Received: 13 December 2017 | Accepted: 14 January 2018
DOI: 10.1002/qre.2287

RESEARCH ARTICLE

WILEY

The performance of $\bar{X}$ control charts for large non-normally distributed datasets

Leo C.E. Huberts | Marit Schoonhoven | Rob Goedhart | Mandla D. Diko | Ronald J.M.M. Does

Department of Operations Management, Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands

Correspondence
Leo C. E. Huberts, Department of Operations Management, Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands.
Email: l.c.e.huberts@uva.nl

Abstract

Because of digitalization, many organizations possess large datasets. Furthermore, measurement data are often not normally distributed. However, when samples are sufficiently large, the central limit theorem may be used for the sample means. In this article, we evaluate the use of the central limit theorem for various distributions and sample sizes, as well as its effects on the performance of a Shewhart control chart for these large non-normally distributed datasets. To this end, we use the sample means as individual observations and a Shewhart control chart for individual observations to monitor processes. We study the unconditional performance, expressed as the expectation of the in-control average run length (ARL), as well as the conditional performance, expressed as the probability that the control chart based on estimated parameters will have a lower in-control ARL than a specified desired in-control ARL. We use recently developed factors to correct the control limits to obtain a specified conditional or unconditional in-control performance. The results in this paper indicate that the $\bar{X}$ control chart should be applied with caution, even with large sample sizes.

KEYWORDS
big data, central limit theorem, conditional performance, Shewhart, statistical process monitoring

1 | INTRODUCTION

Shewhart control charts are commonly used to monitor process data. Typically, the performance of such control charts is heavily dependent on the assumption of normally distributed data. In practice, this assumption is often violated. For example, Alwan[1] analyzed 235 real datasets and concluded that most of these datasets do not meet the assumptions underlying the traditional control charts.

Since recent advances have led to an increase in the amount of available information, one way to work around the violation of the normality assumptions is to gather larger datasets and use subgroup averages instead of individual observations. Because averages are normally distributed under certain conditions, according to the central limit theorem (CLT), this should largely resolve the issue of non-normally distributed data (cf Billingsley[2]).

While the approach of using averages instead of individual observations is suitable for many statistical techniques, the major difference with many other statistical techniques is that in statistical process monitoring (SPM) we are interested in the long tail behavior of the distribution. This means that, even when the statistic is almost normally distributed, small deviations at the long tails can lead to a bad control chart performance in terms of the false alarm rate and the average run length (ARL). In this paper, we

Correction added on 19 June 2018, after first online publication: affiliation and corresponding author details have been corrected

therefore investigate the performance of Shewhart-type $\bar{X}$ control charts for large non-normally distributed datasets using the convolutions of the distributions. To the best of our knowledge, the performance of Shewhart $\bar{X}$ control charts in this setting has not been investigated thus far.

The paper is structured as follows. In the next section, we briefly describe the model and control charts considered in this paper. Subsequently, in Section 3, the CLT is summarized followed by the convolutions of various probability distributions. In Section 4, we investigate the differences between the normal and non-normal convolutions. Next, Section 5 describes the performance of the Shewhart control chart based on large non-normally distributed datasets. Finally, Section 6 provides some concluding remarks.

2 | THE CLASSICAL SHEWHART CONTROL CHART

Because of the increase in data supply and storage, nowadays organizations often possess large datasets. As the CLT states that under certain conditions the sample means are normally distributed when the samples are sufficiently large, we could treat the sample means as individual observations and use a Shewhart control chart for individual observations under normal theory. To construct such a chart, m samples of size n are collected when the process is assumed to be in control. On the basis of these data, the process mean μ is estimated by

$$\bar{\bar{X}} = \frac{1}{m}\sum_{i=1}^{m}\bar{X}_i = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}X_{ij}\right), \tag{1}$$

where X_{ij} is the j-th observation in the i-th subgroup ($i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$), and the process standard deviation σ is estimated from the standard deviation of the sample means $\bar{X}_i$

$$S = \left(\frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}_i - \bar{\bar{X}})^2\right)^{1/2}. \tag{2}$$

An unbiased estimator of the standard deviation of the sample means ($\sigma/\sqrt{n}$) is $S/c_4(m)$, where $c_4(m)$ is defined by

$$c_4(m) = \left(\frac{2}{m-1}\right)^{1/2}\frac{\Gamma(m/2)}{\Gamma((m-1)/2)}.$$

The choice of the estimator of the standard deviation of the sample means is based on Cryer and Ryan.[3] We have also evaluated the alternative and more traditional estimator based on moving ranges (which was also used by Roes et al[4]). However, the use of this estimator has not improved the performance of the Shewhart $\bar{X}$ control chart, which confirms the result of Cryer and Ryan.[3] The control limits based on estimated parameters are given by

$$\widehat{UCL} = \bar{\bar{X}} + kS/c_4(m), \widehat{LCL} = \bar{\bar{X}} - kS/c_4(m), \tag{3}$$

with $\widehat{UCL}$ and $\widehat{LCL}$ the respective upper and lower control chart limits and k the factor used to achieve the desired in-control performance. When the process parameters are known, k is commonly set equal to 3, which yields a false alarm rate of 0.0027 or equivalently an ARL of 370.4. However, when process parameters are unknown, other values can be chosen to match a certain desired performance. Obtaining a desired control chart performance for practitioners in expectation represents the unconditional performance of the control chart. Recently, factors k_u have been derived to ensure that the in-control ARL in expectation (EARL) is equal to a specified value ($EARL_0$) (see Goedhart et al[5]).

Another recent development is to evaluate control chart design on the variation of the in-control ARLs of the individually estimated, also called conditional control charts. Saleh et al[6] investigated the conditional performance of the traditional control charts based on estimated parameters. They show that for estimated control chart limits for $k = 3$ the probability of ending up with an estimated chart that has an in-control conditional ARL (CARL) lower than 370.4 is considerable. Goedhart et al[7] developed new correction factors k_c for control charts in order to ensure that the probability (P_E) that a design delivers an estimated control chart with an in-control CARL lower than a specified value ($CARL_0$) is at most a specified probability (p).

In this article, we study both the unconditional and conditional performance of the control chart constructed with (3) including the newly developed factors, for the cases where the data are non-normally distributed and various sample sizes ($n = 5, 30, 50, 100, 250, 1000$). With this model, we can investigate whether the CLT works well and whether the newly developed correction factors are applicable to large non-normal datasets as well. We consider the normal distribution, the standard uniform distribution, heavy tailed symmetrical distributions (Student's t_4 and t_{10} and the logistic distribution), and skewed distributions (the lognormal, Gamma(5, 1), Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$ and χ^2_{20} distributions).

The distribution of the sample means for any one of these non-normal distributions can be found using the convolution of that non-normal distribution, ie,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i = \frac{1}{n}C_n,$$

where C_n is the convolution of n i.i.d. random variables with distribution F. In the next section we produce the distribution of C_n for the considered non-normal distributions.

Correction added on 19 June 2018, after first online publication: the running header has been corrected

3 | THE DISTRIBUTION OF THE SAMPLE MEAN

Let $X_1, X_2, \ldots, X_n$ be n i.i.d. observations drawn from F, with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then as n tends to infinity, the random variables $\sqrt{n}(\bar{X} - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$ (cf Billingsley[2]), ie,

$$\sqrt{n}\left(\left(\frac{1}{n}C_n\right) - \mu\right) \xrightarrow{d} N(0, \sigma^2).$$

Hence the asymptotic distribution of the sample means is normal under the above restrictions. The exact distribution for finite values of n can be obtained by evaluating the convolution. To assess the performance of the Shewhart control chart for sample means of non-normally distributed samples, we need the distributional properties of the convolution of these samples: $C_n = \sum_{i=1}^{n} X_i$. The convolutions will allow an investigation of the distribution of the sample means of non-normal distributions and a comparison with the asymptotic normal distribution according to the CLT.

The convolutions are given below; further details on the derivations and approximations are given in the appendix.

3.1 | The convolutions

3.1.1 | The normal distribution

The convolution of i.i.d. normal random variables is just a normal distribution, with mean $n\mu$ and variance $n\sigma^2$

$$C_n \sim N(n\mu, n\sigma^2).$$

3.1.2 | The uniform distribution

The convolution of i.i.d. standard uniform random variables has an Irwin-Hall (IH) distribution, which has a piecewise polynomial probability density function with parameter n (see Hall[8]):

$$C_n \sim IH(n).$$

3.1.3 | The Student's t_ν distribution with ν degrees of freedom

For $\nu = 1$, t_1 is equal to a standard Cauchy distribution and its convolution C_n will have a Cauchy distribution as well (see Blyth[9]):

$$C_n \sim Cauchy(0, n),$$

where 0 and n denote the location and scale parameters of the Cauchy distribution respectively. Note that the conditions needed to apply the CLT do not hold for this case, as the Cauchy distribution has no finite mean and variance. For $\nu > 1$, we use an approximation based on the numerical inversion of the characteristic function.

3.1.4 | The logistic distribution

The standardized version of the sum of i.i.d. logistically distributed random variables with $\mu = 0$ and $s = 1$ can be approximated by a Student's t_ν distributed random variable with $\nu = 5n + 4$ degrees of freedom (George and Mudholkar[10]):

$$C_n \dot{\sim} t_{5n+4}.$$

3.1.5 | The lognormal distribution

The distribution of the convolution C_n of the lognormal distribution can be approximated using 2 methods: the Fenton-Wilkinson approximation by Fenton[11] or the Pearson IV approximation by Nie and Chen.[12] The performance of the Pearson IV approximation turns out to be more accurate than the Fenton-Wilkinson approximation as it matches 2 more moments (see Section 3.2). In the sequel, we will use the Pearson IV approximation

$$C_n \dot{\sim} Pearson_{IV}(\lambda, \alpha, m, \nu),$$

with location parameter λ, scale parameter $\alpha > 0$, and shape parameters $m > \frac{1}{2}$, $\nu \neq 0$.

