> DO NOT PUT YOUR FAITH IN WHAT STATISTICS SAY UNTIL YOU HAVE CAREFULLY CONSIDERED WHAT THEY DO NOT SAY.
>
> WILLIAM W. WATT

Measure - Statistics

Measure - Statistics is described in the following topic areas:

- Basic statistics
- Probability
- Process capability

Basic Statistics

Basic Statistics is presented in the following topic areas:

- Basic terms
- Central limit theorem
- Descriptive statistics
- Graphical methods
- Statistical conclusions

Basic Terms

Continuous Distribution:	A distribution containing infinite (variable) data points that may be displayed on a continuous measurement scale. Examples: normal, uniform, exponential, and Weibull distributions.
Discrete Distribution:	A distribution resulting from countable (attribute) data that has a finite number of possible values. Examples: binomial, Poisson, and hypergeometric distributions.
Parameter:	The true numeric population value, often unknown, estimated by a statistic.
Population:	All possible observations of similar items from which a sample is drawn.
Statistic:	A numerical data value taken from a sample that may be used to make an inference about a population.

(Omdahl, 2009)[12]

Other terms are defined in the following content.

Central Limit Theorem

If a random variable, x, has mean μ, and finite variance σ_x^2, as n increases, $\bar{x}$ approaches a normal distribution with mean μ and variance $\sigma_{\bar{x}}^2$. Where, $\sigma_{\bar{X}}^2 = \frac{\sigma_x^2}{n}$ and n is the number of observations on which each mean is based.

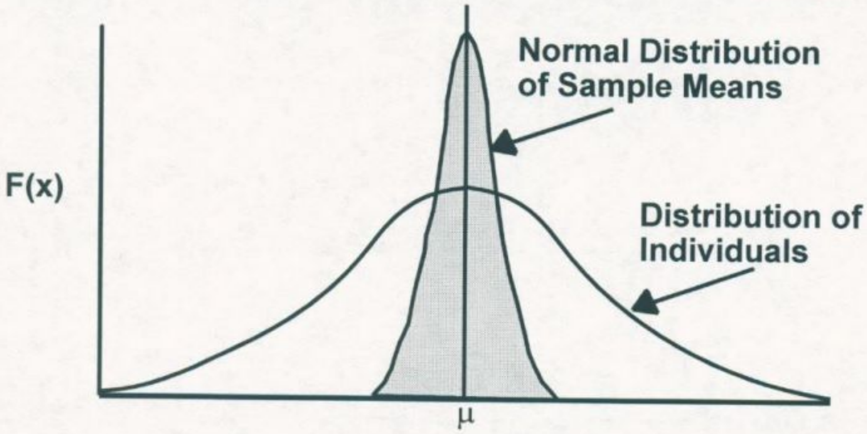


Figure 7.1 Distributions of Individuals Versus Means

The Central Limit Theorem States:

- **The sample means $\bar{X}_i$ will be more normally distributed around μ than individual readings X_j. The distribution of sample means approaches normal regardless of the shape of the parent population. This is why $\bar{X}$ - R control charts work!**

- **The spread in sample means $\bar{X}_i$ is less than X_j with the standard deviation of $\bar{X}_i$ equal to the standard deviation of the population (individuals) divided by the square root of the sample size. $s_{\bar{x}}$ is referred to as the standard error of the mean:**

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \quad \text{Which is estimated by } s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

Example 7.1: Assume the following are weight variation results: $\bar{X}$ = 20 grams and σ = 0.124 grams. Estimate $\sigma_{\bar{X}}$ for a sample size of 4:

Solution:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{0.124}{\sqrt{4}} = 0.062 \text{ grams}$$

Central Limit Theorem (Continued)

The significance of the central limit theorem on control charts is that the distribution of sample means approaches a normal distribution. Refer to Figure 7.2 below:

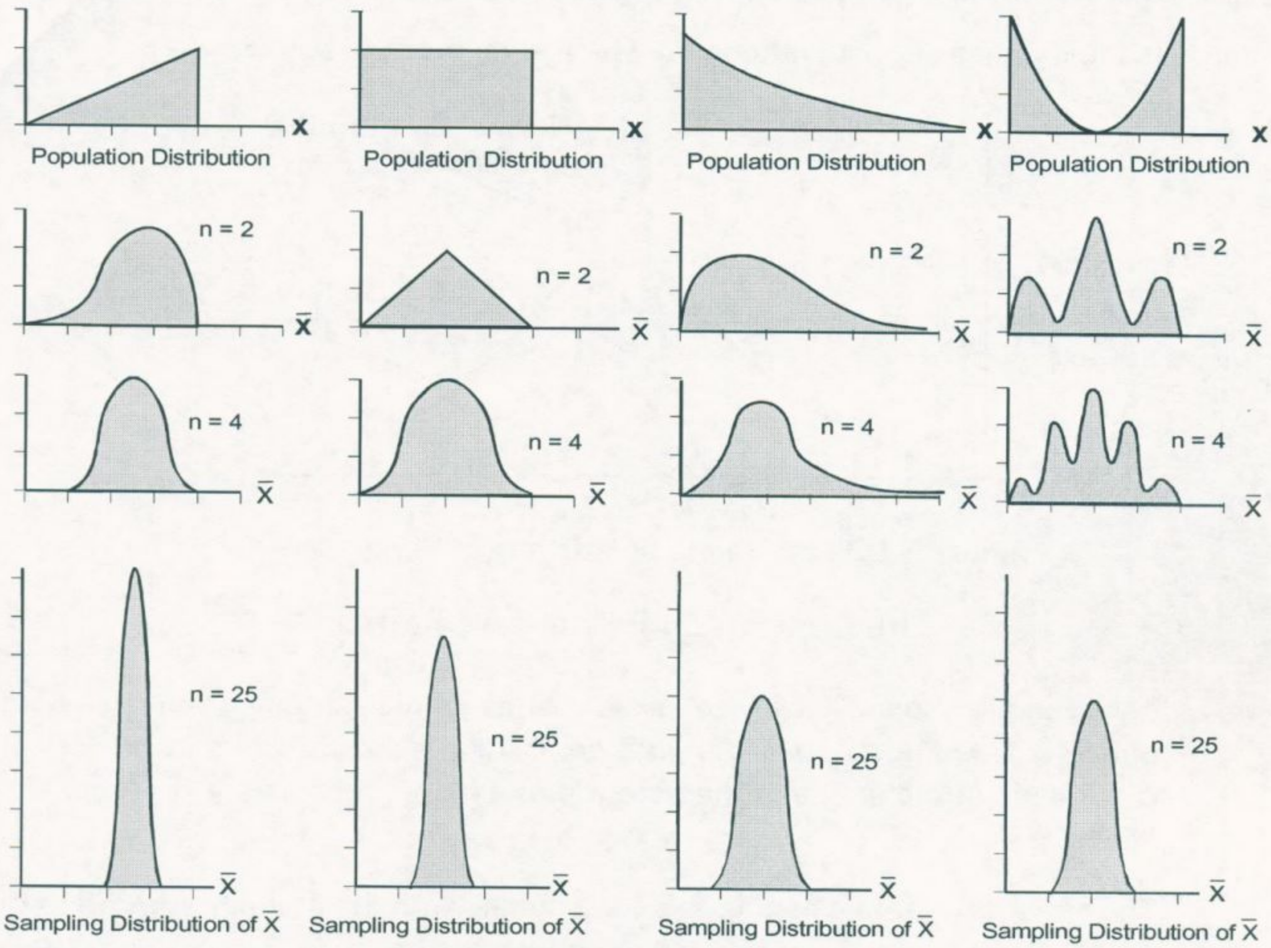


Figure 7.2 Illustration of Central Tendency

(Lapin, 1982)[9]

