes an index of "interestingness". The original index of interestingness proposed by Friedman and Tukey (1974) was designed to reveal clustering. More general indices relate interestingness to nonnormality or minimal entropy (based on similar argumentation, as we will give below for our principle C); see Huber (1985) and Jones and Sibson (1987).

Identify Salient Features

Having displayed the data such that their distribution F_Y is revealed, the second step of EDA is to identify features and properties of F_Y that are salient (such as the bimodality in Figure 2). To provide inquirers with heuristics that have practical value, we must give the term *salient* operational meaning. Below we substantiate the following principle.

C. *Assuming a neutral reference distribution, look for deviations from this reference distribution.*

We provide two heuristics to define such a neutral reference distribution. The first is Shewhart's theory of *assignable causes of variation.* The basic principle for their identification was formulated by him as "... our clue to the existence of assignable causes is anything that indicates nonrandomness" (Shewhart (1939), p. 26). This statement is based on a line of reasoning that is expounded in Shewhart (1931, pp. 121–162). If variation is the aggregate effect of a constant system of causes, none of which predominates the others, the variation will have a normal distribution. Such variation does not give clues for the identification of variables. If one of the causes of variation is dominant, it is an assignable cause, which means that it may leave traces in the data (in the form of deviations from normality) that enable its identification.

Unraveling this line of reasoning, the steps in the argumentation are:

- The data are an additive aggregate of effects (as in Equation (2)).
- Addition of effects dilutes the individual effects (which we seek to identify).
- Addition of effects results in a more and more normal distribution (because, by central limit theory, the convolution of two distributions is closer to normal than the least normal of the original distributions).
- Thus, normality is seen as a byproduct of dilution of effects through addition, and the heuristic says that if F_Y is normal, it is typically uninformative for the identification of underlying effects.

(The last two steps in this argumentation resemble Huber's (1985, Section 5) line of reasoning in a similar context). Typical deviations from normality that an inquirer encounters are bi- or multimodality (corresponding to clusters and outliers in the data) and discontinuities in the density function (corresponding to edges in histograms and scatterplots). We formulate our heuristic as follows.

C1. *Look for deviations from normality.*

The central limit theorem plays an important role in associating normality to dilution of effects. A principle that is not based on the central limit theorem and that is more generic is based on information theory. Following the line of reasoning of Shannon (see Berger (1988)), the uncertainty represented by a probability density $f(x)$ is its entropy $H(f) = -\int \log f(x) dF(x) = -\mathrm{E} \log f(x)$. The concept of entropy is used in various manners in contexts related to hypothesis generation: as a basis for prior distributions in Bayesian pattern discovery (Brand (1999)), as an index of "interestingness" in projection pursuit (Jones and Sibson (1987)), and in the "maximum entropy principle" (Jaynes (1957), Good (1963), and Bard (1988)). This last principle can be described as follows.

Suppose that an inquirer has some information about the distribution of a variable, perhaps one or a few moments, and possibly that it is nonnegative. The maximum entropy principle tells the inquirer in that case to assume the probability distribution f that has maximal entropy $H(f)$ among all distributions satisfying the constraints posed by the information on moments and its support. For a discrete distribution $(p_1, p_2, \ldots, p_k)$ with a finite number k of values, the maximum entropy distribution is the uniform distribution $p_i = 1/k$ for $i = 1, \ldots, k$. The real-valued maximum entropy distribution under the constraints that $\mathrm{E}X = \mu$ and $\mathrm{E}(X - \mu)^2 = \sigma^2$ is the normal distribution. For nonnegative distributions under the constraint that $\mathrm{E}X = \mu$, one finds the exponential distribution. The maximum entropy principle produces the most 'neutral' choice of distribution, in the sense that it is the uniquely maximally noninformative distribution that is consistent with the known constraints (where 'noninformative' is based on the information theoretic definition of information as $-H(f)$).