3.1.6 | The gamma $\Gamma(\alpha, \beta)$ distribution with parameters α and β

If X_i is gamma distributed $X_i \sim \Gamma(\alpha, \beta)$, with parameters α and β, then its convolution is gamma distributed with parameters $n\alpha$ and β

$$C_n \sim \Gamma(n\alpha, \beta).$$

3.1.7 | The chi-squared χ^2_ν distribution with ν degrees of freedom

The convolution distribution of the sum of n i.i.d. chi-squared random variables with ν degrees of freedom is again a chi-squared distribution with $n\nu$ degrees of freedom:

$$C_n \sim \chi^2_{n\nu}.$$

3.2 | Accuracy of the approximated distributions

As reported in the previous section, the convolutions of the Student's t_ν with $\nu > 1$, logistic and lognormal distributions have to be approximated. In the graphs in the left column of Figure 1, the approximated densities of the convolutions for the t_{10}, t_4, logistic and lognormal distributions are plotted and compared with the empirical distribution based on 6 million samples. The graphs in the middle and right columns of Figure 1 zoom in on the 0.135th and 99.865th percentiles of the distributions. The graphs

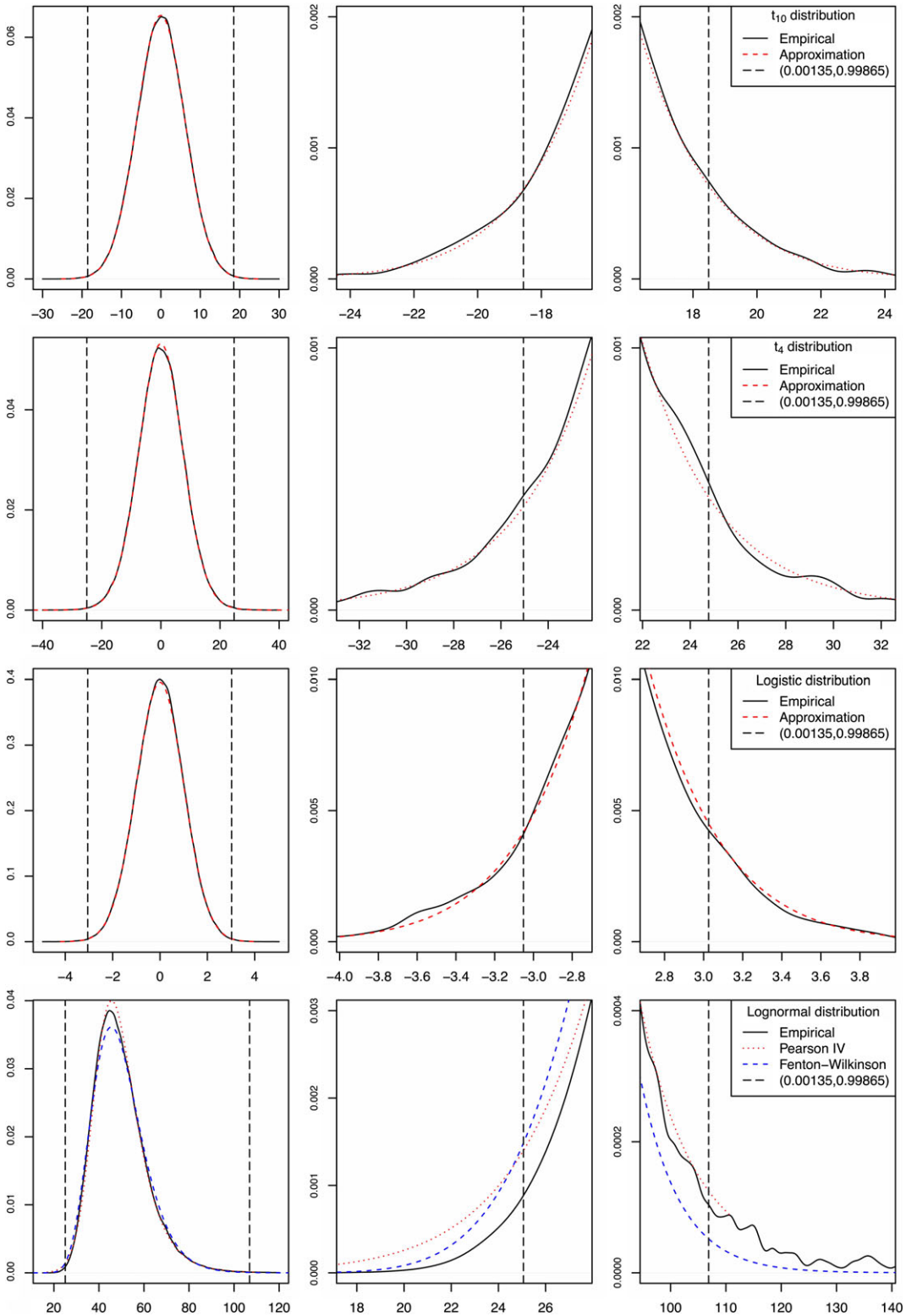

FIGURE 1 Approximated versus empirical densities for $n = 30$ [Colour figure can be viewed at wileyonlinelibrary.com]

show that the approximated t_{10}, t_4, and logistic convolutions are accurate. For the lognormal approximations, we find that the Pearson IV approximation is closer to the empirical distribution than the Fenton-Wilkinson approximation. Thus, we will use the Pearson IV approximation in the sequel.

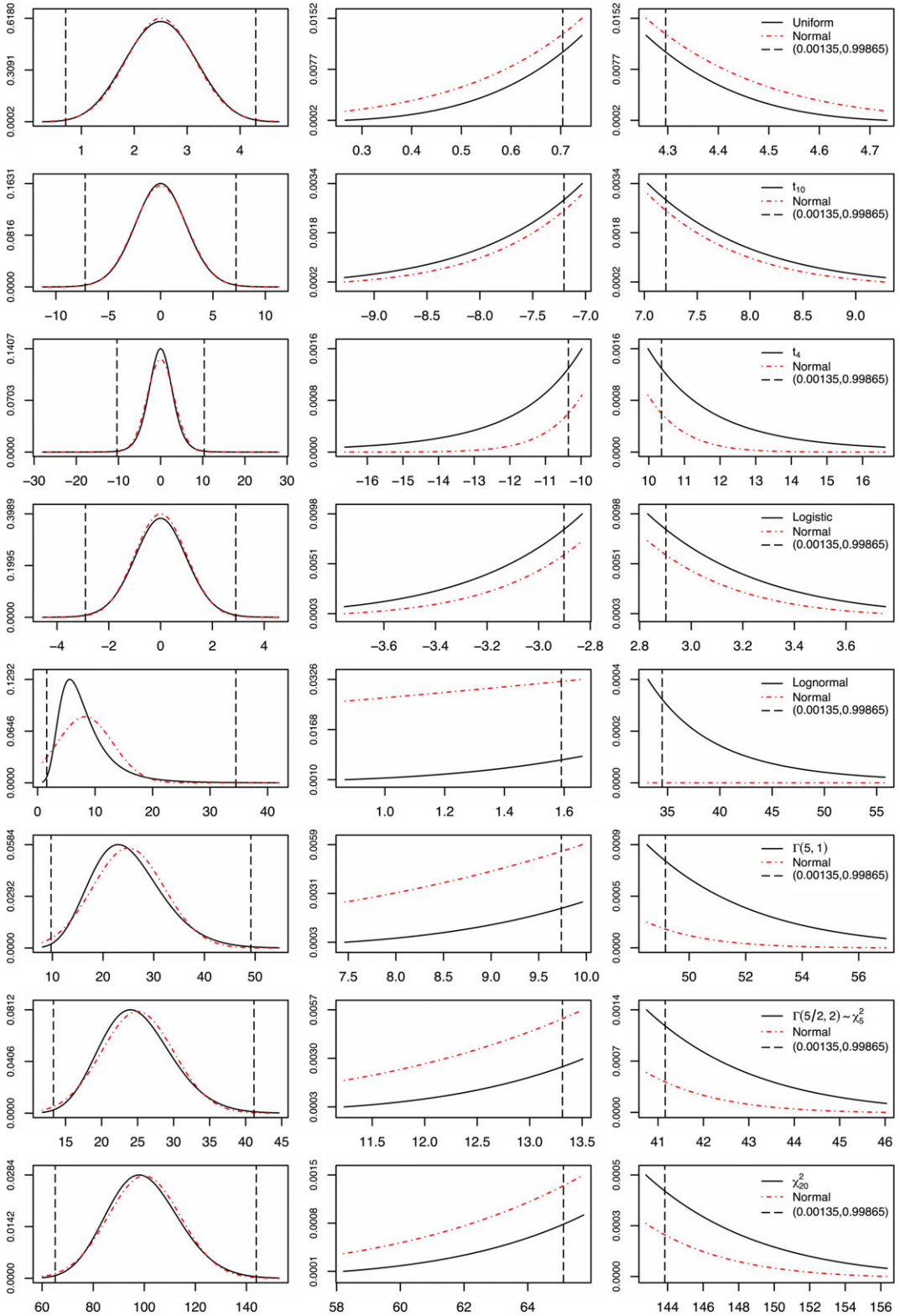

FIGURE 2 Densities of non-normal convolutions versus normal distributions for $n = 5$ and $\alpha = 0.0027$ [Colour figure can be viewed at wileyonlinelibrary.com]

