Control Charts for Skewed Distributions: Johnson's Distributions

Bachioua Lahcene*

Department of Mathematics, Preparatory Year College, University of Hail, P.O. Box. 2440, Hail 41581, Saudi Arabia

Abstract: In this study, some important issues regarding process capability and performance have been highlighted, particularly in case when the distribution of a process characteristic is non-normal. The process capability and performance analysis has become an inevitable step in quality management of modern industrial processes. Determination of the performance capability of a stable process using the standard process capability indices (Cp, Cpk) requires that the quality characteristics of the underlying process data should follow a normal distribution. Statistical Process Control charts widely used in industry and services by quality professionals require that the quality characteristic being monitored is normally distributed. If, in contrast, the distribution of this characteristic is not normal, any conclusion drawn from control charts on the stability of the process may be misleading and erroneous. In this paper, an alternative approach has been suggested that is based on the identification of the best distribution that would fit the data. Specifically, the Johnson distribution was used as a model to normalize real field data that showed departure from normality. Real field data from the construction industry was used as a case study to illustrate the proposed analysis.

Keywords: Statistical Process Control, Shewhart control charts, non-normal data, Johnson System of distributions.

1. INTRODUCTION

Statistical Process Control is a process improvement methodology widely used by modern manufacturing and service organizations. This methodology is mainly based on the use of control charts and frequency distributions of process and quality characteristics data. Common and well established control charts include the Shewhart control chart ($\bar{x}-R$ and $\bar{x}-S$ charts), the cumulative sum control chart (CUSUM) and the exponentially weighted moving average control chart (EWMA). Determination of the performance capability of a stable process using the standard process capability indices (C_p, C_{pk}) requires that the quality characteristics of the underlying process data should follow a normal distribution, it is becoming more critical than ever to assess precisely process losses due to non-compliance of customer specifications. To assess these losses, industry is widely using process capability indices.

Departures from this normality assumption could lead to erroneous results when applying conventional statistical capability measures which are based on the assumption. Many researchers have been investigating solutions to the non-normality problem. In case of data that do not obey normal distribution, the key issue in this analysis was to obtain correct estimate of process performance. When the distribution of a process characteristic is non- normal, process capability indices calculated using conventional methods could often lead to erroneous and misleading interpretation of the process's capability. Typically we assume that the processes follows normal probability distribution ensuring a high percentage of the process measurements falling between ± 3s of the process mean and the total spread amounts to about 6s variations. This article describes the estimation of C_p and C_{pk}, commonly used process capability indices (PCI), in case of nonnormal data using the characteristics of Pearson system distribution.

In process improvement strategies, these control charts are used to monitor product quality and detect special events occurring in the process that may cause out-of-control situations that would lead to an unstable and unpredictable process. Such processes deliver poor quality products to customers. Customers expect suppliers of products and services to provide proof of process control and process capability. organizations' management continuously improve processes, by making them more stable and capable to produce high quality, to meet customer specifications, and to achieve business excellence.

Standards (Shewhart) control charts are designed on the assumption that the process being monitored produces a quality characteristic that can be approximated by a symmetrical normal distribution, when only the innate sources of variability are present in the system. The central limit theorem can be used to approximate distributions to the normal distribution provided that the samples being measured and monitored would be large enough. However, in many industrial situations, this cannot be assured and the process output is not normally distributed and heavy tailed and skewed. Experience has showed that in

*Address correspondence to this author at the Department of Mathematics, Preparatory Year College, University of Hail, P.O. Box. 2440, Hail 41581, Saudi Arabia; E-mail: bachiouala@hotmail.com

some manufacturing processes, such as chemical processes parameters, cutting tool wear processes and some concrete production processes, the distribution are usually skewed. In this case, standard control charts based on normality assumptions can lead to erroneous conclusions regarding the stability and the capability of the process. Such wrong conclusions would cost manufacturing and service organizations big financial losses and lose customers to competitors.

