1 Chebyshev's Theorem

Abstract

Chebyshev's Theorem constrains how measurement data are distributed about their mean. It provides a prediction for the fraction of a population that falls within a specified number of standard deviations of the mean. Unlike the normal distribution and other probability distributions Chebyshev is very general and does not require that any special conditions or requirements are satisfied. Examples of setting specifications and estimating the process fraction defective with Chebyshev are presented.

Key Words

Chebyshev, Tschebyshev, probability distribution, normal distribution, Camp-Meidel

Introduction

The first steps in determining how a distribution behaves are to determine or estimate the population mean and standard deviation, μ and σ. As soon as these values are determined the shape of the histogram is constrained, that is, the data are required to fall in a certain pattern about the mean. This constraint is called Chebyshev's Theorem.

Where the Technique is Used

Chebyshev's theorem is used to construct intervals, such as for tolerances or hypothesis tests, when the distribution of the random variable under consideration is unknown. Chebyshev's Theorem is useful because it always works but its usefulness is compromised because it is a very conservative technique.

Chebyshev's Theorem can be used to:

- Construct an interval for a measurement variable, such as for setting tolerances.
- Construct an acceptance interval for a statistic in a hypothesis test.
- Estimate the fraction of a population that falls outside an interval, such as to estimate the fraction of defective parts produced by a process.

Data

The data must be variables data (measurement data) or attribute data which may be modelled with a continuous probability distribution. If the population mean and standard deviation are unknown then a minimum sample of size $n = 30$ is recommended.

Assumptions

- The data come from a single stable process.
- Two-sided (upper and lower) limits are required.

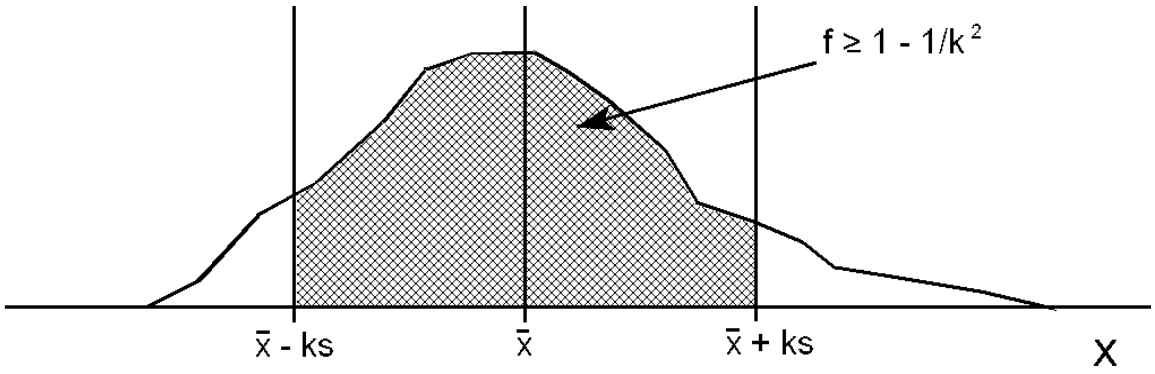

Figure 1: Chebyshev's Theorem

Chebyshev's Theorem

Chebyshev's theorem states, in words: *the fraction of a population that falls within k standard deviations of the mean is at least* $1 - \frac{1}{k^2}$. We can express this as an inequality:

$$f\left(\mu - k\sigma < x < \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}$$

where the leading $f()$ is mathematical shorthand for "the fraction of". Note that the x, which represents all of the possible values in the data set, is snuggled between a lower bound $\mu - k\sigma$ and an upper bound $\mu + k\sigma$. The $\geq$ symbol means "at least" and $1 - \frac{1}{k^2}$ gives the numerical value of the fraction. Sometimes it is more convenient to write $1 - \frac{1}{k^2}$ as $\frac{k^2-1}{k^2}$. Reread the English language statement of Chebyshev's theorem and make sure you understand the equation and what each symbol means.