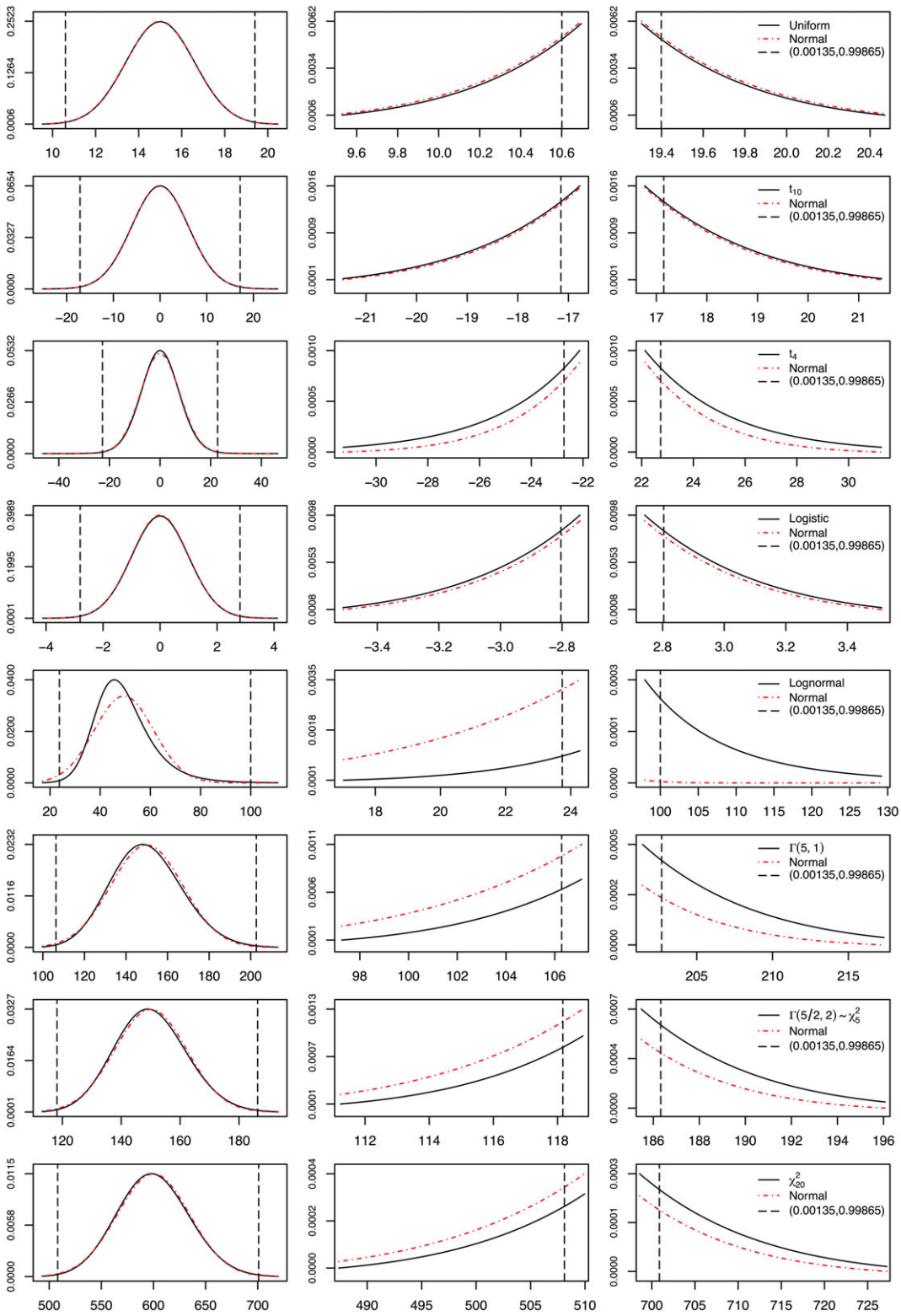

FIGURE 3 Densities of non-normal convolutions versus normal distributions for $n = 30$ and $\alpha = 0.0027$ [Colour figure can be viewed at wileyonlinelibrary.com]

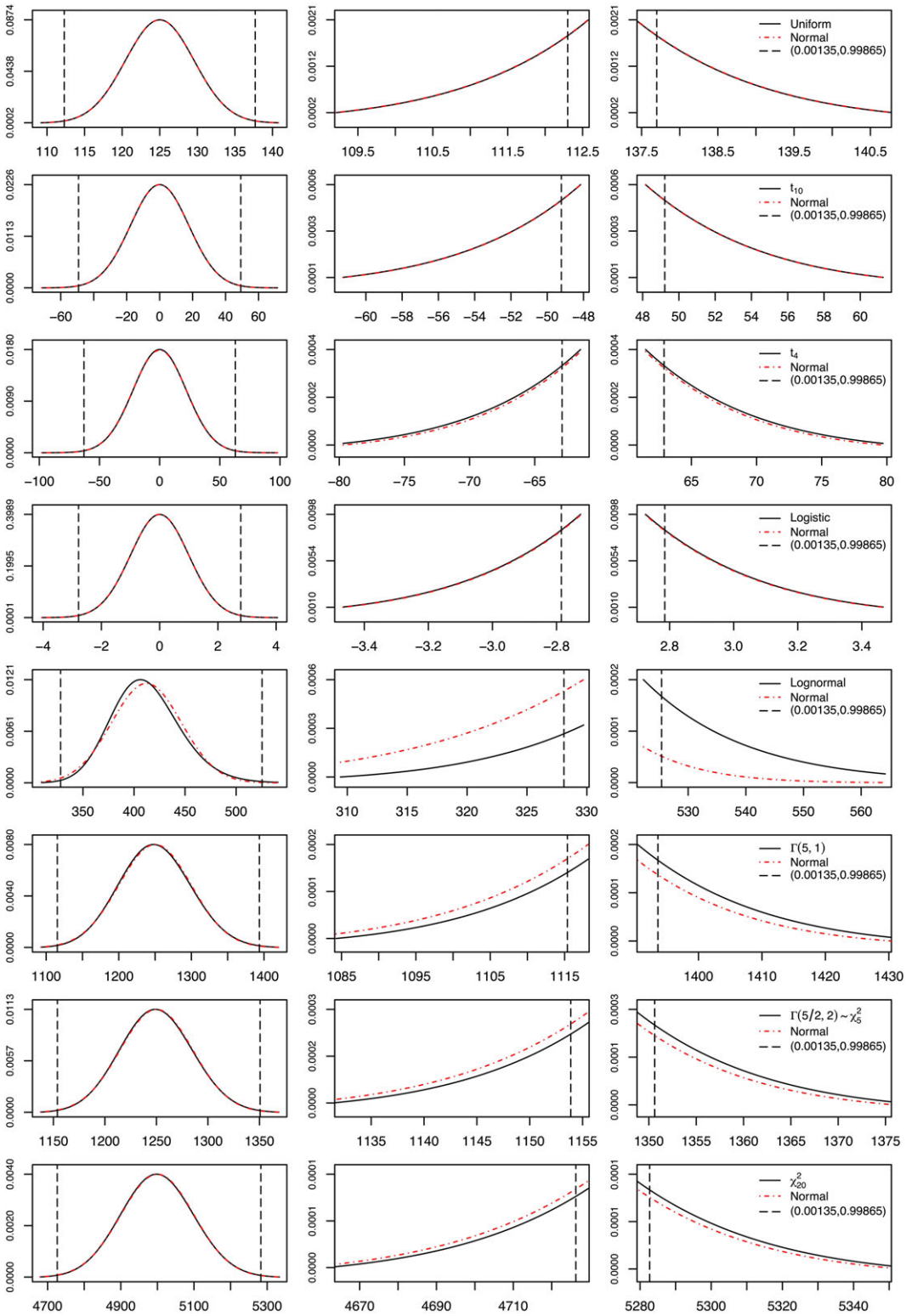

FIGURE 4 Densities of non-normal convolutions versus normal distributions for $n = 250$ and $\alpha = 0.0027$ [Colour figure can be viewed at wileyonlinelibrary.com]

4 | EVALUATION OF THE CENTRAL LIMIT THEOREM

To investigate the differences between the actual distribution of the sample mean and the appropriate normal distribution, we have plotted both distributions and the tail behaviors. In Figures 2 to 4, we have used $n = 5, 30, 250$ and $\alpha = 0.0027$ to investigate the tail behaviors. The graphs on the left give the densities, while the graphs in the middle and on the right zoom in on the 0.135th and 99.865th percentiles of the distributions.

The graphs show that, for a sample size of $n = 30$ or larger, the convolutions of the uniform, t_{10} and logistic distributions, do not deviate much from the normal distribution. The distribution of the t_4 convolution, however, clearly has wider tails than the normal distribution.

The overall distribution of the Gamma convolution is quite close to normal, with gamma$(\frac{5}{2}, 2) \sim \chi^2_5$ closer to normal than gamma(5,1). When we zoom in on the tail behavior, the gamma distributions show skewed tail behavior with narrower tails on the left and wider tails on the right than the normal distribution.

The χ^2_{20} convolution deviates a little from the normal distribution, but less so than the χ^2_5 convolution.

The lognormal convolution shows the largest difference with the normal distribution. The distribution of the lognormal convolution is still strongly skewed for large values of n ($n = 250$).