With the advents in statistical theories and computing facilities, this can be easily solved, by understanding of distributions that provide good model for most non-normal quality characteristics. Such an approach has been reported in the technical literature [1-4]. Derya and Canan (2012) [4] developed standard control charts instead using Weibull, Gamma and lognormal distributions. Sherill and Johnson (2009) [3] showed the possibility to use exponential, Weibull and Lognormal distribution for transforming non-normal data for process control and process capability calculations. The objective of the present paper is to examine the use of the Johnson`s family of distributions to model control charts that can be used for process improvement purposes. A real field case study is presented for ready mixed concrete production plants where process distribution showed a skewed non-normal distribution.

2. JOHNSON'S DISTRIBUTIONS IN QUALITY IMPROVEMENT

Statisticians and quality professionals are often faced with the problem of summarizing a set of data by means of a mathematical function which fits the data and allowsss on obtaining estimates of percentiles. Frequently, statisticians and quality professionals usually have insufficient theoretical grounds for selecting a model like normal, gamma or extreme-value distributions for a "real world" data set [5]. Usually data are obtained and empirical methods are used to draw conclusions and make decisions on process and quality improvement in real business situations. The fitting of empirical distributions to data has a long history, and many different procedures have been advocated. The most common of these is the use of normal distribution. The central limit theorem leads one to expect this distribution to provide reasonable representation for many, but not all, physical phenomena [6].

Although models like gamma, log-normal and beta distributions do lead to a wide diversity of distribution shapes, they still do not provide the degree of generality that is frequently desirable. In 1949, Johnson derived a system of curves that has the flexibility of covering a wide variety of shapes. This system has the practical and theoretical advantages of being able to transform these curves to the normal distribution. The Johnson system is able to closely approximate many of the standard continuous distributions through one of the three functional forms and is thus highly flexible. The Johnson system provides one distribution corresponding to each pair of mathematically possible values of skewness and kurtosis. Any data set can be fitted by a member of the Johnson families such as S_U, S_L, and S_B. This motivated us to use Johnson system for the analysis of micro array data [7]. This family of distributions, published by the statistician N.L. Johnson in 1949, is perhaps the most versatile choice. It is based on a transformation of the standard normal variable, and includes four forms:

1. Unbounded: the set of distributions that go to infinity in both the upper or lower tail.

2. Bounded: the set of distributions that have a fixed boundary on either the upper or lower tail, or both.

3. Log Normal: a border between the Unbounded and Bounded distribution forms.

4. Normal: a special case of the unbounded form.

The fact that the Johnson system involves a transformation of the raw variable to a normal variable allowsss estimates of the percentiles of the fitted distribution to be calculated from the Normal distribution percentiles, for use in control limit calculations (on the Individual-X chart or the $\bar{x} - R$ charts) or for Capability Analysis. Thus, although capability indices and control limits are generally only defined for normal variables, this approach allowsss their calculation for all distribution types [7]. In this study, the authors applied the Johnson system, which includes the S_U, S_L, and S_B distributions, as the Johnson system exhibits the key property of being able to accommodate all theoretically feasible skewness-kurtosis combinations (Figure **1**).

The standard process capability analysis is one of many statistical process control widely used in manufacturing and services engineering. It is based on the assumption that process data are normally distributed. When this condition cannot be guaranteed, either capability indices should be computed based on distributions other than normal, or the data should be transformed so that it conforms better to the normal distribution [1]. Sherill and Johnson (2009) [3], and

many others showed that the use of Box-Cox and the Johnson transformations would help the quality professional to perform correct process analysis using both control charts for process stability and capability indices for process capability to meet customer specifications. In addition, it is worth mentioning here that in a recent study, [8], showed compressive strength of concrete elements in buildings are best modeled using log-normal and the Johnson SB distributions.

3. MATHEMATICAL FORMULATION OF THE JOHNSON'S DISTRIBUTIONS

As stated earlier, when process data exhibit non-normal distribution, it is erroneous to draw standard control charts for process improvement and perform capability analysis. The practical solution is to transform the data and drive them towards normality, using common and well established probability distributions, such as Box-Cox, log-normal or the Johnson distribution. Such an approach has been used in the open literature. Basically the Johnson transformation computes an optimal transformation function from three flexible distribution families (S_U, S_B, and S_L). This makes this transformation more powerful than other distribution [3].

