Some Misconceptions about the Normal Distribution

By Keith M. Bower, M.S.

Reprinted with permission from the American Society for Quality.

As part of a Six Sigma training course, practitioners are introduced to arguably the most important probability distribution in statistics: the *normal* distribution. Statistical procedures are often based upon the assumption that data collected for an analysis are drawn from a normal distribution.

This article discusses three misconceptions regarding the use of the normal distribution in theory and practice:

1. Something is "wrong" if the distribution is non-normal
2. The larger the sample size, the closer it approximates a normal distribution
3. Capability estimates do not depend on normality

The Normal Distribution

A normal distribution can be described solely by the arithmetic mean (μ) and standard deviation (σ). These parameters may be estimated by the sample mean ($\overline{X}$) and sample standard deviation (S) respectively.

A normal distribution is typically expressed in statistical shorthand as N(μ, σ^2). For example, a normal distribution with a mean of 12 and standard deviation of 5 is written N(12, 25).[1]

The probability density function (pdf) for a normal distribution is:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

It is alleged that Abraham de Moivre was the first to propose the normal distribution, in a supplement to a letter dated November 12, 1733. However, the name of Carl Friedrich Gauss is more closely associated with the normal, or *Gaussian*, distribution. The movement away from the term *Gaussian* occurred during the first part of the 20th century. According to Helen M. Walker[2], the origin of the term "normal" is obscure, though the *Encyclopedia of Statistical Sciences*[3] states that C. S. Peirce may have been the first to use it in the mid 1870s.

Misconception 1:
Something is "wrong" if the distribution is non-normal

Often, distributions other than the normal are more appropriate for a given set of data. In particular, when a naturally occurring boundary exists (e.g., zero, with cycle time data), the assumption of normality may not be sensible because the normal distribution has positive probability throughout the entire real number line (i.e., from negative to positive infinity).

Some Six Sigma practitioners are encouraged to discover *why* the data are nonnormal and to continue to look for explanations until normality is obtained. This may be poor advice and frustrate the investigator because, despite best efforts, the assumption of normality frequently cannot (reasonably) be obtained.

The misunderstanding may be due to an unwarranted inference from the name of the distribution itself. Six Sigma practitioners, especially those new to statistical theory, may believe that it is "normal" to see such a distribution in practice. Though the normal distribution may be a reasonable assumption for many processes, it is not reasonable for all processes.

Furthermore, practitioners are occasionally led to believe that an approximately normal distribution implies that a process is in statistical control. Again, the inference is not valid; through control charting a practitioner can assess the stability of a process.[4]

Misconception 2:
The larger the sample size, the closer it approximates a normal distribution

This misconception may be due to a misunderstanding of the central limit theorem (CLT). As discussed by Robert V. Hogg and Johannes Ledolter, the CLT may be stated as follows:

> If $\overline{X}$ is the mean of a random sample $X_1, X_2, \ldots X_n$ from a distribution with mean μ and finite variance $\sigma^2 > 0$, then the distribution of
>
> $$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{(\sum_{i=1}^{n} X_i) - n\mu}{\sigma\sqrt{n}}$$
>
> approaches a distribution that is N(0,1) as n becomes large.[5]

In other words, when sampling from such a distribution (normal or otherwise), as the sample size increases, the distribution of $\overline{X}$ gets closer to a normal distribution. When sampling from a normal distribution, the distribution of $\overline{X}$ will, necessarily, be normal.

Of course, none of this implies that when larger samples are taken from a nonnormal distribution the underlying distribution itself becomes normally distributed. Rather, one would witness a clearer picture of the actual

(nonnormal) distribution itself. The CLT discussed above is involved with assessing the distribution of $\overline{X}$, not X—the individual values themselves.

Misconception 3:
Capability estimates do not depend on normality

Quality practitioners are keenly aware that reasonable estimates of process parameters greatly depend on the stability of the process. Such estimates, for example $\overline{X}$ for μ, and S_{within} for σ_{within}, can then be used to compute capability estimates such as $\hat{C}_p$ and $\hat{C}_{pk}$, where:

$$C_p = \frac{(USL - LSL)}{6\sigma_{within}} \quad \text{and} \quad C_{pk} = \min\left\{\frac{(\mu - LSL)}{3\sigma_{within}}, \frac{(USL - \mu)}{3\sigma_{within}}\right\}$$

These capability indices are widely employed in the automotive industry and beyond. Unfortunately, practitioners may be tempted to use the parameter estimates in a situation where the normal distribution is not an adequate fit to the continuous dataset. As was shown by Steven E. Somerville and Douglas C. Montgomery, such estimates of process capability can be markedly different from the advertised rate, depending on the actual form of the underlying distribution.[6]

Again, one argument states that with increasing sample sizes, the normality assumption is not an important factor in capability analyses. Of course, this argument can be countered with the discussion in misconception 2.

With nonnormal data, several strategies can lead to meaningful capability estimates. For example, practitioners may seek to transform the data, and use the normal distribution with the resulting values, or use an alternative distribution.

Implications for Statisticians and Trainers

There are many instances in which the assumption of normality is extremely important, e.g., Bartlett's test for equal variances. There are also situations in which the procedure may be robust to the assumption of normality, depending on the data collection procedure. Examples include t-tests and the ANOVA procedure.[7]

In exposing the three misconceptions discussed here, this discussion does not intend to discount them completely. The intent, rather, is to increase awareness of the limitations of extreme and dogmatic thinking regarding the normal distribution.

References

[1] For more information on these terms, see Robert V. Hogg and Allen T. Craig, *Introduction to Mathematical Statistics,* 5th ed. (New Jersey: Prentice-Hall, Inc., 1995), 140.

[2] Helen M. Walker, *Studies in the History of Statistical Method With Special Reference to Certain Educational Problems* (Baltimore: Williams & Wilkins Company, 1929), 185.

[3] Samuel Kotz, Norman L. Johnson and Campbell B. Read, eds., *Encyclopedia of Statistical Sciences* (New York: John Wiley & Sons, Inc., 1985), 6:348.

[4]For more information on normality and control charting, see Douglas C. Montgomery, *An Introduction to Statistical Quality Control*, 4th ed. (New York: John Wiley & Sons, Inc., 2001), 232, 254-255.

[5]Robert V. Hogg and Johannes Ledolter, *Applied Statistics for Engineers and Physical Scientists*, 2nd ed. (New York: Macmillan Publishing Company, 1992), 154.

[6]Steven E. Somerville and Douglas C. Montgomery, "Process Capability Indices and Non-Normal Distributions," *Quality Engineering* 9, no.2 *(*1996): 305-316.

[7]For further information on this aspect of robustness to normality, see George E. P. Box, William G. Hunter, and J. Stuart Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building* (New York: John Wiley & Sons, Inc., 1978), 95-101, 188; and Joshua M. Tebbs and Keith M. Bower, "Some Comments on the Robustness of Student t Procedures," *Journal of Engineering Education* 92, no. 1 (2003): 91-94.