A graphical representation of Chebyshev's theorem is possible. Consider the distribution or histogram of the population shown in Figure 1. The horizontal axis is the measurement axis for the variable x. The upper and lower bounds on x are indicated. If the total area under the histogram is scaled to 1 then the shaded area within these bounds corresponds to the fraction of the population that falls in that range. Compare the English language statement and the Chebyshev inequality to the picture and make sure you understand how they are related.

Example 1: *A part is produced with upper and lower specs of* $USL/LSL = 0.800 \pm 0.040$ *inches. If the mean part dimension is* $\mu = 0.800$ *inches and the standard deviation of the parts is* $\sigma = 0.010$ *inches estimate the fraction of the product that falls within the spec.*

Solution: *Since the lower spec falls* 4 *standard deviations below the mean and the upper spec falls* 4 *standard deviations above the mean we need to determine the fraction of the parts that fall within* $k = 4$ *standard deviations of the mean. From Chebyshev we have:*

$$f\left(LSL < x < USL\right) \geq 1 - \frac{1}{k^2}$$

By substituting the appropriate numerical values into this equation we find:

$$f\left(0.760 < x < 0.840\right) \geq 1 - \frac{1}{4^2} = \frac{15}{16}$$

or at least $\frac{15}{16}$*ths of the parts produced fall in spec. A sketch of the situation is shown in Figure 2.*

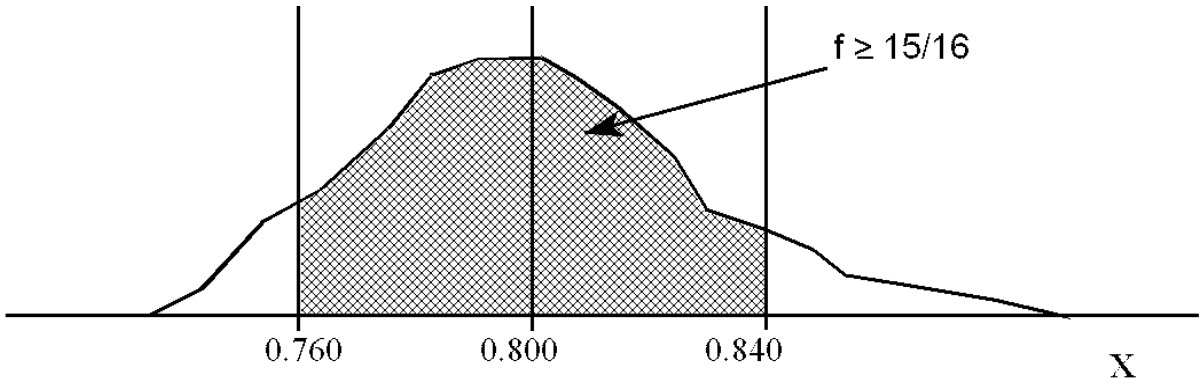

Figure 2: Chebyshev's Solution for Example 1

Chebyshev with Standardized Units (z units)

Since every problem we encounter has a different mean and standard deviation we require a general method of describing how an x value falls within the distribution of x values. The accepted technique is to use standardized units, also called z units or z values. Every possible x value has its corresponding z value. If the population mean μ and the population standard deviation σ are known then the z value for an x value is found from the transformation equation:

$$z = \frac{x - \mu}{\sigma}$$

For the lower bound $\mu - k\sigma$ the corresponding z value is:

$$z = \frac{(\mu - k\sigma) - \mu}{s} = -k$$

Similarly, the upper bound $\mu + k\sigma$ has its corresponding z value:

$$z = \frac{(\mu + k\sigma) - \mu}{s} = +k$$

The z transform can be used to express Chebyshev's theorem from its original as:

$$f\left(-k < z < k\right) \geq 1 - \frac{1}{k^2}$$

This form makes the English language statement of Chebyshev easier to understand: *the fraction of the values that fall within k standard deviations of the mean (above or below) is at least* $1 - \frac{1}{k^2}$. This new form permits a very important modification to be added to Figure 1. Consider the addition of the second measurement scale, the z scale, in Figure 3. With both the x and the z scales represented it is possible to talk in terms of the actual measurement units (the x's) or the standardized units (the z's).