In Figure 7.2, a variety of population distributions approach normality for the sampling distribution of $\bar{X}$ as n increases. For most distributions, but not all, a near normal sampling distribution is attained with a sample size of 4 or 5.

Descriptive Statistics

Descriptive statistics include measures of central tendency, measures of dispersion, probability density function, frequency distributions, and cumulative distribution functions.

Measures of Central Tendency

Measures of central tendency represent different ways of characterizing the central value of a collection of data. Three of these measures will be addressed here: mean, mode, and median.

The Mean (X-bar, $\bar{X}$)

The mean is the total of all data values divided by the number of data points.

$$\bar{X} = \frac{\sum X}{n}$$

Where:

$\bar{X}$ is the mean
$\sum$ means summation
X represents each number
n is the sample size

Example 7.2: For the following 9 numbers, find $\bar{X}$.

5 3 7 9 8 5 4 5 8

Answer: 6

The arithmetic mean is the most widely used measure of central tendency.

Advantages of using the mean:

- It is the center of gravity of the data
- It uses all data
- No sorting is needed

Disadvantages of using the mean:

- Extreme data values may distort the picture
- It can be time-consuming
- The mean may not be the actual value of any data points

Measures of Central Tendency (Continued)

The Mode

The mode is the most frequently occurring number in a data set.

Example 7.3: (9 Numbers). Find the mode of the following data set:

5 3 7 9 8 5 4 5 8

Answer: 5

Note: It is possible for groups of data to have more than one mode.

Advantages of using the mode:

- No calculations or sorting are necessary
- It is not influenced by extreme values
- It is an actual value
- It can be detected visually in distribution plots

Disadvantage of using the mode:

- The data may not have a mode, or may have more than one mode

The Median (Midpoint)

The median is the middle value when the data is arranged in ascending or descending order. For an even set of data, the median is the average of the middle two values.

Examples 7.4: Find the median of the following data set:

(10 Numbers) 2 2 2 3 4 6 7 7 8 9

(9 Numbers) 2 2 3 4 5 7 8 8 9

Answer: 5 for both examples

Measures of Central Tendency (Continued)

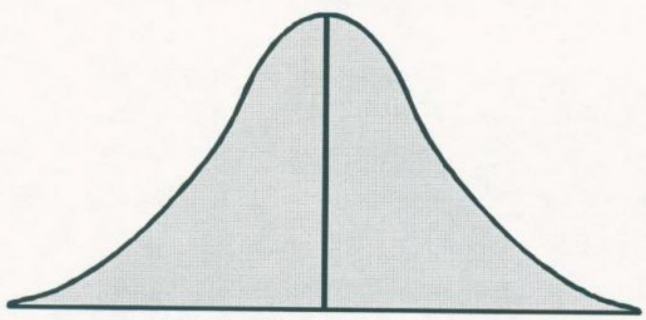
Advantages of using the median:

- Provides an idea of where most data is located
- Little calculation required
- Insensitivity to extreme values

Disadvantages of using the median:

- The data must be sorted and arranged
- Extreme values may be important
- Two medians cannot be averaged to obtain a combined distribution median
- The median will have more variation (between samples) than the average

For a Normal Distribution



MEAN = MEDIAN = MODE

For a Skewed Distribution

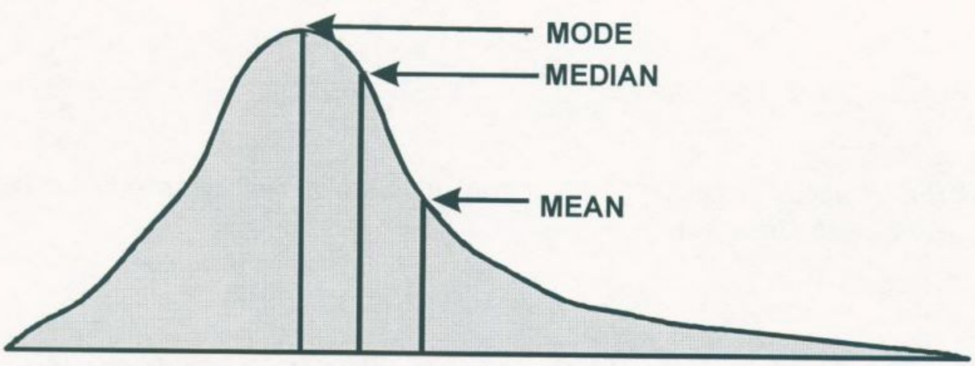


Figure 7.3 A Comparison of Central Tendency in Normal and Skewed Distributions

Measures of Dispersion

Other than central tendency, the other important parameter to describe a set of data is spread or dispersion. Three main measures of dispersion will be reviewed: range, variance, and standard deviation.

Range (R)

The range of a set of data is the difference between the largest and smallest values.

Example 7.5: (9 Numbers). Find the range of the following data set:
5 3 7 9 8 5 4 5 8
Answer: 9 - 3 = 6

Variance (σ^2, s^2)

The variance, σ^2 or s^2, is equal to the sum of the squared deviations from the mean, divided by the sample size. The formula for variance is:

$$\text{Population, } \sigma^2 = \frac{\sum (X - \mu)^2}{N} \qquad \text{Sample, } s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

The variance is equal to the standard deviation squared.

Standard Deviation (σ, s)

The standard deviation is the square root of the variance.

$$\text{Population, } \sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} \qquad \text{Sample, } s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

N is used for a population and n - 1 for a sample to remove potential bias in relatively small samples (less than 30).