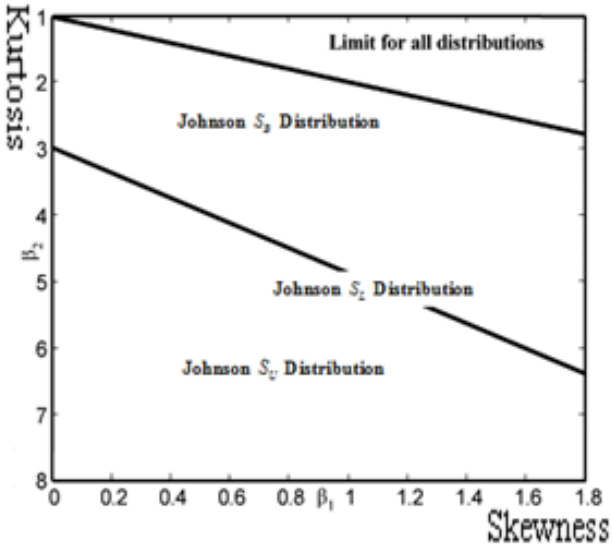

Figure 1: The Skewness and Kurtosis Plane for the Johnson Distributions.

These translations transform any continuous random variable X into a standard normal variable Z using general form:

$$Z = a + bg(\frac{X-\mu}{\sigma}) \tag{1}$$

Where: a and b are shape parameters, μ is a location parameter, and $g(x)$ is a function defining the Johnson system of families, determined as:

$$g(x) = \begin{cases} \ln(x), & \text{for the lognormal family,} \\ \ln\left(x + \sqrt{x^2+1}\right), & \text{for the unbounded family,} \\ \ln\left(\frac{x}{1-x}\right), & \text{for the bounded family,} \\ x, & \text{for the normal family.} \end{cases}$$

As discussed in [5], the above system has the flexibility to match any feasible set of values for the mean, variance, skewness, and kurtosis coefficients. With this system, the skewness and kurtosis also uniquely identify the appropriate form for the (g) function.

3.1. Johnson's Translation System

Johnson proposed three normalizing transformations having the general form:

$$Z = \gamma + \sigma\, f\left(\frac{X-\mu}{\lambda}\right), \tag{2}$$

Where $f(.)$ denotes the transformation function, Z is a standard normal random variable γ and σ are shape parameters, λ is a scale parameter and μ is a location parameter. Without loss of generality, it is assumed that $\sigma \succ 0$ and $\lambda \succ 0$.

The first transformation proposed by Johnson defines the lognormal system of distributions denoted by S_L:

$$Z = \gamma + \sigma \ln\left(\frac{X-\mu}{\lambda}\right) = \overset{*}{\gamma} + \sigma \ln(X-\mu), \quad X \succ \mu, \tag{3}$$

The bounded system of distributions S_B is defined by:

$$Z = \gamma + \sigma \ln\left(\frac{X-\mu}{\mu+\lambda-X}\right) = \quad \mu \prec X \prec \mu+\lambda, \tag{4}$$

S_B curves cover bounded distributions. The distributions can be bounded on the lower end, or the upper end, or both. This family covers gamma distributions, beta distributions and many others.

The unbounded system of distributions S_U is defined by:

$$Z = \gamma + \sigma \ln\left[\left(\frac{X-\mu}{\lambda}\right) + \left\{\left(\frac{X-\mu}{\lambda}\right)^2 + 1\right\}^{1/2}\right] = \gamma + \sigma \sinh^{-1}\left(\frac{X-\mu}{\lambda}\right), \quad -\infty \prec X \prec +\infty \tag{5}$$

The S_U curves are unbounded and cover the t and normal distributions, among others.