Note that when we consider a relatively small sample size ($n = 5$), there are large differences for all distributions. This indicates that the normal approximation is not good enough for small sample sizes.

5 | CONTROL CHART PERFORMANCE

5.1 | Simulation procedure

To evaluate the control chart performance, we conduct 10 000 simulation runs for each parameter combination. For each simulation run

1. A dataset consisting of m samples of size n is generated. On the basis of these data, μ is estimated by $\overline{\overline{X}}$ and $\sigma/\sqrt{n}$ is estimated by $S/c_4(m)$, using (1) and (2). Next, $\widehat{UCL}$ and $\widehat{LCL}$ can be determined using (3). Factor k_u is based on Goedhart et al[5] and factor k_c on Goedhart et al.[7]
2. For each dataset, the conditional false alarm rate (CFAR) is calculated as $CFAR = 1 - P(\widehat{LCL} < \bar{X} < \widehat{UCL}) = 1 - P(n\widehat{LCL} < C_n < n\widehat{UCL})$ using the convolutions of Section 3.1. The CARL is given by $1/CFAR$.

When we perform the above procedure, we end up with 10 000 CARLs of individually estimated control charts. When k_u is used, the EARL is estimated by averaging the 10 000 CARLs of the simulated control charts. When k_c is used, the exceedance probability (P_E) is obtained by determining the percentage of CARLs lower than a specified value ($CARL_0$). Both the unconditional and conditional results were verified using the empirical distribution of the non-normal distributions.

We expect that the higher $EARL_0$ or $CARL_0$, the larger the sample size should be to ensure that the performance of the control charts is as desired. This is because the higher these values are, the more our interest moves towards the long tail of the distribution of the sample means, where minor deviations from the normal approximation have more impact on the performance. For this reason, we consider various values for $EARL_0$ and $CARL_0$, namely, 1000, 370.4, and 100.

Finally, as we expect that the correction factors are more accurate when the sample size (n) is larger, we consider a broad range of values, namely, $n = 5, 30, 50, 100, 250, 1000$. For the amount of samples m, we take values $m = 30, 50, 100, 200$.

5.2 | Unconditional performance

In this section, we present the simulation results of the control charts based on (3) and k_u as defined in Goedhart et al.[5] Tables 1 to 3 present the results for an $EARL_0$ equal to 1000, 370.4, and 100, respectively. Each table presents the EARL and 5th, 50th and 95th percentiles of the CARL distribution.

Each table shows that the larger the sample size (n), the closer the EARL is to its desired value $EARL_0$ and so the more applicable is the correction factor. Increasing the number of samples (m) also reduces the deviation in performance with respect to the case of normally distributed data, but the impact of m is less strong than the impact of n, as was to be expected. Also, the value of $EARL_0$ is of influence: the higher $EARL_0$, the larger the sample size should be to obtain a performance that resembles the performance under normality. This can be explained as the relative difference between the distributions of the means based on the non-normal and normal distributions is the largest in the tails of the distributions. To give an example, for the case $EARL_0 = 1000$, the t_{10} and logistic distributions require a sample size of 100 or larger in order to obtain a reasonable in-control performance with the use of the given correction factors while, for the case $EARL_0 = 100$, a sample size of 30 is sufficient to obtain the desired EARL values.

As discussed in Section 4, the uniform distribution is the only distribution that has a convolution distribution with

thinner tails than the normal distribution on both sides. This produces extremely large EARL values for small n. Furthermore, as the uniform distribution is bounded by an interval, conditional control limits have been generated that produce a CFAR of zero for small values of n giving an infinite CARL. Tables 1 to 3 show the amount of infinite values we found for the uniform distribution within the second parentheses.

TABLE 1 EARL (5th, 50th, 95th percentile of CARL) with $EARL_0 = 1000$

n	Distribution	$m = 30$	$m = 50$	$m = 100$	$m = 200$
5	Normal	871 (47,300,3216)	996 (98,478,3207)	1005 (206,690,2816)	1009 (344,832,2256)
	Uniform[a]	1.9^*10^{12} (60,656, 83126)(4837)	4.5^*10^9 (155,1554, 113257)(3213)	9.2^*10^9 (434,3284, 102822)(1036)	81054 (995,4941, 58812)(95)
	t_{10}	448 (42,213,1516)	500 (79,305,1535)	536 (151,422,1295)	550 (232,486,1082)
	t_4	157 (31,101,395)	169 (48,127,388)	185 (77,155,360)	187 (102,170,312)
	Logistic	409 (41,201,1377)	452 (79,295,1292)	495 (147,392,1195)	506 (216,451,985)
	Lognormal	98 (16,48,233)	90 (23,56,204)	82 (32,64,172)	82 (40,69,146)
	Gamma(5, 1)	431 (40,200,1490)	441 (73,275,1321)	429 (127,341,1014)	427 (187,378,835)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	289 (35,149,941)	298 (59,198,850)	297 (97,241,684)	290 (134,262,544)
	χ^2_{20}	589 (43,235,2024)	584 (82,347,1814)	592 (157,451,1503)	588 (235,514,1190)
30	Normal	978 (47,307,3430)	996 (101,479,3263)	1001 (203,682,2886)	998 (341,823,2215)
	Uniform	1136 (48,334,4000)(0)	1252 (104,549,4260)(0)	1205 (230,792,3490)(0)	1206 (386,965,2829)(0)
	t_{10}	748 (49,280,2648)	857 (94,439,2812)	866 (195,620,2332)	876 (316,748,1881)
	t_4	312 (41,180,911)	363 (72,255,946)	396 (132,326,845)	401 (191,367,718)
	Logistic	715 (45,269,2544)	861 (94,433,2798)	850 (194,606,2314)	858 (311,723,1842)
	Lognormal	121 (25,77,315)	127 (38,93,293)	124 (54,106,244)	128 (70,115,212)
	Gamma(5, 1)	722 (47,270,2581)	805 (96,424,2508)	793 (183,568,2113)	800 (289,679,1700)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	637 (46,252,2241)	682 (92,386,2142)	681 (169,510,1784)	678 (264,582,1405)
	χ^2_{20}	847 (48,281,2773)	884 (96,441,2973)	891 (197,633,2368)	884 (313,743,1911)
50	Normal	911 (48,296,3256)	1040 (100,475,3233)	999 (211,694,2745)	995 (341,823,2222)
	Uniform	996 (48,307,3385)(0)	1159 (103,510,3742)(0)	1087 (216,738,3078)(0)	1088 (359,895,2446)(0)
	t_{10}	832 (48,293,2886)	881 (96,452,2939)	924 (195,651,2597)	912 (318,763,2013)
	t_4	378 (43,208,1230)	427 (83,294,1201)	471 (147,384,1048)	491 (224,445,916)
	logistic	788 (46,287,2783)	886 (97,447,2990)	898 (190,643,2388)	914 (322,765,2007)
	Lognormal	148 (28,90,403)	151 (43,112,358)	159 (63,130,303)	155 (83,140,259)
	Gamma(5, 1)	818 (47,288,2987)	858 (95,438,2886)	884 (199,626,2374)	866 (312,725,1895)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	797 (46,265,2556)	766 (91,412,2403)	782 (182,554,2103)	785 (291,669,1664)
	χ^2_{20}	850 (48,291,2930)	926 (99,457,3136)	925 (193,660,2563)	941 (323,783,2098)
100	Normal	974 (48,304,3349)	932 (97,475,3080)	992 (205,693,2746)	1009 (342,846,2234)
	Uniform	992 (49,305,3339)(0)	1003 (100,490,3393)(0)	1052 (215,715,2901)(0)	1056 (352,871,2360)(0)
	t_{10}	920 (47,295,3145)	950 (101,482,3130)	960 (201,666,2644)	969 (333,812,2143)
	t_4	473 (44,235,1599)	554 (90,353,1670)	592 (165,464,1396)	622 (258,553,1241)
	logistic	970 (46,287,2953)	922 (99,459,3090)	947 (207,660,2573)	956 (330,795,2076)
	Lognormal	394 (32,115,562)	205 (51,149,515)	209 (82,179,425)	211 (109,192,361)
	Gamma(5, 1)	841 (49,289,3049)	898 (96,460,2939)	928 (200,657,2516)	939 (322,778,2077)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	822 (47,287,2858)	874 (96,453,2937)	865 (192,625,2280)	864 (314,728,1882)
	χ^2_{20}	864 (49,293,3217)	949 (100,471,3186)	959 (205,666,2614)	957 (332,801,2102)
250	Normal	947 (48,300,3168)	1003 (97,477,3285)	996 (203,695,2782)	1004 (338,832,2233)
	Uniform	945 (47,299,3524)(0)	987 (99,477,3258)(0)	1000 (210,701,2827)(0)	1018 (342,839,2285)(0)
	t_{10}	876 (48,294,3282)	939 (101,485,3180)	996 (203,685,2795)	977 (334,819,2145)
	t_4	605 (46,270,2204)	699 (91,407,2192)	759 (182,571,1960)	770 (293,663,1583)
	logistic	906 (47,295,3170)	970 (100,478,3229)	975 (204,681,2660)	983 (335,808,2170)
	Lognormal	372 (37,165,1011)	334 (68,225,929)	363 (112,281,771)	342 (162,312,621)
	Gamma(5, 1)	900 (48,297,3030)	954 (97,472,3135)	984 (205,676,2680)	972 (334,814,2123)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	925 (48,293,3086)	933 (99,467,2989)	940 (199,663,2567)	955 (335,799,2089)
	χ^2_{20}	877 (48,299,3002)	976 (102,472,3167)	1002 (205,681,2783)	985 (338,818,2171)

(Continues)

TABLE 1 (Continued)

n	Distribution	$m = 30$	$m = 50$	$m = 100$	$m = 200$
1000	Normal	1122 (47,299,3286)	996 (99,476,3170)	1002 (206,690,2789)	1002 (340,830,2227)
	Uniform	914 (49,294,3172)(0)	985 (102,482,3304)(0)	1006 (210,701,2787)(0)	977 (338,829,2226)(0)
	t_{10}	852 (48,301,3291)	979 (102,489,3187)	978 (204,683,2691)	1001 (345,837,2210)
	t_4	812 (47,289,2927)	852 (96,445,2892)	907 (200,644,2425)	913 (332,772,1983)
	logistic	925 (47,300,3082)	976 (99,476,3302)	1000 (213,689,2784)	991 (334,830,2195)
	Lognormal	596 (43,240,2014)	632 (88,365,1969)	646 (163,482,1640)	647 (257,564,1304)
	Gamma(5, 1)	929 (49,301,3246)	1028 (101,487,3305)	983 (207,684,2746)	998 (340,831,2227)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	889 (47,295,3191)	991 (98,470,3345)	990 (203,683,2783)	981 (338,822,2180)
	χ^2_{20}	974 (48,300,3222)	1057 (98,477,3392)	1007 (202,695,2796)	991 (336,821,2169)

[a]The amount of infinite CARL values we found is indicated within the second parentheses.