Coefficient of Variation (COV)

The coefficient of variation equals the standard deviation divided by the mean and is expressed as a percentage.

$$\text{COV} = \frac{s}{\bar{X}}(100\%) \text{ or COV} = \frac{\sigma}{\mu}(100\%)$$

The Classic Method of Calculating Standard Deviation

Calculate the standard deviation of the following data set using the formula:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Example 7.6: Determine s from the following data:

SAMPLE	X	$\bar{X}$	$(X - \bar{X})$	$(X - \bar{X})^2$
1	162	146	+16	256
2	176	146	+30	900
3	160	146	+14	196
4	142	146	- 4	16
5	125	146	-21	441
6	159	146	+13	169
7	145	146	- 1	1
8	167	146	+21	441
9	114	146	-32	1024
10	120	146	-26	676
11	119	146	-27	729
12	180	146	+34	1156
13	154	146	+ 8	64
14	125	146	-21	441
15	142	146	- 4	16
	$\sum X = 2190$			

$\Sigma (X - \bar{X})^2 = 6526$

Calculate the average: $\bar{x} = \frac{\Sigma x}{n} = \frac{2190}{15} = 146$

- **Compute the deviation** $(x - \bar{x})$
- **Square each deviation** $(x - \bar{x})^2$
- **Sum the squares of the deviations** $\Sigma (x - \bar{x})^2$
- **Calculate standard deviation:**

$$S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{6526}{14}} = \sqrt{465} = 21.6$$

Summary:
$\bar{X} = 146$
$n = 15$
$s = 21.6$
$R = 66$

s is the standard deviation of the sample (21.6) which is used as an estimate for the population from which the sample was taken.

Shortcut Formula for Standard Deviation

$$S = \sqrt{\frac{n\left(\Sigma x^2\right) - \left(\Sigma x\right)^2}{n(n-1)}}$$

This formula will yield the same results as shown on the previous page. It is called a "shortcut" because it is convenient to use with some computers and calculators when working with messy data.

Determine $\overline{X}$ and s Using a Calculator

Formerly the authors attempted to instruct students on how to determine $\overline{X}$ and standard deviation on a Sharp calculator. However, many varieties of Texas Instrument, Casio, Hewlett Packard, and Sharp calculators can accomplish this task. The functions on all of these calculators are subject to change. It should be recognized that most technical people determine the mean and dispersion using a calculator. The following general procedures apply:

1. Turn on the calculator. Put it in statistical mode.
2. Enter all observation values following the model instructions.
3. Determine the sample mean ($\overline{X}$).
4. Determine the population standard deviation, σ, or the sample standard deviation, s.

Alternative Methods to Determine Standard Deviation

Standard deviation can be determined using probability paper. However, since the advent of computer programs, this is rarely done. Standard deviation can also be estimated from control charts using $\overline{R}$. This technique is discussed later in this Section and relates to the determination of process capability.

The control chart method of estimating standard deviation makes the big assumption that the process being charted is in control and many processes are not. Using a calculator or software program to determine s from individual data is often more accurate.

Probability Density Function

The probability density function, f(x), describes the behavior of a random variable. Typically, the probability density function is viewed as the "shape" of the distribution. It is normally a grouped frequency distribution. Consider the histogram for the length of a product shown in Figure 7.4.

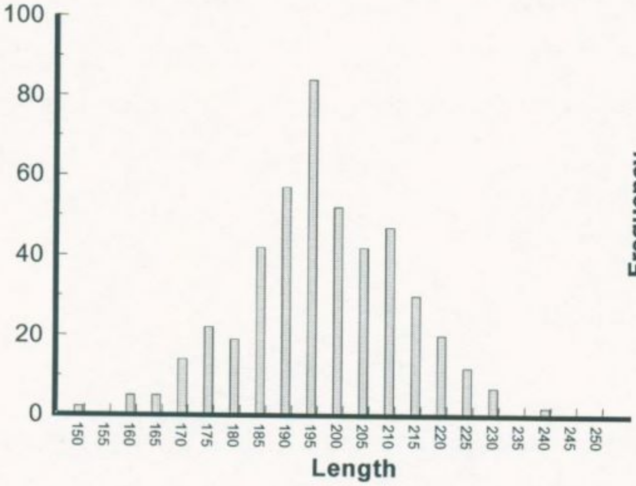


Figure 7.4 Example Histogram

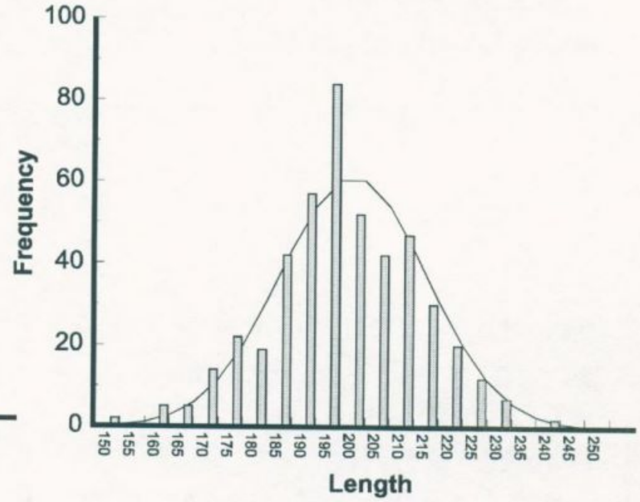


Figure 7.5 Histogram with Overlaid Model

A histogram is an approximation of the distribution's shape. The histogram shown in Figure 7.4 appears symmetrical. Figure 7.5 shows this histogram with a smooth curve overlaying the data. The smooth curve is the statistical model that describes the population; in this case, the normal distribution.

When using statistics, the smooth curve represents the population. The differences between the sample data represented by the histogram and the population data represented by the smooth curve are assumed to be due to sampling error. In reality, the difference could also be caused by lack of randomness in the sample or an incorrect model.

Probability Density Function (Continued)

The probability density function is similar to the overlaid model in Figure 7.5. The area below the probability density function to the left of a given value, x, is equal to the probability of the random variable represented on the x-axis being less than the given value x. Since the probability density function represents the entire sample space, the area under the probability density function must equal one. Since negative probabilities are impossible, the probability density function, f(x), must be positive for all values of x.

Stating these two requirements mathematically for continuous distributions with $f(x) \geq 0$:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

For discrete distributions for all values of n with $f(x) \geq 0$:

$$\sum_{0}^{n} f(x) = 1$$

Figure 7.6 demonstrates how the probability density function is used to compute probabilities. The area of the shaded region represents the probability of a single product drawn randomly from the population having a length less than 185. This probability is 15.9% and can be determined by using the standard normal table (z table) discussed later in this Section.