3.2. Johnson's Family of Distributions

The Johnson family of distributions is made up of three distributions, Johnson S_U, Johnson S_B and lognormal. It covers any specified average, standard deviation, skewness and kurtosis. Together they form 4-parameter family distributions that cover the entire skewness-kurtosis region other than the impossible region. The Johnson S_U distribution covers the area above the lognormal curve and the Johnson S_B covers the area below the lognormal curve. A family of distributions is several distributions combined so that they cover a well defined region in a skewness and kurtosis plot (lognormal family of distributions, negative lognormal and normal distributions,..). Readers can find detailed developments about the Johnson family of distributions in reference books [6].

This family of distributions is usually parameterized as a function of skewness and kurtosis. Skewness is a measure of non symmetry in the data, so for a normal distribution it takes the value of zero. Negative values for the skewness indicate that data are skewed left, and positive values indicate that data are skewed right. On the other hand, kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. The kurtosis for a normal distribution is 3.0. A kurtosis value larger than 3.0 indicates a "peaked" distribution and a kurtosis value less than 3.0 indicates a "flat" distribution. Thus, both can be seen as measures of shape of the distributions.

When the data is not normally distributed, capability analysis can still produce useful results by using nonparametric indices, or by transforming the data so that it conforms better to normal distribution than its original form. In order to find whether a process is in control, quality control charts like R and X Bar charts can be used. Earlier work of Lovelace and Swain (2009) [9] has been extended for this distributional assumption. Quantiles are estimated by probability plotting technique and then control limits are obtained to determine whether the process is in statistical control or not [9].

4. APPLICATION OF JOHNSON'S SYSTEM OF DISTRIBUTIONS FOR REAL FIELD DATA

To illustrate the above analysis, real field data from the construction industry business was chosen as a case study. Data from Ready mixed concrete plants were gathered and analyzed using Minitab 16 statistical software. The observed quality characteristic was the compressive strength (kgf/cm^2) of concrete as defined by international quality standards [10]. The gathered data consisted of 22 samples of concrete with a nominal specification 350 kgf/cm^2. The sampling process consists of a sample size of 3 spanning over a period of 30 days. These data are presented in Table **1**.

Initial analysis of the data of the concrete using standard $\overline{X}$-chart (Figure **2**) showed that the process is out of statistical control; this would mean the existence of special causes of variation affecting the process.

Table 1: Data for Compressive Strength for Ready Mixed Concrete (Kgf/cm^2)

Sample	Cylinder 1	Cylinder 2	Cylinder 3
1	353.8	363	360.6
2	357.8	358.7	370.9
3	365.2	360	356.6
4	340.4	335.2	330.1
5	359.6	358.1	351.2
6	368.1	366.7	369.3
7	357.9	355.0	350.6
8	337.8	352.6	361.6
9	359.1	349.2	363.7
10	361.1	358.2	358.3
11	358.3	345.7	341.7
12	357.3	359.2	356.9
13	352.6	363.1	374.6
14	360.8	356.2	352.7
15	347.5	339.8	354.3
16	358.2	359.5	353.9
17	375.2	372.5	370.2
18	357.5	359.5	348.9
19	343.2	355.8	362.4
20	362.1	356.6	359.1
21	365.2	362	359.4
22	361.3	346.8	339.0

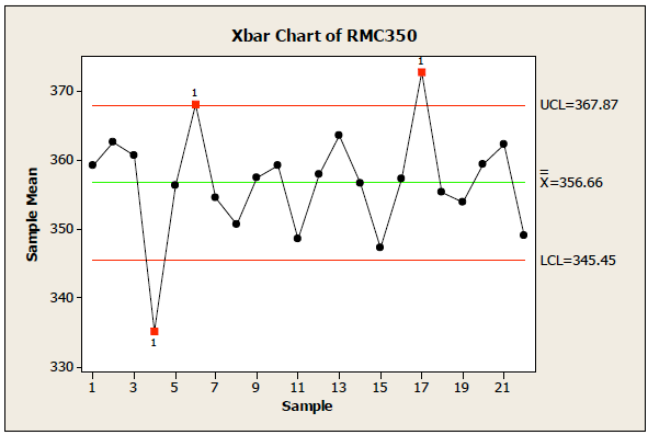

Figure 2: Standards $\bar{x}$ chart for the concrete compressive strength.