In Section 4, we already indicated large differences between the normal distribution and the distributions of the lognormal and t_4 convolutions and small deviations compared with the uniform, t_{10}, logistic, Gamma(5, 1), Gamma($\frac{5}{2}$, 2) $\sim$ χ^2_5, and χ^2_{20} convolutions. The EARL results confirm these hypotheses, as for all values of n and m the lognormal EARL values are consistently far below the desired $EARL_0$, indicating the strong skewness as observed in the analysis of the convolutions.

5.3 | Conditional performance

In this section, we present the results of the control charts based on (3) with k_c such that the probability of having an in-control CARL lower than a specified value ($CARL_0$) is equal to p (cf Goedhart et al[7]). We set $p = 10\%$. Tables 4 to 6 present the realized exceedance probabilities P_E for a specified $CARL_0$ of 1000, 370.4, and 100, respectively. Each table presents the results for various sample sizes ($n = 5, 30, 50, 100, 250, 1000$), various numbers of samples ($m = 30, 50, 100, 200$), and various distributions (normal, uniform, t_{10}, t_4, logistic with $\mu = 0$ and $s = 1$, lognormal with $\mu = 0$ and $\sigma = 1$, Gamma(5, 1), Gamma($\frac{5}{2}$, 2) $\sim$ χ^2_5 and χ^2_{20}).

As for the unconditional case, the tables show that the larger the sample size (n), the closer P_E is to its desired value $p(10\%)$, and so the better the applicability of the control charts. Also, the value for $CARL_0$ has an impact: the lower the $CARL_0$, the closer the control chart performance is to the desired performance. This can be explained by the increase in relative difference further in the tails of the distributions.

The normal approximation is worst in the case of the lognormal distribution, as we see that the deviation of P_E with respect to $p = 10\%$ is the largest. A very large sample size (n) is needed to guarantee a desired conditional performance. In the case of $CARL_0 = 100$, a sample size of 1000 gives reasonable P_E values, also for the lognormal distribution, while for $CARL_0 = 1000$ and 370.4 even a sample size of 1000 is not large enough to ensure the right exceedance probabilities.

Interestingly, increasing m actually increases P_E for the non-normal distributions in most situations. For example, the t_4 distribution for $CARL_0 = 370.4$ and $n = 50$ has a P_E of 17.2% for $m = 30$. With m increased to 200, for t_4 now 40.3% of the CARLs are below the desired $CARL_0 = 370.4$. This can be explained by a decrease in parameter estimate variation and thus a decrease in the constant k_c, causing tighter control limits.

6 | SUMMARY AND CONCLUDING REMARKS

In this paper, we have studied the applicability of the CLT to large non-normal datasets. According to the CLT, sufficiently large samples should lead to normally distributed sample averages. However, since SPM is concerned with the far tail of the distribution, it was unclear whether the convergence to normality would be sufficient.

In this research, we have thus investigated whether the charting constants that are designed for normally distributed data can also be applied to large non-normal datasets. In particular, we have applied the Shewhart control chart for individual observations to monitor the sample means of non-normally distributed datasets.

The study demonstrates that the appropriateness of the control charting constants, also for non-normally distributed data, depends on various factors. These factors include the sample size (n), the number of samples (m), the specified desired performance of the control chart, and the degree of the deviation from normality. When the deviation from normality is moderate (as is the case for the uniform, t_{10}, logistic, Gamma(5, 1), Gamma($\frac{5}{2}$) $\sim$ χ^2_5, and χ^2_{20} distributions), a sample size of 100 is large enough to ensure appropriate use of the correction factors.

However, when the deviation from normality is substantial due to heavy tails (t_4) or substantial skewness (lognormal), the correction factors are not applicable even when

TABLE 2 EARL (5th, 50th, 95th percentile of CARL) with $EARL_0 = 370.4$

n	Distribution	$m = 30$	$m = 50$	$m = 100$	$m = 200$
5	Normal	378 (32,155,1224)	365 (56,217,1142)	375 (103,283,945)	368 (148,321,737)
	Uniform[a]	271351 (36,251,7283)(2202)	15844 (74,430,6530)(679)	3061 (157,663,5127)(32)	1292 (279,827,3571)(0)
	t_{10}	224 (28,121,727)	245 (49,165,693)	246 (81,202,561)	252 (118,228,469)
	t_4	128 (22,70,276)	113 (34,85,258)	119 (51,102,226)	121 (66,109,197)
	Logistic	211 (28,117,683)	225 (47,155,623)	230 (80,190,506)	238 (113,217,434)
	Lognormal	75 (14,39,177)	67 (19,45,169)	64 (26,51,132)	63 (33,55,114)
	Gamma(5, 1)	236 (29,123,786)	229 (47,156,642)	224 (77,185,498)	224 (107,204,410)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	171 (26,101,523)	174 (42,124,462)	170 (63,143,363)	167 (84,154,294)
	χ^2_{20}	292 (30,138,966)	284 (51,177,854)	278 (86,222,656)	278 (126,249,529)
30	Normal	344 (32,155,1167)	366 (57,219,1123)	366 (101,282,907)	368 (150,322,746)
	Uniform	406 (31,167,1413)(0)	430 (58,239,1355)(0)	417 (109,310,1073)(0)	418 (166,357,874)(0)
	t_{10}	324 (31,149,1085)	340 (54,206,1047)	340 (94,264,836)	340 (145,299,675)
	t_4	193 (27,111,543)	197 (48,144,510)	209 (75,174,440)	212 (106,193,367)
	Logistic	338 (31,145,1052)	323 (55,204,986)	336 (96,263,817)	338 (145,297,669)
	Lognormal	89 (20,59,230)	92 (29,70,210)	89 (40,77,170)	90 (51,83,149)
	Gamma(5, 1)	318 (30,147,1059)	328 (56,202,958)	331 (95,262,793)	331 (141,294,645)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	302 (31,144,1042)	304 (54,193,906)	302 (90,241,715)	301 (132,266,586)
	χ^2_{20}	349 (31,151,1166)	344 (56,208,1035)	352 (99,271,871)	347 (146,304,690)
50	Normal	336 (31,152,1112)	370 (56,220,1127)	366 (101,282,905)	368 (151,322,744)
	Uniform	373 (31,156,1244)(0)	407 (58,226,1237)(0)	389 (105,295,984)(0)	390 (157,339,788)(0)
	t_{10}	366 (30,148,1096)	352 (57,213,1065)	349 (98,273,867)	349 (148,306,694)
	t_4	213 (29,120,639)	225 (49,159,591)	236 (80,195,500)	243 (116,220,434)
	Logistic	327 (31,149,1127)	344 (55,210,1051)	344 (97,269,835)	352 (149,308,695)
	Lognormal	130 (21,67,272)	104 (32,79,240)	105 (46,91,204)	104 (58,97,172)
	Gamma(5, 1)	350 (31,147,1187)	342 (56,206,1043)	345 (100,267,841)	348 (145,309,691)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	334 (31,145,1119)	324 (55,204,980)	329 (96,256,785)	323 (137,286,630)
	χ^2_{20}	358 (32,152,1197)	352 (55,212,1083)	356 (99,274,876)	358 (147,314,719)
100	Normal	371 (31,153,1233)	367 (56,215,1127)	369 (102,282,927)	367 (149,323,739)
	Uniform	373 (32,156,1247)(0)	371 (56,221,1152)(0)	382 (105,290,948)(0)	384 (156,334,774)(0)
	t_{10}	349 (30,154,1245)	356 (56,215,1057)	364 (101,279,897)	358 (149,312,711)
	t_4	253 (29,129,758)	271 (50,182,755)	279 (88,225,623)	282 (128,253,529)
	Logistic	334 (31,153,1157)	346 (56,209,1073)	361 (101,275,907)	359 (147,317,716)
	Lognormal	147 (24,83,377)	134 (36,98,331)	131 (54,114,261)	133 (72,123,222)
	Gamma(5, 1)	372 (31,156,1275)	361 (55,211,1067)	359 (102,277,896)	360 (149,317,713)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	338 (31,149,1153)	356 (55,212,1088)	348 (98,269,856)	347 (146,304,694)
	χ^2_{20}	366 (31,152,1164)	364 (56,216,1134)	367 (102,281,909)	362 (148,315,726)
250	Normal	372 (32,154,1227)	371 (55,220,1152)	370 (103,283,924)	371 (153,325,740)
	Uniform	364 (31,154,1314)(0)	368 (56,216,1119)(0)	370 (103,286,935)(0)	374 (152,326,759)(0)
	t_{10}	361 (31,154,1232)	357 (56,214,1089)	373 (102,286,918)	365 (151,321,731)
	t_4	295 (30,141,956)	308 (54,197,904)	313 (94,253,735)	317 (137,282,607)
	Logistic	362 (33,154,1201)	357 (54,217,1084)	366 (101,279,923)	365 (150,319,730)
	Lognormal	206 (27,106,580)	187 (44,134,491)	191 (70,159,409)	191 (95,175,334)
	Gamma(5, 1)	365 (31,160,1258)	362 (58,213,1122)	367 (100,282,905)	367 (150,320,747)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	359 (31,157,1220)	364 (55,215,1092)	361 (100,280,889)	359 (149,314,721)
	χ^2_{20}	360 (32,156,1220)	360 (56,217,1120)	364 (98,278,905)	368 (153,320,746)
1000	Normal	376 (32,151,1235)	358 (56,215,1100)	372 (102,285,947)	367 (153,322,730)
	Uniform	356 (32,152,1199)(0)	368 (57,219,1199)(0)	373 (103,287,926)(0)	369 (151,323,745)(0)
	t_{10}	367 (32,154,1220)	368 (56,218,1141)	365 (101,278,914)	369 (152,321,744)
	t_4	323 (32,150,1150)	339 (56,210,1013)	357 (100,276,876)	353 (148,310,706)
	Logistic	347 (31,153,1155)	363 (56,215,1125)	371 (100,286,904)	365 (150,319,735)
	Lognormal	282 (29,135,926)	290 (52,184,845)	288 (88,230,671)	287 (127,258,544)
	Gamma(5, 1)	367 (31,154,1189)	369 (58,217,1165)	373 (103,281,923)	367 (151,321,747)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	341 (31,153,1190)	371 (57,222,1149)	369 (101,284,911)	369 (150,322,750)
	χ^2_{20}	357 (31,155,1225)	361 (56,213,1110)	374 (100,280,959)	367 (149,319,736)