The relevance for EDA is that this principle provides the inquirer with a *neutral reference distribution* (such as the normal or exponential distribution, depending on the situation). If the distribution F_Y of the data deviates from this reference distribution, this is interesting because it means that F_Y has a smaller entropy than the reference distribution, suggesting that F_Y contains more information than just the information about the support and first moments. Alwan et al. (1998) use this idea to develop an information theoretic framework for statistical process control. They provide neutral reference distributions for a number of situations, and they measure deviation from this reference distribution by the Kullback–Leibler function (see Alwan et al. (1998)).

C2. *Assuming a neutral reference distribution (in the sense of maximum entropy), look for deviations from this reference distribution.*

Note that heuristics C1 and C2 are equivalent in the case of real-valued data with known (or estimated) first two moments, be it that they are based on essentially different lines of reasoning. Note also that both principles are heuristics and not algorithms. There is no intrinsic reason that they should work and, in fact, many deviations from normality or the maximum entropy distribution will not be informative.

In case the dataset has more structure than just 1-dimensional Y-data (strata, time order, multiple dimensions), one needs a number of refinements to principle C. These refinements are based on the idea that the neutral reference case implies that the data are independent and identically distributed (i.i.d.) across strata, time, and against variables.

C3 (Data Plus Time Order). *Look for deviations from i.i.d.*

The principle is illustrated from the following example.

Example 4: Excessive Variation in a Cutting Process

This case, presented in detail in Bisgaard (1988), concerns a process in which products are conveyed to a knife by a belt. The project tackled the problem of excessive variability in the distance between individual products, which resulted in deviations in the dimensions of the cut products. A factorial experiment did not result in breakthroughs because the studied factors had no relevant effects. But residual plots showed that the distance between products has a cyclical pattern in time. After much but fruitless detective work, the breakthrough came from the autocorrelation function, which showed that the cycle has a period of 40 or 80. One of the operators hypothesized that perhaps the number of products on one loop of the conveyer belt is 80. This led to studying the belt in more detail. Subsequent experimentation proved that the belt was the cause of the problem and, in particular, the belt's flexibility.

The i.i.d. situation is taken as the neutral reference. Deviations from i.i.d. indicate that the data are potentially informative for EDA purposes. As stated before, time acts as a container variable, confounding the effects of many X-variables that change over time, but were not measured. Deviations from i.i.d. could have the form of one or a few change points (such as jumps in the mean) or of a continual evolution (a trend, or cyclical patterns). The next principle is the analogue for the situation of stratified data.

C4 (Stratified Data). *Look for between-strata differences in distribution.*

Example 3 (Throughput time of invoicing process) illustrates the principle. The next refinement of principle C. reads as follows.

C5 (Multivariate Data). *Look for correlations among variables.*

The same idea (that the neutral reference distribution is i.i.d.) applied to multivariate data implies that the variables are independent. Correlations among variables indicate that the dataset is potentially informative and that the inquirer might discover which variables may be suitable Y-variables and which variables may be X-variables.

A final variant of principle C combines data with context knowledge.

C6. *Look for discrepancies between a priori perception and the data's distribution.*

If the data's mean, spread, or another feature of their distribution is at odds with what the inquirer expected to find, that is a salient feature, as the following example illustrates.

Example 5: Bank Notes Staying Behind in a Cash Center

In a bank's cash center, bank notes are counted, the corresponding amount is credited to the owner's account, and the bank notes are shipped to the National Bank. There is a single daily shipment to the National Bank and bank notes that are not yet through the process at the moment of shipment stay in the cash center for an additional day. The bank loses one day's interest on this amount.

Aiming to reduce the daily amount of money that does not make the shipment to the National Bank, an inquirer studied a dataset showing the daily amounts of money staying overnight in the cash center. The data showed few salient features, except that the amount of money staying behind was never near zero, but always substantially higher. This struck the inquirer as odd because he had expected that, on quiet days, all bank notes would have been processed long before they needed to be shipped. A discussion with various people involved pinpointed the cause: the cashier was told long ago always to keep around two million euros' worth of bank notes in reserve. This policy, now outdated, was immediately corrected.