Example 2: *The voltage at a critical point on a circuit board has a mean of* 28 *volts and a standard deviation of* 0.4 *volts. Estimate the fraction of the circuit boards that fall within the spec of* $USL/LSL = 29.4/26.6$ *volts.*

Solution: *The situation is shown in Figure 4. We must determine k, the number of standard deviations that the spec limits fall from the mean. This is most easily done using the z transform. For the lower spec limit we have:*

$$z = \frac{26.6 - 28}{0.4} = -3.5$$

and for the upper spec limit we have:

$$z = \frac{29.4 - 28}{0.4} = 3.5$$

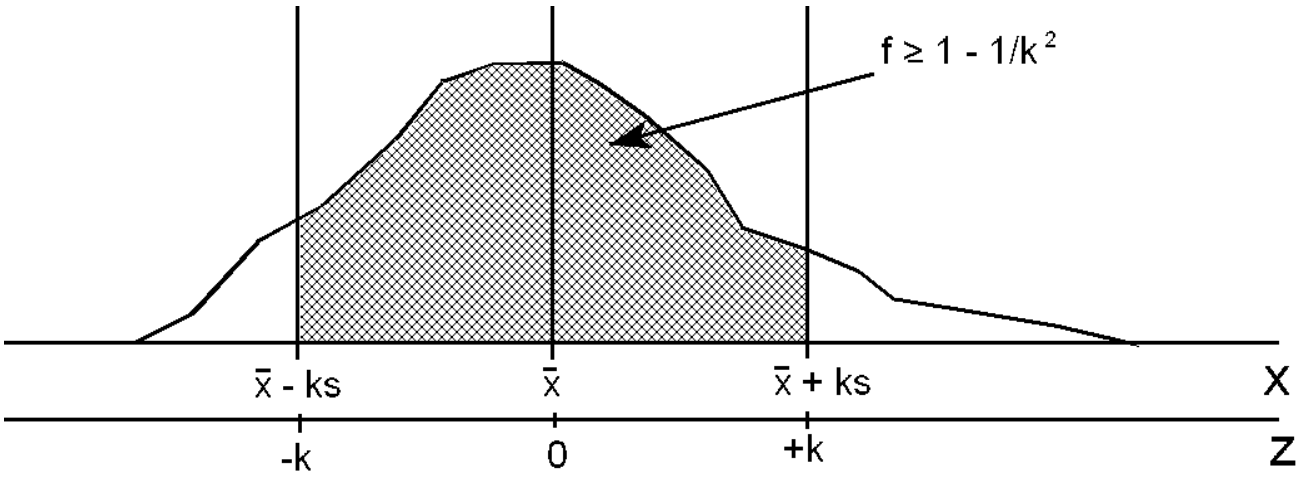

Figure 3: Chebyshev with Standard (z) Units

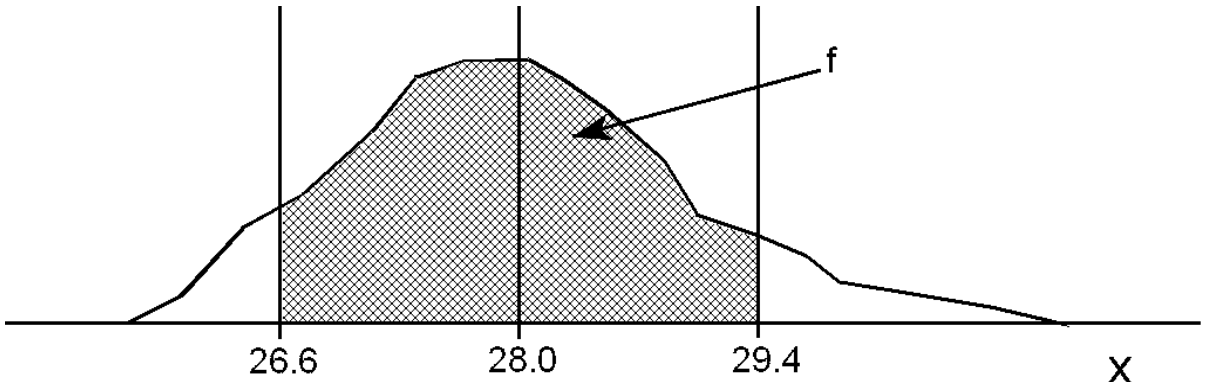

Figure 4: Chebyshev's Solution for Example 2

Apparently $k = 3.5$ *so the fraction of the parts that fall within the spec is* $f \geq 1 - \frac{1}{3.5^2} = 0.918$. *We can summarize this problem by writing Chebyshev's inequality:*