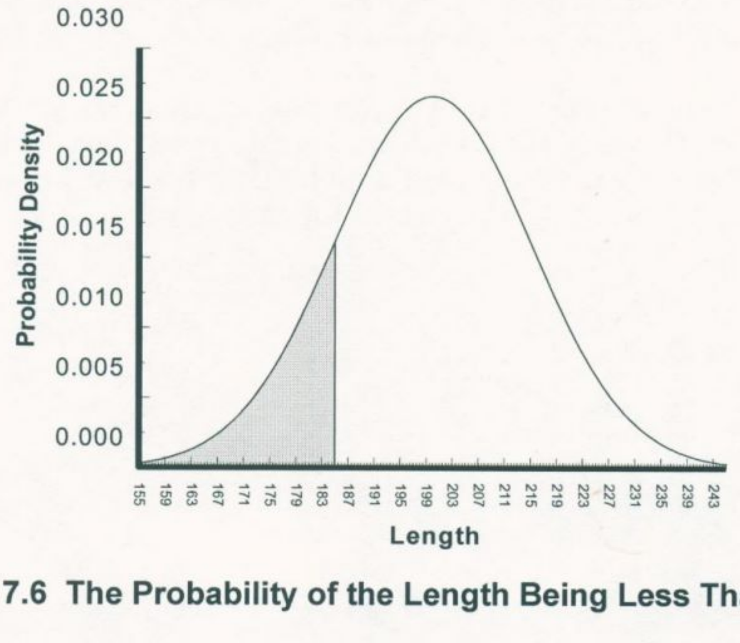


Figure 7.6 The Probability of the Length Being Less Than 185

Cumulative Distribution Function

The cumulative distribution function, F(x), denotes the area beneath the probability density function to the left of x. This is demonstrated in Figure 7.7.

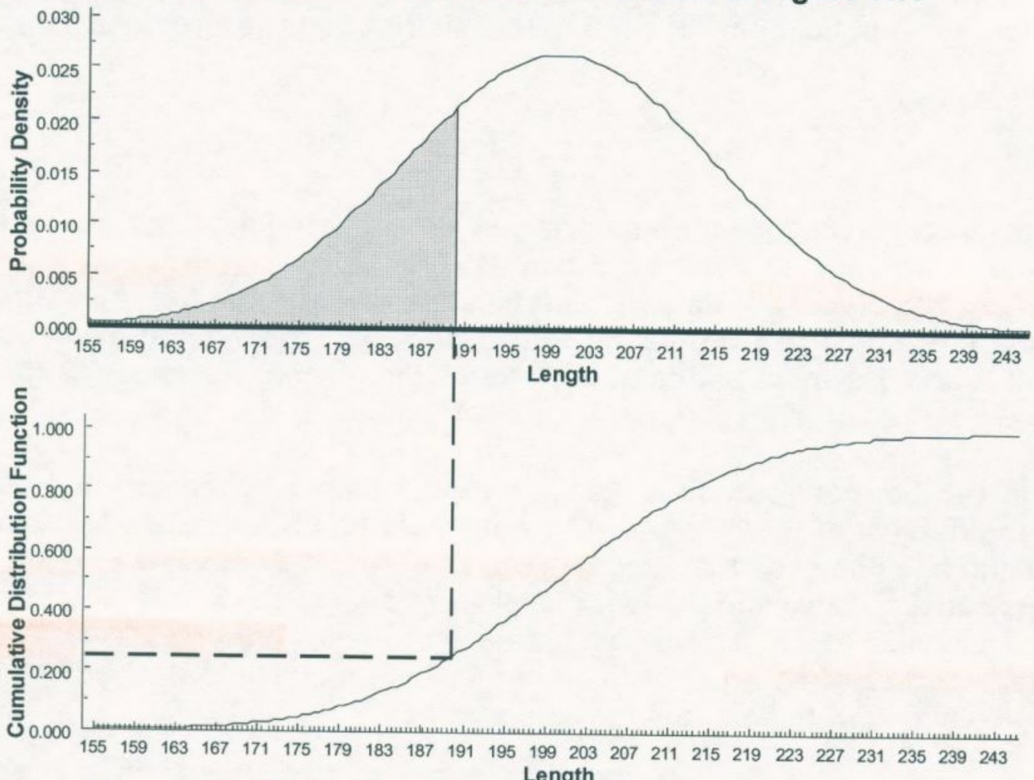


Figure 7.7 Cumulative Distribution Function for Length

The area of the shaded region of the probability density function in Figure 7.7 is 0.2525 which corresponds to the cumulative distribution function at x = 190.

Mathematically, the cumulative distribution function is equal to the integral of the probability density function to the left of x.

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Example 7.7: A random variable has the probability density function f(x) = 0.125x, where x is valid from 0 to 4. The probability of x being less than or equal to 2 is:

Solution: $F(2) = \int_{0}^{2} 0.125x\, dx = \frac{0.125x^2}{2} \Big|_{0}^{2} = 0.25$

Graphical Methods

Graphical methods include boxplots, stem and leaf plots, scatter diagrams, run charts, histograms, and normal probability plots. Additional information on properties and applications of probability distributions is provided later in this Section.

Boxplots

One of the simplest and most useful ways of summarizing data is the boxplot. This technique is credited to John W. Tukey (1977)[16]. The boxplot is a five number summary of the data. The data median is a line dividing the box. The upper and lower quartiles of the data define the ends of the box. The minimum and maximum data points are drawn as points at the end of lines (whiskers) extending from the box. A simple boxplot is shown in Figure 7.8 below.

Boxplots can be more complex. See Figure 7.9 below. They can be notched to indicate variability of the median. The notch widths are calculated so that if two median notches do not overlap, the medians are different at a 5% significance level. Boxplots can also have variable widths, proportional to the log of the sample size. Outliers can also be identified as points (asterisks) more than 1.5 times the interquartile distance from each quartile. Some computer programs can automatically generate boxplots for data analysis.