[a]The amount of infinite CARL values we found is indicated within the second parentheses.

TABLE 3 EARL (5th, 50th, 95th percentile of CARL) with $EARL_0 = 100$

n	Distribution	$m = 30$	$m = 50$	$m = 100$	$m = 200$
5	Normal	100 (16,59,305)	101 (25,73,261)	99 (39,85,206)	100 (51,92,176)
	Uniform[a]	187 (18,73,585)(151)	165 (29,96,480)(2)	155 (47,119,382)(0)	150 (67,131,299)(0)
	t_{10}	80 (16,53,226)	83 (24,64,212)	83 (35,73,165)	83 (46,78,138)
	t_4	59 (14,39,142)	60 (20,46,129)	60 (27,52,112)	60 (35,55,95)
	Logistic	78 (16,51,219)	78 (23,61,193)	80 (34,70,159)	81 (45,76,133)
	Lognormal	48 (11,29,127)	104 (14,33,111)	46 (19,36,89)	44 (24,38,75)
	Gamma(5, 1)	92 (16,56,267)	89 (24,66,221)	87 (36,75,176)	87 (46,81,148)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	81 (16,53,236)	79 (23,61,196)	76 (33,67,150)	76 (43,72,124)
	χ^2_{20}	96 (17,59,284)	93 (25,70,234)	93 (37,80,192)	94 (50,87,161)
30	Normal	99 (17,58,297)	102 (26,74,261)	101 (38,86,214)	101 (52,93,175)
	Uniform	105 (17,62,326)(0)	108 (26,77,287)(0)	105 (40,89,226)(0)	106 (54,97,189)(0)
	t_{10}	97 (17,59,287)	96 (25,72,244)	97 (38,83,200)	96 (50,89,165)
	t_4	80 (16,51,207)	85 (23,61,180)	79 (34,69,152)	80 (45,74,130)
	Logistic	93 (16,58,280)	94 (25,69,237)	94 (37,82,194)	96 (51,89,164)
	Lognormal	57 (14,38,143)	54 (19,43,122)	55 (25,47,101)	53 (31,49,85)
	Gamma(5, 1)	97 (17,59,295)	97 (25,71,249)	97 (38,83,204)	98 (51,91,170)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	97 (17,59,287)	95 (25,70,246)	96 (38,83,199)	95 (50,88,163)
	χ^2_{20}	100 (17,58,309)	99 (26,72,260)	99 (39,85,207)	99 (52,91,170)
50	Normal	102 (17,59,301)	97 (26,72,244)	99 (39,85,206)	101 (52,93,174)
	Uniform	101 (17,59,300)(0)	105 (26,74,273)(0)	102 (39,87,216)(0)	102 (53,94,177)(0)
	t_{10}	99 (17,59,301)	98 (25,71,257)	97 (38,84,203)	98 (51,90,169)
	t_4	88 (16,53,225)	85 (24,64,197)	84 (35,72,163)	85 (47,78,140)
	Logistic	96 (17,57,292)	96 (25,70,253)	97 (39,83,199)	97 (51,90,165)
	Lognormal	65 (15,41,149)	64 (20,46,133)	58 (27,51,106)	57 (34,54,92)
	Gamma(5, 1)	96 (17,58,292)	100 (26,73,266)	98 (38,84,204)	99 (51,91,171)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	106 (16,59,304)	98 (25,72,255)	98 (39,83,203)	97 (50,90,168)
	χ^2_{20}	99 (17,58,308)	99 (26,72,262)	99 (39,85,207)	99 (51,92,173)
100	Normal	97 (17,58,298)	98 (25,73,252)	99 (38,84,208)	100 (52,92,175)
	Uniform	101 (17,60,305)(0)	100 (25,73,262)(0)	101 (40,86,212)(0)	102 (53,94,177)(0)
	t_{10}	96 (17,58,283)	98 (25,73,254)	99 (38,84,210)	99 (52,92,171)
	t_4	93 (16,55,248)	88 (25,66,216)	91 (37,78,182)	91 (48,84,155)
	Logistic	100 (17,59,304)	99 (25,72,263)	98 (39,84,205)	99 (51,91,173)
	Lognormal	73 (15,44,181)	66 (21,52,150)	76 (30,58,124)	65 (38,61,104)
	Gamma(5, 1)	101 (17,59,293)	101 (26,73,268)	99 (38,84,209)	99 (52,92,171)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	98 (16,58,295)	100 (26,73,260)	99 (39,85,204)	99 (51,92,170)
	χ^2_{20}	99 (17,59,303)	99 (26,72,259)	100 (39,85,208)	100 (52,93,171)
250	Normal	99 (17,59,299)	102 (25,74,263)	100 (38,85,212)	100 (52,92,175)
	Uniform	101 (16,59,319)(0)	100 (25,72,258)(0)	100 (39,86,212)(0)	100 (52,92,176)(0)
	t_{10}	100 (17,59,302)	99 (25,72,258)	100 (39,86,211)	99 (52,92,172)
	t_4	93 (17,58,284)	100 (24,69,232)	96 (38,81,192)	95 (50,87,163)
	Logistic	102 (17,60,312)	101 (25,73,257)	100 (38,85,212)	99 (51,92,173)
	Lognormal	95 (15,50,223)	79 (23,60,190)	78 (34,68,155)	77 (43,72,126)
	Gamma(5, 1)	99 (17,60,307)	99 (26,74,257)	99 (39,85,207)	100 (52,92,177)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	101 (16,60,308)	99 (25,73,259)	100 (39,84,216)	98 (52,91,171)
	χ^2_{20}	101 (17,59,315)	100 (25,73,259)	100 (39,85,209)	100 (52,93,176)
1000	Normal	102 (17,60,312)	101 (25,72,264)	100 (39,85,211)	101 (52,92,176)
	Uniform	99 (17,59,299)(0)	100 (26,73,262)(0)	101 (39,86,211)(0)	100 (52,92,174)(0)
	t_{10}	103 (17,59,310)	101 (26,74,266)	100 (39,85,209)	100 (52,93,173)
	t_4	98 (16,59,298)	99 (25,72,254)	98 (39,84,205)	98 (51,91,170)
	Logistic	98 (17,59,305)	100 (25,73,255)	100 (38,85,209)	100 (52,93,172)
	Lognormal	91 (16,56,261)	93 (25,68,241)	92 (38,79,189)	92 (49,85,158)
	Gamma(5, 1)	104 (17,60,316)	98 (26,72,262)	99 (39,85,208)	101 (52,93,173)
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	100 (17,60,307)	99 (25,73,259)	100 (38,85,205)	99 (51,92,171)
	χ^2_{20}	99 (17,58,296)	99 (25,73,251)	100 (39,86,211)	100 (52,92,173)

[a]The amount of infinite ARL values we found is indicated within the second parentheses