$$f\left(26.6 < x < 29.4\right) \geq 0.918$$

or

$$f\left(-3.5 < z < 3.5\right) \geq 0.918$$

In words, the fraction of the circuit boards that fall between 26.6 *and* 29.4 *volts (or within* 3.5 *standard deviations of the mean) is at least* 91.8%. *An updated picture of the situation is shown in Figure 5.*

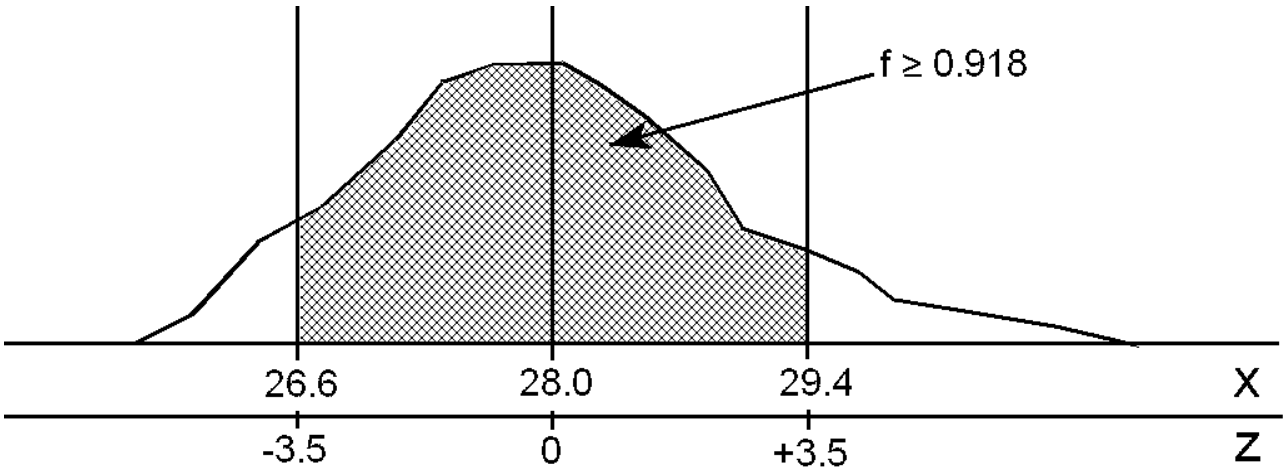

Figure 5: Chebyshev's Solution for Example 2 With Standard (z) Units

Using Chebyshev to Set a Specification

So far the examples cited use Chebyshev to determine the fraction of the parts that fall within the spec. In many cases a designer knows $\bar{x}$ and s and he must determine the spec to guarantee that no more than a certain fraction of the material produced is defective. Chebyshev can still be used to solve this problem.

Example 3: *The mean and standard deviation of the fill gas pressure in light bulbs are* $\mu = 42000$ *and* $\sigma = 140$ *pascals. Determine the specs that will guarantee that no more than* 1% *of the light bulbs will be defective.*

Solution: *We must first determine the* k *value for the Chebyshev inequality. Since we wish to have no more than* 1% *defectives we need at least* 99% *of the product to be in spec, or* $f \geq 0.99$*. Since Chebyshev says* $f \geq 1 - \frac{1}{k^2}$ *we must have* $1 - \frac{1}{k^2} = 0.99$*. Solving this equation for* k *we find that* $k = 10$ *so the desired spec is:*

$$f\left(LSL < x < USL\right) \geq 1 - \frac{1}{k^2}$$

$$f\left(\mu - k\sigma < x < \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}$$

$$f\left(42000 - 10 \cdot 140 < x < 42000 + 10 \cdot 140\right) \geq 1 - \frac{1}{10^2}$$

$$f\left(40600 < x < 43400\right) \geq 0.99$$

The situation is summarized in Figure 6.