out-of –control shown from the $\bar{x}$ control chart was drawn based on the assumption of normally distributed concrete data. Is this assumption correct? If not what would be the best distribution that fits these real field data. To answer this question, distributions identification was carried out for the data, and the outcome is presented in Figure **3** as probability plots. From this figure, it can be seen that the compressive strength of concrete does fit neither the normal, nor the exponential, nor the Weibull, nor the lognormal distributions. It is very obvious that the exponential distribution is a poor model for the concrete data. The Johnson distribution would be an alternative for the model [3, 8]. The transformed data by the Johnson system are illustrated in Figure **4**, where it can be seen that this distribution shown as a mixture would be the best model of these concrete data. From this figure, it can be seen that within the interval percentile ranging from 1.054 to 98.94, would be the best fit of the data. Normality within this interval can be guaranteed. These correspond to the lower control limit and the upper control limit for the normalized data which are UCL=375.2 (kgf/cm^2) and LCL=330.1 (kgf/cm^2). These control limits will be used as the new control limits for the $\bar{x}$ chart as shown in Figure **5**. It is clearly shown that the control chart with the new control limits indicate totally the opposite of the early conclusion drawn from the standard control chart. The process is shown to be in statistical control.

5. CONCLUSIONS

Most statistical process control charts require that the quality characteristic being monitored is normally distributed. If, in contrast, the quality distribution of the quality characteristic of interest is not normal, the conclusions drawn from control charts on the stability of the process may be misleading and highly erroneous.