Interpret Salient Features

The third step in the process of EDA, the step from identified salient features to hypotheses, is what turns descriptive statistics into EDA. Salient features in the data are the fingerprints of the effects of variables; upon identification of salient features, it is up to the project leader to relate them to possible variables. How people discover things is a topic on the intersection of artificial intelligence and the cognitive sciences. Having long denied that there could be such a thing as a logic of discovery, modern philosophy of science is making progress on this subject; see Thagard (1992), for example.

The simplest form of discovery is purely data driven and has the inquirer generalize features found in the data. For example, upon observing that, as long as records go back, the sun has risen each morning, the inquirer could generalize this pattern and hypothesize that the sun will rise each morning.

A form of discovery that often leads to more interesting hypotheses is explanation driven and it is called *abduction* (originally introduced by Peirce (1931–1935, CP 5.189), but see Niiniluoto (1999) or De Mast and Bergman (2006) for an introduction). Loosely said, abduction means that the inquirer compares conceptual combinations to his observations until all the pieces seem to fit together and a possible explanation pops up. The driving principle is explanatory coherence (Thagard (2004)), which could be said to be the extent to which the pieces fit together and is based on the extent to which an idea explains a wide range of observations, is consistent with context knowledge, and is simple (meaning: parsimonious, with only a limited number of parameters or side assumptions). The observation that the sun rises each morning combined with other observations and ideas could lead to the notion of the solar system because the hypothesis of celestial bodies revolving around a sun is a simple theory, which explains the phenomenon of day and night, and is coherent with a lot of other observations.

Coherence-driven discovery, finally, consists of reasoning to overcome apparent contradictions (Magnani (2000)). Copernicus apparently constructed the heliocentric model of the solar system specifically to resolve contradictions in the then prevailing geocentric model of Ptolemy (Thagard (1992), p. 196).

Note that the term *abduction* is used rather than *induction* because the latter is reserved for a complete inference (resulting in a refuted or corroborated hypothesis). For example, the whole sequence of hypothesis generation (possibly by abduction) followed by empirically testing the hypothesis, leading to a refutation or acceptance of the hypothesis, is an example of an inductive inference (see Maher (1998)).

It should be clear that the third step in EDA—interpretation of salient features—heavily depends on context knowledge. Interpretation of salient features considers how data are connected to things in the world and thus requires an ontology of what sort of things are in the world (and this, Mulaik (1985) argues, makes this the work of the scientist [or in our context, the engineer] instead of the statistician). The practical ramification for inquirers is that identified salient features in datasets should be discussed with people who have intimate knowledge of the process under study. Example 2 (eccentricity of pins on cellphone components) illustrates the point: upon identification of the bimodality of the distribution, the inquirer called in the help of the operators to arrive at the hypothesis that properties of the molds might be causal influence factors. We formulate the fourth EDA principle:

D. Identified salient features should be paired with context knowledge in order to interpret them.

The Use of Automatic Procedures in EDA

It is hard to conceive that the third step in the process of EDA (the interpretation of salient features) can be automatized. The display of the data, however, and the identification of salient features could be aided by automatic algorithms (think of procedures for cluster analysis, runs tests, outlier detection, and the various tests for assignable causes in control charts). The use of such algorithms has various benefits:

- Humans tend to see too many patterns (that is, they tend to mistake artefacts of noise as a pattern). Guidelines, such as control limits and runs rules in a control chart, give the inquirer a sense of what degree of salientness could be expected in mere noise.
- In high dimensional or large data sets, human perception gets lost, and there is not really an alternative for relying on automated procedures, at least as a first step (cf. data mining, principal components analysis, projection pursuit).
- The human pattern recognition capabilities can be stimulated by enhancing the presentation of the data by signaling salient features that can be identified automatically.