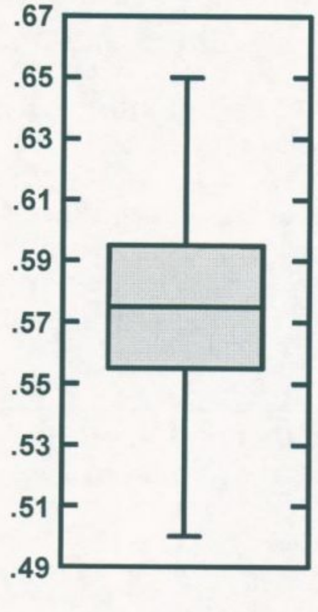


Figure 7.8 A Simple Boxplot

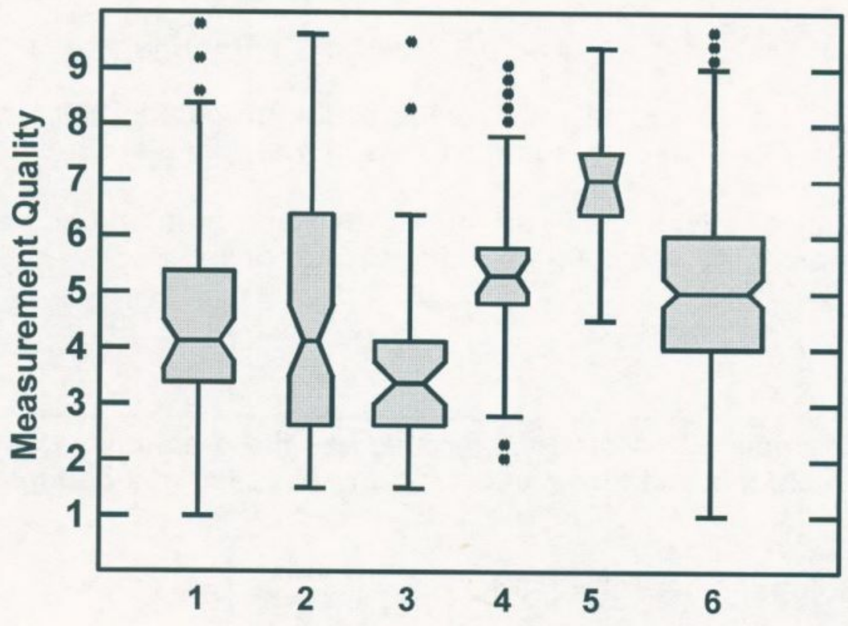


Figure 7.9 Complex Boxplots

Stem and Leaf Plots

The stem and leaf diagram (Tukey, 1977)[16] is a convenient, manual method for plotting data sets. These diagrams are effective in displaying both variable and categorical data sets. The diagram consists of grouping the data by class intervals, as stems, and the smaller data increments as leaves. Stem and leaf plots permit data to be read directly, whereas, histograms lose the individual data values as frequencies within class intervals.

Example 7.8: Shear Strength, 50 observations:

51.4	49.9	46.5	47.5	42.5	40.8	46.8	47.2	49.1	43.6
43.1	48.0	46.8	43.1	44.6	52.2	49.6	48.8	47.3	45.8
45.6	48.9	47.7	46.4	45.2	46.4	46.5	50.6	42.6	47.5
44.9	50.4	43.2	44.4	40.2	43.8	44.1	48.9	42.4	45.4
49.6	48.8	44.2	51.0	47.4	47.0	47.0	48.6	49.8	50.3

Show the above data in a histogram format.

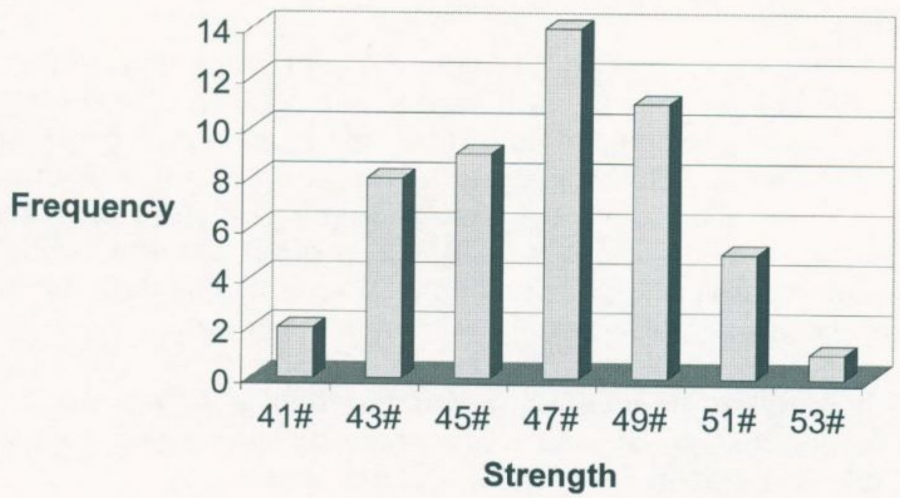


Figure 7.10 Shear Strength Histogram

Stem and Leaf Plots (Continued)

Example 7.8 (continued): Show the same data in a stem and leaf diagram:

```
                 5
                 2
                538
             69 8709
             1688591
            514644966
          8 6212408644
Leaf      2 48245068302
          -------------
Stem      0123456789012
          4444444444555
```

Figure 7.11 Shear Strength Stem and Leaf Plot

Scatter Diagrams

A scatter diagram is a graphic display of many XY coordinate data points which represent the relationship between two different variables. It is also referred to as a correlation chart. For example, temperature changes cause contraction or expansion of many materials. Both time and temperature in a kiln will affect the retained moisture in wood. Examples of such relationships on the job are abundant. Knowledge of the nature of these relationships can often provide a clue to the solution of a problem. Scatter diagrams can help determine if a relationship exists and how to control the effect of the relationship on the process.

In most cases, there is an independent variable and a dependent variable. Traditionally, the dependent variable is represented by the vertical axis and the independent variable is represented by the horizontal axis.

The ability to meet specifications in many processes is dependent upon controlling two interacting variables and, therefore, it is important to be able to control the effect one variable has on another. For instance, if the amount of heat applied to plastic liners affects their durability, then control limits must be set to consistently apply the right amount of heat. Through the use of scatter diagrams, the proper temperature can be determined ensuring a quality product.

Scatter Diagrams (Continued)

The dependent variable can be controlled if the relationship is understood. Correlation originates from the following:

- A cause-effect relationship
- A relationship between one cause and another cause
- A relationship between one cause and two or more other causes

Not all scatter diagrams reveal a linear relationship. The examples below definitely portray a relationship between the two variables, even though they do not necessarily produce a straight line. If a center line can be fitted to a scatter diagram, it will be possible to interpret it. To use scatter diagrams, one must be able to decide what factors will best control the process within the specifications.