TABLE 4 P_E with $CARL_0 = 1000$ and $p = 10\%$

n	Distribution	$m = 30$	$m = 50$	$m = 100$	$m = 200$
5	Normal	8.9	9.5	9.4	9.3
	Uniform	2.9	1.6	0.7	0
	t_{10}	18.8	22.4	31.7	45.2
	t_4	83.7	93.5	98.7	99.6
	Logistic	20.2	25.2	35.8	52.6
	Lognormal	97.6	99	99.7	99.9
	Gamma(5, 1)	28.3	36.8	53.4	73.6
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	44	57.9	78.2	94.3
	χ^2_{20}	18.6	23.7	31.8	45.5
30	Normal	9	9.1	9.7	9.5
	Uniform	8.9	7.9	6.5	5.7
	t_{10}	10.6	11.3	12.5	13.3
	t_4	31.2	42.5	61.8	82.2
	Logistic	10.9	11.1	12.1	14.9
	Lognormal	92.3	97.2	99.6	100
	Gamma(5, 1)	12.8	14.2	16	19.5
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	15.3	18.2	24.2	32.3
	χ^2_{20}	11	11.3	12.5	14.3
50	Normal	9	9.4	9	9.4
	Uniform	9.3	8.9	8.2	7.8
	t_{10}	9.8	10.3	10.9	11.8
	t_4	21.7	30	42.6	62.3
	Logistic	10.1	10.5	11.5	12.2
	Lognormal	85.8	94.3	99.3	99.9
	Gamma(5, 1)	11.3	11.8	13.4	15.5
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	12.6	14.8	17.9	22.2
	χ^2_{20}	10	10.2	12.1	12.5
100	Normal	9.4	9.2	9.5	9.6
	Uniform	9.2	9.5	9	8.4
	t_{10}	9.1	9.7	10.1	10.9
	t_4	16.1	19.4	26.7	37
	Logistic	9.5	9.6	10.2	10.9
	Lognormal	67.5	82.5	95.9	99.8
	Gamma(5, 1)	9.5	10.3	11.7	12
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	11	11.5	13.6	16.1
	χ^2_{20}	9.7	9.9	10.5	11.4
250	Normal	9.5	9.8	8.8	9.4
	Uniform	9.8	10.1	9.5	9.5
	t_{10}	8.9	9.2	9.4	10.2
	t_4	12.1	13.3	16.2	20
	Logistic	9	9.4	9.5	10
	Lognormal	37.4	49.9	70.8	89.2
	Gamma(5, 1)	9.1	10	10.6	11
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	9.8	10.6	9.8	11.8
	χ^2_{20}	9.6	9.8	9.8	10
1000	Normal	8.8	9.4	9.3	9.6
	Uniform	10.5	8.8	9.1	10.3
	t_{10}	9.2	9.5	9.7	9.6
	t_4	10	10.5	11.5	12.8
	Logistic	9.2	9.2	9.6	10
	Lognormal	15.5	19.2	25.1	34.5
	Gamma(5, 1)	9.2	9.4	9.7	9
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	8.6	9.4	9.3	10.1
	χ^2_{20}	9.6	9.3	8.7	9.8

TABLE 5 P_E with $CARL_0 = 370.4$ and $p = 10\%$

n	Distribution	$m = 30$	$m = 50$	$m = 100$	$m = 200$
5	Normal	8.9	9.6	9.2	9.8
	Uniform	4	2.6	1.3	0.3
	t_{10}	15.5	18.6	24.8	34.1
	t_4	62.1	78	91.6	97.9
	Logistic	16.9	21.7	27.6	39.6
	Lognormal	93.4	96.7	98.7	99.7
	Gamma(5, 1)	22.6	27.9	39.4	54.4
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	34.6	44.3	63.6	82.3
	χ^2_{20}	15.8	18.6	23.6	30.3
30	Normal	9.3	9.4	9.7	9.6
	Uniform	9.3	8.2	7.1	6.4
	t_{10}	10	10.3	11.5	12.6
	t_4	22.4	28.1	40.6	57.3
	Logistic	10.1	11.1	11.5	14.2
	Lognormal	80.9	90.7	97.7	99.6
	Gamma(5, 1)	11.5	12	13.4	15.5
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	13.1	14.5	19.4	22.2
	χ^2_{20}	10	10.9	11.2	12.3
50	Normal	8.8	9.5	9.7	9.7
	Uniform	9.5	9.2	8.8	8.5
	t_{10}	9.6	10.4	10.8	10.6
	t_4	17.2	21.3	28.3	40.3
	Logistic	9.9	10.3	10.6	11.2
	Lognormal	71	83.8	95.7	99.4
	Gamma(5, 1)	9.9	11	12.4	13.6
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	11.7	12.6	15.3	16.9
	χ^2_{20}	10.3	10.1	11	11.2
100	Normal	9.1	9.3	9.7	10
	Uniform	9.3	9.7	9.2	8.7
	t_{10}	9.1	9.6	10.3	10.8
	t_4	13.2	15.6	19.6	25.9
	Logistic	9.2	9.8	10.5	10.2
	Lognormal	52.2	66.4	84.8	96.3
	Gamma(5, 1)	9.5	10.3	10	11.1
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	9.9	11.4	12	12.9
	χ^2_{20}	9.1	9.4	10.4	9.8
250	Normal	8.8	9.5	9.8	9.5
	Uniform	9.9	10.1	9.6	9.7
	t_{10}	9.3	9.3	9.3	10.1
	t_4	11.6	12.4	13.8	15.8
	Logistic	9	9.9	10.1	9.5
	Lognormal	28	37.2	52.8	71
	Gamma(5, 1)	9.1	9.6	10	10
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	9.6	9.6	10.3	11
	χ^2_{20}	9.4	9.1	9.9	10.1
1000	Normal	9.7	9	9.5	10.1
	Uniform	10.5	8.9	9.1	10.3
	t_{10}	8.9	8.9	9.5	9.4
	t_4	9.2	10.1	10.7	11.5
	Logistic	9.3	9.1	9.3	9.8
	Lognormal	13.7	16.6	19.3	24.4
	Gamma(5, 1)	9.5	9.2	9.6	10
	Gamma($\frac{5}{2}$, 2) $\sim \chi^2_5$	9.3	9.5	9.5	9.8
	χ^2_{20}	8.9	9.2	9.3	9.7

TABLE 6 P_E with $CARL_0 = 100$ and $p = 10\%$

n	Distribution	$m = 30$	$m = 50$	$m = 100$	$m = 200$
5	Normal	9.3	9.8	9.2	9.7
	Uniform	5.9	4.4	3	1.5
	t_{10}	13.6	14.3	17.9	22.9
	t_4	35.6	45.3	60.9	76.5
	Logistic	14.5	16.7	19.6	26.5
	Lognormal	77	83.6	91.5	95.9
	Gamma(5, 1)	15.9	17.1	20.3	25.4
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	22.8	25.6	34.6	44.6
	χ^2_{20}	12	12.5	14.2	16.1
30	Normal	9.1	9.2	9.5	10.4
	Uniform	9.7	8.9	7.8	7.6
	t_{10}	9.4	9.9	11	11.2
	t_4	15.5	17.9	22.5	29.5
	Logistic	9.9	10.1	10.8	12
	Lognormal	52	63.1	77.8	91.1
	Gamma(5, 1)	9.9	10.2	11.2	11.5
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	10.9	12.1	12.6	13.6
	χ^2_{20}	9.5	9.8	10.2	10.8
50	Normal	9.2	9.2	10	9.6
	Uniform	9.8	9.7	9.5	9
	t_{10}	9.3	9.8	10.5	10.4
	t_4	13.8	15.1	17.8	22.1
	Logistic	10.1	10	9.8	11.5
	Lognormal	41.3	53	68.6	84.1
	Gamma(5, 1)	9.2	9.5	10.1	11.1
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	10.4	11.3	11.3	12
	χ^2_{20}	9.9	9.9	9.5	9.9
100	Normal	9.1	9.2	9.5	9.3
	Uniform	9.5	9.9	9.4	9.1
	t_{10}	9.1	9.7	10	9.8
	t_4	11.8	11.5	14.2	16.4
	Logistic	9.7	9.8	10.7	9.8
	Lognormal	29.2	36.9	49.6	66.5
	Gamma(5, 1)	9.5	9.6	9.7	10.4
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	9.5	9.9	9.9	10.5
	χ^2_{20}	9.7	9.5	9.5	9.8
250	Normal	9	9.5	9.3	9.5
	Uniform	10	10.2	9.8	9.9
	t_{10}	9.2	9.5	9.4	10
	t_4	10.5	10.8	12	13.4
	Logistic	9	9	9.3	9.5
	Lognormal	18	21.4	28.1	35.9
	Gamma(5, 1)	9.3	9.9	9.8	9.8
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	8.9	9.5	9.5	10
	χ^2_{20}	9.3	9.4	9.3	10
1000	Normal	9.4	10	9.3	9.4
	Uniform	10.6	8.9	9.1	10.3
	t_{10}	9.2	10	9.3	10.3
	t_4	9.5	10.1	10.3	10.3
	Logistic	8.7	9.4	9.1	10.2
	Lognormal	11.6	12.6	14.3	16.2
	Gamma(5, 1)	9	8.8	9.7	9.8
	Gamma$(\frac{5}{2}, 2) \sim \chi^2_5$	9	9.8	9.2	9.7
	χ^2_{20}	9.4	9.3	9.7	9.3

the sample size (n) is 1000. The implications are especially relevant within the field of SPM, where estimation of accurate tail behavior is important. The results indicate that the $\bar{X}$ control chart under normal theory should be used with caution, even with relatively large datasets.