One should be aware, however, that such algorithms are necessarily limited in their versatility. In general, they can do little beyond screening the data for predefined patterns, missing the versatility of human pattern-recognition faculties. For example, the various techniques generally understood as change-point analysis (Lombard (1998)) or the various approaches to pattern recognition (Kuncheva and Whitaker (2005)) all come down to the application of classifiers. That is, rules that classify (features of) the data as being "normal" or "salient". A simple example of such a classifier is to mark residuals in model fitting as "unusual" if they are larger than two times the error standard deviation. There is no real discovery going on here, only the application of a predefined rule. There are many generic classifiers, but for specific applications, these classifiers could be tailor made (see such disciplines as machine learning and discriminant analysis).

Besides automatically flagging salient features in the data, automatic procedures could also be used to provide displays of the data designed to make certain predefined forms of salient features stand out more clearly. Think of a dendrogram in hierarchical clustering (designed to reveal features in the data related to clustering) or a CUSUM (cumulative sum) plot (designed to reveal drifts and shifts in the mean in time-series data).

Final Remarks

Based on the EDA principles developed in this paper (see Table 1) and the discussion above, Table 2 gives an overview of techniques useful for EDA. Note that it focuses on numerical data. EDA based on categorical data follows the same general process and principles, but techniques and specific principles are different. A future paper will address this topic.

In this final section, we use the developed ideas to clarify a number of issues related to EDA. The first point is repeated because of its importance, namely, that EDA is only suitable for the discovery of factors that actually vary during data collection. For this reason, it should be complemented by other approaches. The study of a large number of EDA applications led us to the impression that statisticians tend to overrate the effectiveness of EDA for discovery. Although there are abundant showcases in the literature where EDA made all the difference in a project, the quality-improvement projects that we have been involved in suggest that EDA is an important, but not the most important, approach to hypothesis generation. Discovery in general does not follow the Baconian spirit that discoveries come from looking at data, but is a more involved interaction between observations, background knowledge/process knowledge, and detective work that is more penetrating than EDA (Mulaik (1985)). Often, EDA is only the first step in the discovery of X-variables, focusing attention to a certain feature of the data's distribution, but without providing sufficient clues for the inquirer to be able to guess what the cause may

TABLE 1. EDA Principles

A The purpose of EDA is the identification of dependent (Y-) and independent (X-) variables that may prove to be of interest for understanding or solving the problem under study

B Display the data such that their distribution is revealed

B1 (Stratified data) Display the data such that both their within-stratum and across-stratum distributions are revealed

B2 (Data plus time order) Display the data such that their distributions within time intervals and across time are revealed

B3 (Multivariate data) Project the data onto a 2- or 3-dimensional subspace and display the data's distribution over this subspace

C Assuming a neutral reference distribution, look for deviations from this reference distribution

C1 Look for deviations from normality

C2 Look for deviations from the maximum entropy distribution

C3 (Data plus time order) Look for deviations from i.i.d. (independent and identical distributions)

C4 (Stratified data) Look for between-strata differences in distribution

C5 (Multivariate data) Look for correlations among variables

C6 Look for discrepancies between a priori perception and the data's distribution

D Identified salient features should be paired with context knowledge in order to interpret them

TABLE 2. An Overview of EDA Techniques

Y-data only

Display data: histogram, plots of the empirical cumulative distribution function (ogive, probability plot)
Salient features: deviations from normality or another suitable neutral reference distribution
Automatic procedures: outlier detection procedures, cluster analysis

Y-data plus time order

Display data: time series plot, boxplot per time unit, individual value plot per time unit
Salient features: deviations from i.i.d. across time; deviations from normality or another neutral reference distribution within time units
Automatic procedures: control charts (including control limits and runs tests), CUSUM and Cuscore charts, change-point analysis techniques