Questions and Answers

Why is Chebyshev's theorem so important?

Chebyshev does not require that any underlying assumptions about the shape or nature of the distribution be satisfied. It always works, even if the distribution under study is completely pathological.

How good is the estimate that Chebyshev provides?

Not very good. Chebyshev gives the worst case analysis. That's the penalty for always working. With more data and more study you can generally do much better than Chebyshev, but if you don't know anything at all about how the data behave Chebyshev's always there to save you.

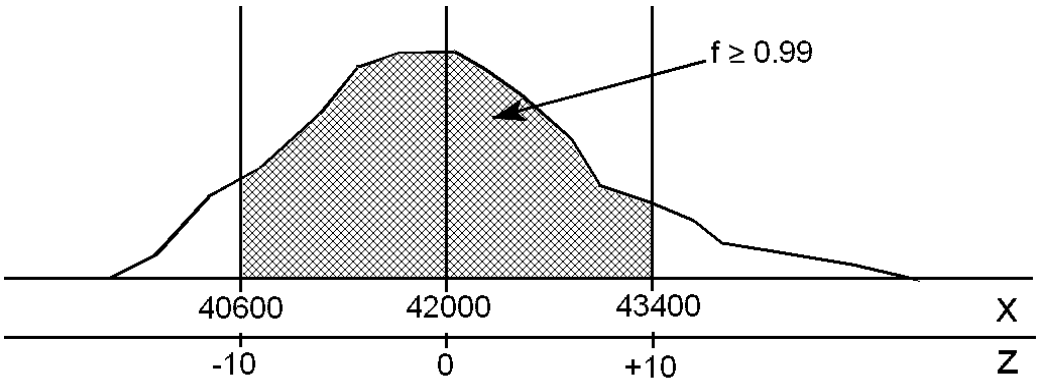

Figure 6: Chebyshev's Solution for Example 3

If the estimate provided by Chebyshev is so bad what other choices do I have?

There are many choices but they all require that you know more about the how the data behave. If the distribution is monomodal, that is, if the histogram is smooth and has only a single peak then the Camp-Meidel inequality can be used:

$$f\left(\mu - k\sigma < x < \mu + k\sigma\right) \geq 1 - \frac{1}{2.25k^2}$$

In many situations the normal (bell-shaped) distribution applies to the data. The normal distribution is well known but people commonly apply it *without checking to see if the normal distribution is appropriate.* Before you can use the normal distribution you must test to see if it is a good match to the data.

How do the Chebyshev, Camp-Meidel, and normal distribution predictions compare?

The following table shows the predicted fraction of the population that falls within k standard deviations of the mean for each model:

k	Chebyshev	Camp-Meidel	Normal
1	NA	NA	0.6826
2	≥ 0.750	≥ 0.889	0.9545
3	≥ 0.889	≥ 0.951	0.9973
4	≥ 0.938	≥ 0.972	0.9999

How do I tell if my data follow the normal distribution?

A graphical test called a normal probability plot can be used. To do this test you will need a sample of at least 30 measurements from the process under study. For larger data sets there are several goodness of fit tests, like the Chi-Square Test, Wilk-Shapiro, Lilliefors, Kolmogorov-Smirnov, and others.

How often can I expect to use Chebyshev?

In most cases Chebyshev doesn't tell you enough so it doesn't get used very often. You use it when you don't know how the data behave and you've got no other choice. Chebyshev's still important; it should always be there in the back of your mind as a safety net. Remember, it can't get any worse than Chebyshev.

How do I use Chebyshev if the mean is not centered in the spec limits?

You can treat one side of the distribution at a time. If you're statistically savvy you may want to look up Markov's Inequality. Chebyshev is derived from Markov.

References

Freund and Simon, *Modern Elementary Statistics*, 9th Ed., Prentice-Hall, 1997.
Johnson, *Miller and Freund's Probability and Statistics for Engineers*, 5th Ed., Prentice-Hall, 1994.
Freund and Walpole, *Mathematical Statistics*, 3rd Ed., Prentice-Hall, 1980.