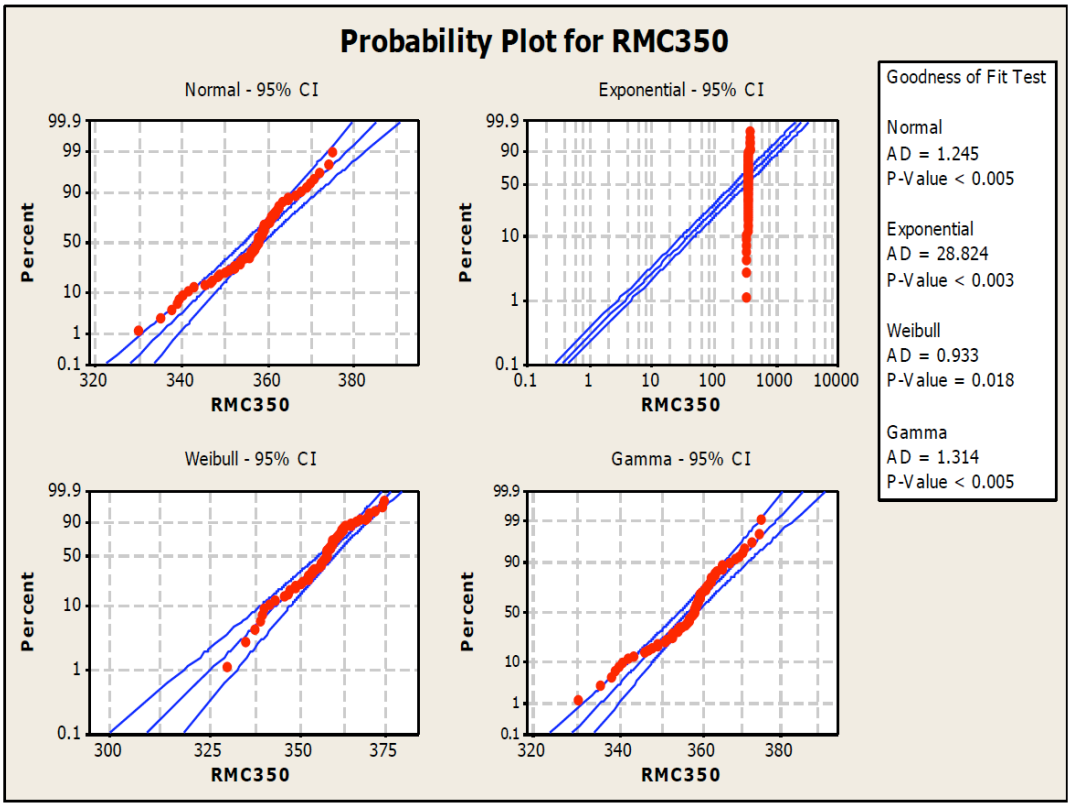

Figure 3: Probability Plots for the Concrete Compressive Strength.

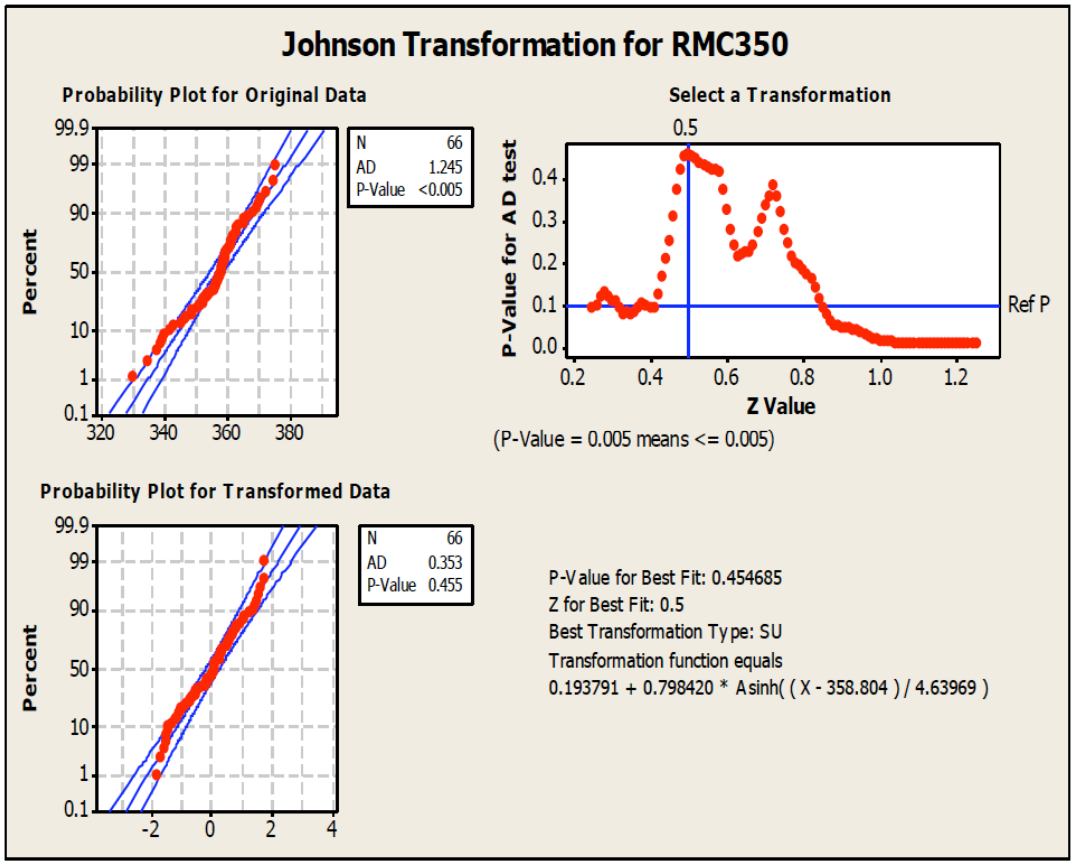

Figure 4: Probability Plots for the Johnson Transformed data of Concrete Strength.