Stratified Y-data

Display data: boxplots per stratum, individual value plot per stratum
Salient features: large between-stratum differences in distribution; deviations from normality or another neutral reference distribution within strata

Multivariate data

Display data: scatter plot of projected data, contour plot of the empirical density projected onto a plane
Salient features: correlations among variables, deviations from normality or another neutral reference distribution
Automatic procedures: principal components analysis (or related procedures), projection pursuit

be. Typically, EDA stimulates and gives direction to the subsequent use of other discovery tools, such as brainstorming, autopsy, pairwise comparison, knowledge pooling, and the other approaches discussed in De Mast and Bergman (2006).

Role of Probabilistic Argumentation

Procedures for the automatic identification of salient features tend to involve probabilistic arguments. Control limits in a control chart, for instance, are related to a tail probability of a reference distribution; and in runs tests (such as *six consecutive observations all increasing or decreasing*), the parameters (here: *six*) are based on a probabilistic argument (see Nelson (1985)). It is, however, good to realize that probability plays no role in identifying the pattern, only in quantifying how salient it is. How long a pattern should be maintained to be signaled as "salient" (is four on a row a trend? or five? or six?) can be given a probabilistic basis. But the form of the pattern (a sequence of increasing or decreasing points) has no probabilistic basis. Although such measures of salience bear a superficial resemblance to significance levels of hypothesis tests, they are different in purpose and principle.

Inferential Studies Camouflaged as EDA

Sometimes, hypotheses identified during an EDA are taken as true facts, and thus hypothesis generation becomes inference. Looking at how inquirers work, it is obvious that real data analyses are somewhere on a continuum, with EDA and CDA as extremes. Is there a problem with that? Contrary to a popular belief, there is no problem *on principle* in using the same data to generate a hypothesis and test it; there is only the complication that p-values, computed in the standard way, do not represent the significance level of such a test, and a correct computation of the significance level of such tests is computationally arduous (see Mayo (1996) for a detailed discussion of this issue). If one is willing to make inferences without probabilistic specification of their significance, one can use data to inspire a hypothesis and then use the same data as evidence (and, in fact, an inferential pattern called "inference to the best explanation" proceeds in this way—see Thagard (2004)).

A serious problem with this procedure, however, is that data that are used for EDA typically do not lend themselves to being a basis for inferences in the strong sense. The typically informal way in which data for EDA are collected gives no guarantee that they are representative for the population under study or that the influence of nuisances is controlled. This is no problem (or is even undesired) for EDA, but it makes EDA data a dangerous basis for inferences.

References

Alwan, L. C.; Ebrahimi, N.; and Soofi, E. S. (1998). "Information Theoretic Framework for Pprocess Control". *European Journal of Operational Research* 111, pp. 526–542.

Bard, Y. (1988). "Maximum Entropy Principle, Classical Approach". In: *Encyclopedia of Statistical Sciences*, S. Kotz and N. Johnson (eds.), Vol. 5, pp. 336–338.

Berger, T. (1988). "Information Theory and Coding Theory". In: *Encyclopedia of Statistical Sciences*, S. Kotz. and N. Johnson (eds.), Vol. 4, pp. 124–141.

Bisgaard, S. (1988). "The Quality Detective: A Case Study". CQPI Report no. 32, Center for Quality and Productivity Improvement, University of Wisconsin.

Bisgaard, S. (1996). "The Importance of Graphics in Problem Solving and Detective Work". *Quality Engineering* 9(1), pp. 157–162.

Brand, M. (1999). "Pattern Discovery Via Entropy Minimization". In: *Artificial Intelligence and Statistics*, D. Heckerman and C. Whittaker (eds.). Kaufman, Oxford.

Chernoff, H. (1973). "The Use of Faces to Represent Points in K-Dimensional Space Graphically". *Journal of the American Statistical Association* 68(342), pp. 361–368.