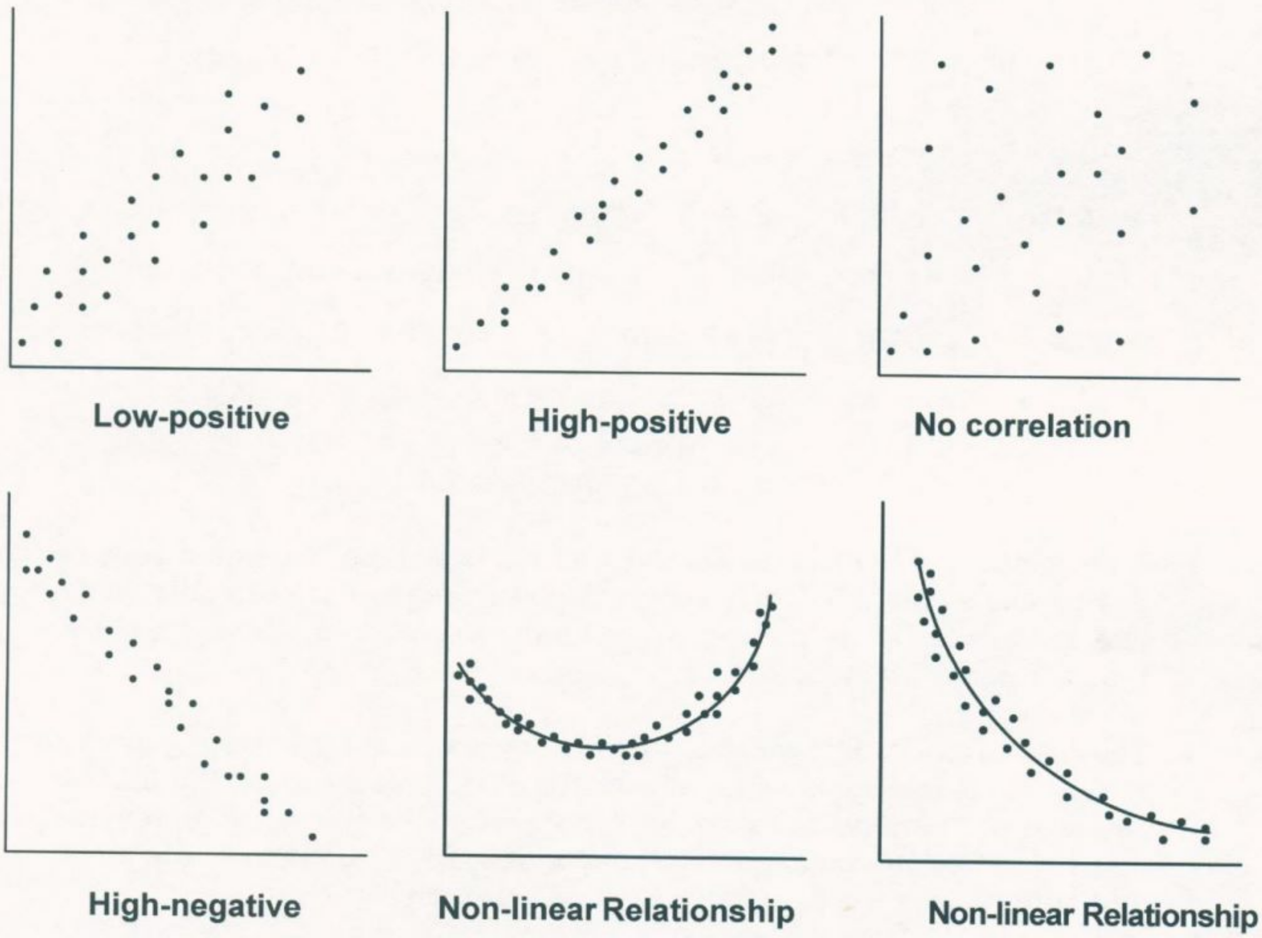


Figure 7.12 Scatter Diagram Examples

Scatter Diagrams (Continued)

Sample Correlation Coefficient

A sample correlation coefficient "r" can be calculated to determine the degree of association between two variables.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]}}$$

Interpret the Relationship Between Two Variables

r = –1.0	strong negative	when X increases, Y decreases
r = –0.5	slight negative	when X increases, Y generally decreases
r = 0	no correlation	the two variables are independent
r = +0.5	slight positive	when X increases, Y generally increases
r = +1.0	strong positive	when X increases, Y increases

Concluding Comments

- A correlation analysis seeks to uncover relationships. Common sense must be liberally applied. There is such a thing as a nonsense correlation whereby two variables that are not related can show correlation. For example, every time the car is washed, it rains.
- The line of "best fit" can be obtained by calculating a "regression line." However, to determine whether a relationship exists or not, the line can be "eyeballed." Simply draw a straight line through the points attempting to have approximately one-half above and one-half below. Study the points for trends and confirm that the line drawn fits appropriately.
- Scatter diagrams should always be analyzed prior to making decisions in correlation statistics.

Run (Trend) Charts

The average human brain is not good at comparing more than a few numbers at a time. Therefore, a large amount of data is often difficult to analyze unless it is presented in some easily digested format.

Data can be presented in either summary (static) or time sequence (dynamic) fashion. Important elements of most processes can change over time. These changes can be presented graphically by use of control charts or by the use of run or trend charts. For many business activities, trend charts will show patterns that indicate if a process is running normally or whether desirable or undesirable changes are occurring.

It should be noted that normal convention has time increasing across the page (from left to right) and the measurement value increasing up the page. Dependent upon the process measurement, values can be "good" if they go up or down the page or remain as close as possible to some target value. Consider the following examples:

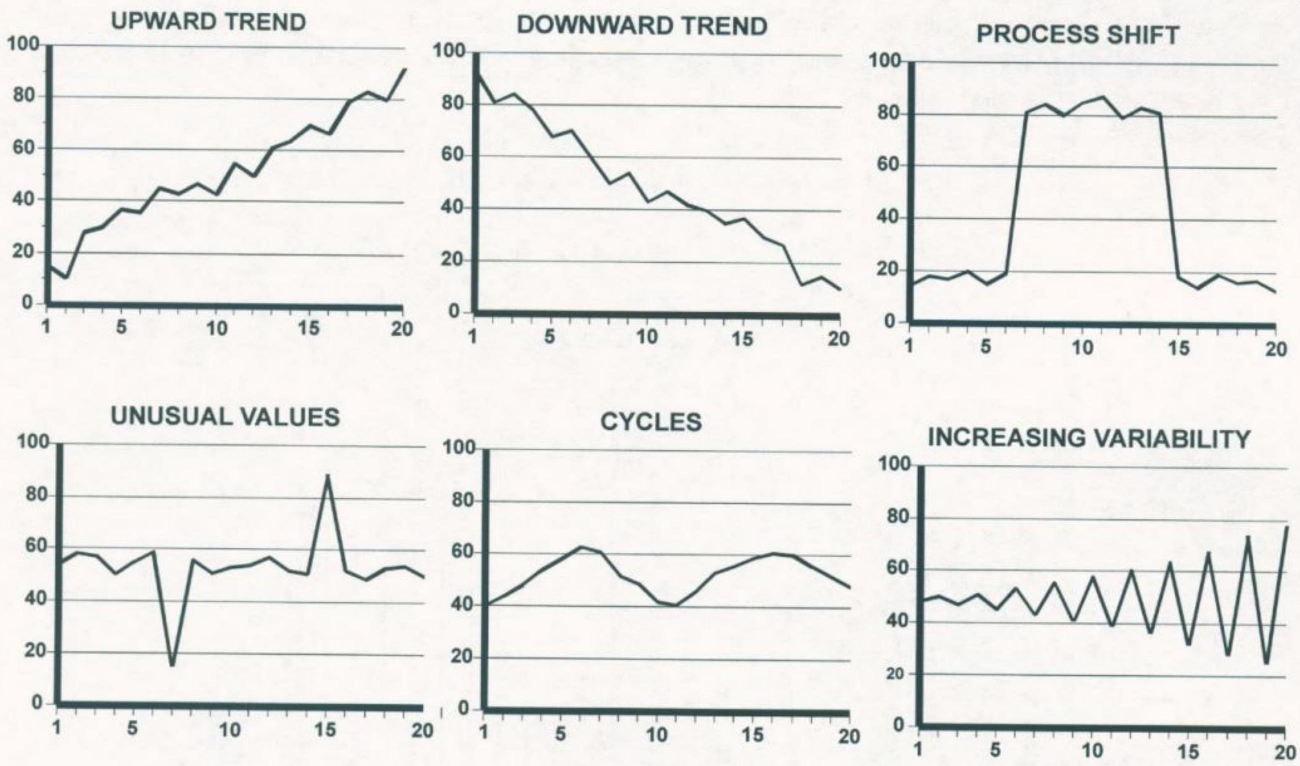


Figure 7.13 Examples of Trends

Histograms

Histograms have the following characteristics:

- Frequency column graphs that display a static picture of process behavior. Histograms require a minimum of 50-100 data points.

- A histogram is characterized by the number of data points that fall within a given bar or interval. This is commonly referred to as "frequency."

- A stable process is frequently characterized by a histogram exhibiting unimodal or bell-shaped curves. A stable process is predictable.

- An unstable process is often characterized by a histogram that does <u>not</u> exhibit a bell-shaped curve. Obviously other more exotic distribution shapes (like exponential, lognormal, gamma, beta, Weibull, Poisson, binomial, hypergeometric, geometric, etc.) exist as stable processes.

- When the bell curve is the approximate distribution shape, variation around the bell curve is chance or natural variation. Other variation is due to special or assignable causes.

Histogram Construction

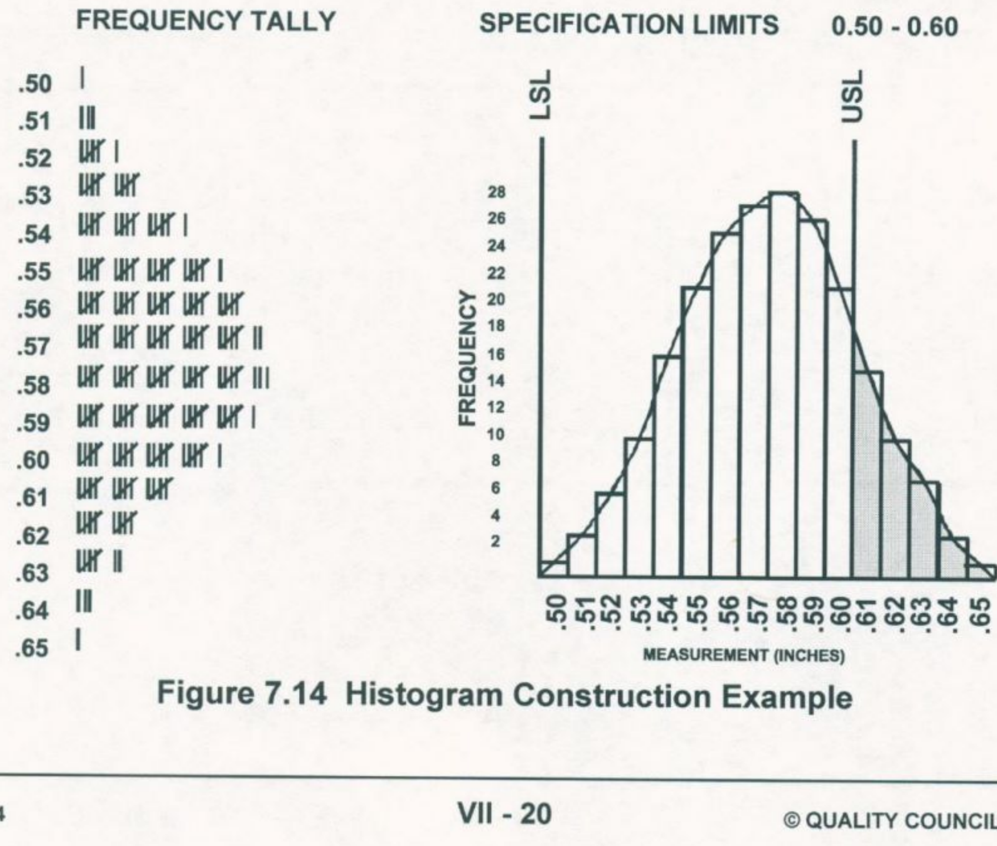


Figure 7.14 Histogram Construction Example

Histogram Examples

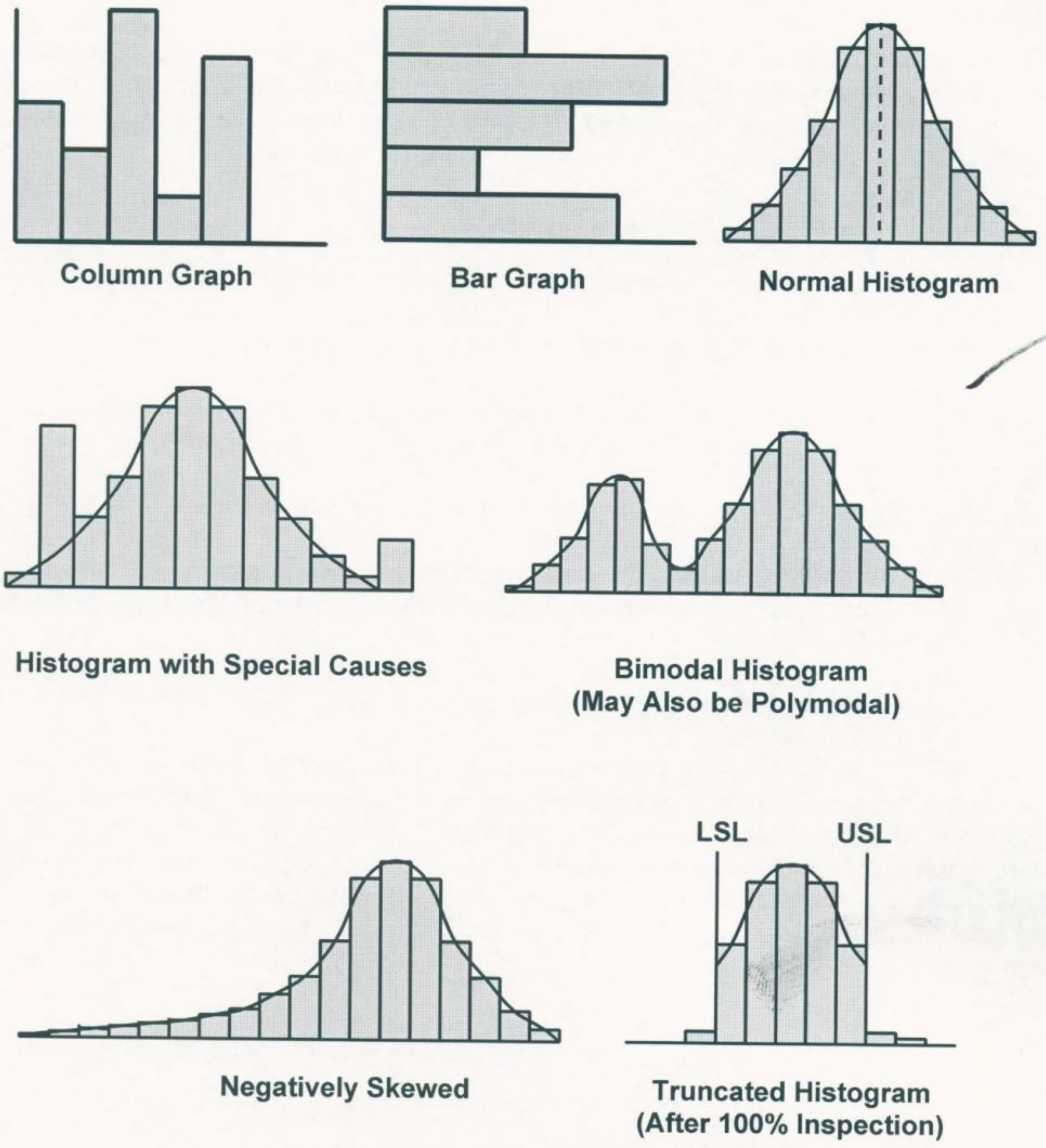


Figure 7.15 Histogram Examples

Drawing Valid Statistical Conclusions *

Analytical (Inferential) Studies

The objective of statistical inference is to draw conclusions about population characteristics based on the information contained in a sample. Statistical inference in a practical situation contains two elements: (1) the inference and (2) a measure of its validity. The steps involved in statistical inference are:

- Define the problem objective precisely
- Decide if the problem will be evaluated by a one-tail or two-tail test
- Formulate a null hypothesis and an alternate hypothesis
- Select a test distribution and a critical value of the test statistic reflecting the degree of uncertainty that can be tolerated (the alpha, α, risk)
- Calculate a test statistic value from the sample information
- Make an inference about the population by comparing the calculated value to the critical value. This step determines if the null hypothesis is to be rejected. If the null is rejected, the alternate must be accepted.
- Communicate the findings to interested parties

Everyday, in our personal and professional lives, individuals are faced with decisions between choice A or choice B. In most situations, relevant information is available; but it may be presented in a form that is difficult to digest. Quite often, the data seems inconsistent or contradictory. In these situations, an intuitive decision may be little more than an outright guess. While most people feel their intuitive powers are quite good, the fact is that decisions made on gut-feeling are often wrong.

* A substantial portion of the material throughout this Section is from the *CQE Primer* (2012)[17]. The student should note that portions of this subject are covered in more depth in Section VIII.

Drawing Valid Statistical Conclusions (Continued)

Null Hypothesis and Alternate Hypothesis

The null hypothesis is the hypothesis to be tested. The null hypothesis directly stems from the problem statement and is denoted as H_0.

The alternate hypothesis must include all possibilities which are not included in the null hypothesis and is designated H_1.

Examples of null and alternate hypothesis:

Null hypothesis:	H_o: $Y_a = Y_b$	H_o: $A \leq B$
Alternate hypothesis:	H_1: $Y_a \neq Y_b$	H_1: $A > B$

<u>A null hypothesis can only be rejected, or fail to be rejected, it cannot be accepted because of a lack of evidence to reject it.</u>

Test Statistic

In order to test a null hypothesis, a test calculation must be made from sample information. This calculated value is called a test statistic and is compared to an appropriate critical value. A decision can then be made to reject or not reject the null hypothesis.