REFERENCES

1. Alwan LC. The problem of misplaced control limits. *J R Stat Soc*. 1995;44(3):269-278.
2. Billingsley P. *Probability and Measure*, 32nd edn. New York: Wiley; 1995.
3. Cryer JD, Ryan TP. The estimation of sigma for an X chart: MR/d_2 or S/c_4. *J Qual Technol*. 1990;22(3):187-192.
4. Roes KC, Does RJMM, Schurink Y. Shewhart-type control charts for individual observations. *J Qual Technol*. 1993;25(3): 188-198.
5. Goedhart R, Schoonhoven M, Does RJMM. Correction factors for Shewhart X and $\bar{X}$ control charts to achieve desired unconditional ARL. *Int J Prod Res*. 2016;54(24):7464-7479.
6. Saleh NA, Mahmoud MA, Keefe MJ, Woodall WH. The difficulty in designing Shewhart $\bar{X}$ and X control charts with estimated parameters. *J Qual Technol*. 2015;47:127-138.
7. Goedhart R, Schoonhoven M, Does RJMM. Guaranteed in-control performance for the Shewhart X and $\bar{X}$ control charts. *J Qual Technol*. 2017;49(2):155-171.
8. Hall P. The distribution of means for samples of size N drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*. 1927;19(3-4):240-245.
9. Blyth CR. Convolutions of Cauchy distributions. *Am Math Mon*. 1986;93(8):645-647.
10. George EO, Mudholkar GS. On the convolution of logistic random variables. *Metrika*. 1983;30(1):1-13.
11. Fenton LF. The sum of log-normal probability distributions in scatter transmission systems. *IRE Trans Commun Syst*. 1960;8(1):57-67.
12. Nie H, Chen S. Lognormal sum approximation with type IV Pearson distribution. *IEEE Commun Lett*. 2007;11(10):790-792.8.
13. Walker GA, Saw JG. The distribution of linear combinations of t-variables. *J Am Stat Assoc*. 1978;73(364):876-878.
14. Witkovsky V. On the exact computation of the density and of the quantiles of linear combinations of t and F random variables. *J Stat Plann Inference*. 2001;94(1):1-13.
15. Gil-Pelaez J. Note on the inversion theorem. *Biometrika*. 1951;38(3-4):481-482.
16. Mehta NB, Wu J, Molisch AF, Zhang J. Approximating a sum of random variables with a lognormal. *IEEE Trans Wireless Commun*. 2007;6(7):2690-2699.

Leo C. E. Huberts is a PhD student in 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 big data in statistical 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.

Rob Goedhart is a PhD student in 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 control charting techniques with estimated parameters.

Mandla D. Diko obtained his MSc in Statistics from the University of Pretoria. He is currently working as a PhD student in the Department of Operations Management of the University of Amsterdam, the Netherlands. His research interests are in statistical process/quality control.

Ronald J.M.M. Does is a Professor of Industrial Statistics at the University of Amsterdam, Director of the Institute for Business and Industrial Statistics, Head of the Department of Operations Management, and Director of the Institute of Executive Programmes 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, health care engineering, and operations management methods.

How to cite this article: Huberts LCE, Schoonhoven M, Goedhart R, Diko MD, Does RJMM. The performance of $\bar{X}$ control charts for large non-normally distributed datasets. *Qual Reliab Engng Int*. 2018;34:979–996. https://doi.org/10.1002/qre.2287

APPENDIX A

Below, we give further details on the derivations of the convolutions given in Section 3.1.

A.1 | The normal distribution

The convolution of i.i.d. normal random variables can be found using the moment generating function approach. The moment generating function of a convolution of normally distributed variables $X \sim N(\mu, \sigma^2)$ is

$$M_{C_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right) = \exp\left(\sum_{i=1}^{n} \mu_i t + \sum_{i=1}^{n} \frac{\sigma_i^2 t^2}{2}\right)$$

which is just the moment generating function of a normal distribution, with mean $n\mu$ and variance $n\sigma^2$ and hence

$$C_n \sim N(n\mu, n\sigma^2).$$

A.2 | The uniform distribution

As shown by Hall,[8] the convolution of i.i.d. standard uniform random variables has a piecewise polynomial probability density function of degree $n-1$

$$f_X(x; n) = \frac{1}{2(n-1)!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (x-k)^{n-1} \mathrm{sgn}(x-k),$$

which we denote as the $IH(n)$ distribution.

A.3 | The Student's t_ν distribution with ν degrees of freedom

There is no closed form of the convolution of Student's t_ν distributed random variables $X \sim t_\nu$ for $\nu > 1$(see Walker and Saw[13]), but approximations do exist. We use an approximation based on the numerical inversion of the characteristic function given by Witkovsky.[14] The characteristic function of the sum of Student's t_ν distributed random variables, C_n, equals $\phi_{C_n}(t) = \phi_X^n(t)$, where the characteristic function of a single Student's t_ν distributed random variable equals

$$\phi_X(t) = \frac{1}{2^{\frac{\nu}{2}-1}\gamma(\frac{\nu}{2})} \left(\nu^{1/2}|\lambda t|\right)^{\nu/2} K_{\nu/2}\{\nu^{1/2}|\lambda t|\},$$

in which $K_\alpha\{z\}$ denotes the modified Bessel function of the second kind. The distribution function $F_{C_n} = \Pr\{C_n \leq x\}$ of C_n is found using the inversion formula of Gil-Pelaez[15]

$$F_{C_n}(x) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{sin(tx)\phi_{C_n}(t)}{t} dt.$$

A.4 | The logistic distribution

Now assume a logistic distribution for the random variable: $X \sim logistic(\mu = 0, s = 1)$. The standardized version of the sum of X_i can be written as

$$Z = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} = \sum_{i=1}^{n} \frac{X_i}{\sqrt{\frac{n\pi^2}{3}}} = \sqrt{\frac{3}{n}} \frac{C_n}{\pi},$$

which distribution can be approximated by

$$Z \sim \sqrt{\frac{\nu-2}{\nu}} t_\nu,$$

with $\nu = 5n + 4$ degrees of freedom. For more details on this approximation see George and Mudholkar.[10]

A.5 | The lognormal distribution

The characteristic and moment generating function of the lognormal distribution are undefined. The distribution of the convolution C_n can be approximated by 2 methods. In the first place, the Fenton-Wilkinson approximation will be used, as it is said to perform well in the tails of a lognormal distribution (see Mehta et al[16]). Secondly, an approximation based on the type IV Pearson distribution will be used.

A.6 | The Fenton-Wilkinson approximation

Consider the sum of lognormal (LN) random variables X_i, where each $X_i \sim LN(\mu, \sigma^2)$ with the expectation $E(X_i) = exp(\mu + 0.5\sigma^2)$ and variance $Var(X_i) = (exp(\sigma^2) - 1)exp(2\mu + \sigma^2)$. The expectation and variance of C_n are $E(C_n) = nE(X_i)$ and $Var(C_n) = nVar(X_i)$. The Fenton-Wilkinson approximation is a lognormal PDF with parameters μ_{C_n} and $\sigma^2_{C_n}$ such that $exp(\mu_{C_n} + 0.5\sigma^2_{C_n}) = nE(X_i)$ and $(exp(\sigma^2_{C_n}) - 1)exp(2\mu_{C_n} + \sigma^2_{C_n}) = nVar(X_i)$. Solving for μ_{C_n} and $\sigma^2_{C_n}$ results in a lognormal distribution for the sum: $C_n \sim LN(\mu_{C_n}, \sigma^2_{C_n})$.

A.7 | The type IV Pearson approximation

The type IV Pearson approximation was developed by Nie and Chen[12] and equates the first 4 central moments $(\mu_1, \mu_2, \mu_3, \mu_4)$ of the sum of lognormal distributions to the 4 parameters of the Pearson IV distribution. Denote the sum of lognormal random variables by C_n, where each $X_i \sim LN(\mu, \sigma^2)$.

Where the Fenton-Wilkinson approximation only uses the first 2 moments as parameters for a lognormal distribution to represent the sum of lognormal random variables C_n, the Pearson IV method uses 4 moments to approximate

C_n. The Pearson IV distribution is given by

$$f_{C_n}(x) = \frac{\left|\frac{\Gamma(m+\frac{\nu}{2}i)}{\Gamma(m)}\right|^2}{\alpha B(m-\frac{1}{2},\frac{1}{2})}\left[1+(\frac{x-\lambda}{\alpha})^2\right]^{-m} e^{-\nu arctan(\frac{x-\lambda}{\alpha})},$$

with location parameter λ, scale parameter $\alpha > 0$ and shape parameters $m > \frac{1}{2}$, $\nu \neq 0$. These 4 parameters can be found using the first 4 central moments of the sum of lognormal random variables C_n.

A.8 | The gamma $\Gamma(\alpha, \beta)$ distribution with parameters α and β

If X is gamma distributed $X_i \sim \Gamma(\alpha_i, \beta)$, then the moment generating function approach can be used to find the distribution of the convolution

$$M_{C_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n}\left(\frac{1}{1-\beta}\right)^{\alpha_i} = \left(\frac{1}{1-\beta}\right)^{\sum_{i=1}^{n}\alpha_i},$$

which is the moment generating function of a $\Gamma(\sum_{i=1}^{n}\alpha_i, \beta)$ distributed random variable. Therefore, $C_n \sim \Gamma(n\alpha, \beta)$.

A.9 | The chi-squared χ^2_ν distribution with ν degrees of freedom

Now assume that the distribution of X is chi-squared with ν degrees of freedom: $X \sim \chi^2_\nu$. The moment generating function of the chi-squared distribution can be used to find the convolution distribution of the sum of n i.i.d. random samples of X

$$M_{C_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n}(1-2t)^{-\frac{\nu}{2}} = (1-2t)^{\sum_{i=1}^{n}\nu_i/2},$$

which is the moment generating function of a $\chi^2_{n\nu}$ distribution, and therefore $C_n \sim \chi^2_{n\nu}$.