Cleveland, W. S. and McGill, R. (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". *Journal of the American Statistical Association* 79(387), pp. 531–554.

De Mast, J. (2003). "Quality Improvement from the Viewpoint of Statistical Method". *Quality and Reliability Engineering International* 19(4), pp. 255–264.

De Mast, J. and Bergman, M. (2006). "Hypothesis Generation in Quality Improvement Projects: Approaches for Exploratory Studies". *Quality and Reliability Engineering International* 22(7), pp. 839–850.

Ehrenberg, A. S. C. (1981). "The Problem of Numeracy". *American Statistician* 35(2), pp. 67–71.

Friedman, J. H. and Tukey, J. W. (1974). "A Projection Pursuit Algorithm for Exploratory Data Analysis". *IEEE Transactions on Computers* C-23, pp. 881–890.

Good, I. J. (1963). "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables". *Annals of Mathematical Statistics* 34, pp. 911–934.

Good, I. J. (1983). "The Philosophy of Exploratory Data Analysis". *Philosophy of Science* 50, pp. 283–295.

Hoaglin, D. C.; Mosteller, F.; and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, NY.

Huber, P. J. (1985). "Projection Pursuit". *The Annals of Statistics* 13(2), pp. 435–475.

Iglewicz, B. and Hoaglin, D. C. (1987). "Use of Boxplots for Process Evaluation". *Journal of Quality Technology* 19(4), pp. 180–190.

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics". *Physical Review* 106(4), pp. 620–630.

Jones, M. C. and Sibson, R. (1987). "What Is Projection Pursuit". *Journal of the Royal Statistical Society, Series A* 150, pp. 1–36.

Kuncheva, L. I. and Whitaker, C. J. (2005). "Pattern Recognition". In: *Encyclopedia of Statistics in Behavioral Sciences*, B. S. Everitt and D. C. Howell (eds.), Vol. 3, pp. 1532–1535.

Lombard, F. (1998). "Changepoint Analysis (Update)". In: *Encyclopedia of Statistical Sciences*, S. Kotz, C. B. Read, and D. L. Banks (eds.), Update Vol. 2, pp. 113–120.

Magnani, L. (2000). *Abduction, Reason and Science: Processes of Discovery and Explanation*. Springer, New York, NY.

Maher, P. (1998). "Inductive Inference". In: *Routledge Encyclopedia of Philosophy*, E. Craig (ed.), Vol. 4, pp. 755–759.

Mallows, C. (2006). "Tukey's Paper after 40 Years", with discussion. *Technometrics* 48(3), pp. 319–336

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. The University of Chicago Press, Chicago, IL.

Mulaik, S. A. (1985). "Exploratory Statistics and Empiricism". *Philosophy of Science* 52, pp. 410–430.

Nelson, L. S. (1985). "Interpreting Shewhart X-Bar Control Charts". *Journal of Quality Technology* 17(2), pp. 114–116.

Niiniluoto, I. (1999). "Defending Abduction". *Philosophy of Science* 66(supplemental), pp. S436–S451.

Peirce, C. S. (1931–1935). *Collected Papers 1–5*, C. Hartshorne, P. Weiss (eds.). Harvard University Press, Cambridge, MA.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand Reinhold, Princeton.

Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. The Graduate School of the Department of Agriculture, Washington [reprinted by Dover Publications, New York, 1986].

Snee, R. and Pfeifer, C. G. (1988). "Graphical Representations of Data". In: *Encyclopedia of Statistical Sciences*, S. Kotz. and N. Johnson (eds.), Vol. 3, pp. 488–511.

Thagard, P. (1992). *Conceptual Revolutions*. Princeton University Press, Princeton, NJ.

Thagard, P. (2004). "Rationality and Science". In: *Handbook of Rationality*, A. Mele and P. Rawlings (eds.). Oxford University Press, Oxford.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, PA.