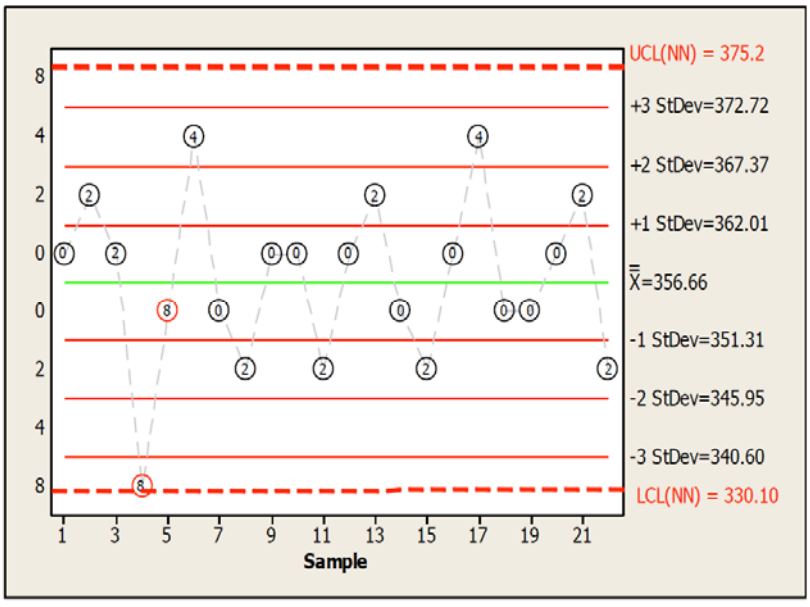

Figure 5: $\bar{x}$ -chart with the new Control Limits using Johnson transformations.

In this paper, an alternative approach has been suggested that is based on the identification of the best distribution that would fit the data. Specifically, the Johnson distribution was used as a model to normalize real field data that showed departure from normality.

Real field data from the construction industry was used as a case study to illustrate the analysis. The assumption of normality when the data were not normally distributed led to conclude that the monitored process was out of statistical control, indicating that

some special causes are present in the process, which would require some intervention from management on the process to get rid of the special cause of variation to occur again. This would certainly cost the organization some cost. However, when the data were transformed and brought to normality through Johnson transformations, and the new control limits calculated, the new control chart indicated no sign of special causes of variation.

REFERENCES

[1] Farnum NR. Using Johnson Curves to Describe Non-Normal Process Data. Quality Eng 1996; 9(2): 329-336. http://dx.doi.org/10.1080/08982119608919049

[2] Chou Y, Polansky AM, Mason RL. Transforming Non-normal Data to Normality in Statistical Process Control. J Quality Technol 1998; 30: 133-141.

[3] Sherill RW, Johnson LA. Calculated Decisions. Quality Progr 2009; 42(1): 30-35.

[4] Derya K, Canan H. Control Charts for Skewed Distributions: Weibull, Gamma and Lognormal. Metodoloski zvezki 2012; 9(2): 95-106.

[5] Johnson NL. Systems of frequency curves generated by methods of translation. Biometrika 1949; 36: 149-176. http://dx.doi.org/10.1093/biomet/36.1-2.149

[6] Hahn J. Gerald and Shapiro S. Samuel, Statistical models in Engineering, John Wiley and Sons 1967.

[7] Johnson NL, Kotz S, Balakrishnan N. Continuous Univariate Distributions, Second Edition, New York: John Wiley & Sons 1994.

[8] Kilink K, Celik A, Tuncan M, Tuncan A, Arslan G, Arioz O. Statistical distributions of in situ microcore concrete strength. Construct Build Mater 2012; 26(1): 393-403. http://dx.doi.org/10.1016/j.conbuildmat.2011.06.038

[9] Lovelace CR, Swain JJ. Process capability analysis methodology for zero bound, non-normalprocess data. Quality Eng 2009; 21: 190-202. http://dx.doi.org/10.1080/08982110802643173

[10] ACI Committee 214, Evaluation of Strength Test Results of Concrete (ACI 214R-02) 2005.

Received on 13-10-2014 Accepted on 12-01-2015 Published on 21-05-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.02.8