Types of Errors

When formulating a conclusion regarding a population based on observations from a small sample, two types of errors are possible:

- <u>Type I error:</u> This error results when the null hypothesis is rejected when it is, in fact, true.

- <u>Type II error:</u> This error results when the null hypothesis is not rejected when it should be rejected.

The degree of risk (α) is normally chosen by the concerned parties (α is normally taken as 5%) in arriving at the critical value of the test statistic.

Drawing Valid Statistical Conclusions (Continued)

Enumerative (Descriptive) Studies

Enumerative data is data that can be counted. For example: the classification of things, the classification of people into intervals of income, age, health. A census is an enumerative collection and study. Useful tools for tests of hypothesis conducted on enumerative data are the chi square, binomial, and Poisson distributions.

Deming, in 1975, defined a contrast between enumeration and analysis:

Enumerative study: A study in which action will be taken on the universe.

Analytical study: A study in which action will be taken on a process to improve performance in the future.

Numerical descriptive measures create a mental picture of a set of data. The measures calculated from a sample are called statistics. When these measures describe a population, they are called parameters.

Measures	Statistics	Parameters
Mean	$\bar{X}$	μ
Standard Deviation	s	σ

Table 7.16 Statistics and Parameters

Table 7.16 shows examples of statistics and parameters for the mean and standard deviation. These two important measures are called central tendency and dispersion.

Summary of Analytical and Enumerative Studies

Analytical studies start with the hypothesis statement made about population parameters. A sample statistic is then used to test the hypothesis and either reject or fail to reject the null hypothesis. At a stated level of confidence, one is then able to make inferences about the population.

Probability

Probability is presented in the following topic areas:

- **Basic concepts**
- **Commonly used distributions**
- **Other distributions**

Most quality theories use statistics to make inferences about a population based on information contained in samples. The mechanism one uses to make these inferences is probability.

Conditions for Probability

The probability of any event, E, lies between 0 and 1. The sum of the probabilities of all possible events in a sample space, S, = 1.

Simple Events

An event that cannot be decomposed is a simple event, E. The set of all sample points for an experiment is called the sample space, S.

If an experiment is repeated a large number of times, N, and the event, E, is observed n_E times, the probability of E is approximately:

$$PE \approx \frac{n_E}{N}$$

Example 7.9: The probability of observing 3 on the toss of a single die is:

$$PE_3 = \frac{1}{6}$$

Example 7.10: What is the probability of getting 1, 2, 3, 4, 5, or 6 by throwing a die?

$$PE_T = P(E_1) + P(E_2) + P(E_3) + P(E_4) + P(E_5) + P(E_6)$$

$$PE_T = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

Compound Events

Compound events are formed by a composition of two or more events. They consist of more than one point in the sample space. For example, if two dice are tossed, what is the probability of getting an 8? A die and a coin are tossed. What is the probability of getting a 4 and tail? The two most important probability theorems are additive and multiplicative (covered later in this Section). For the following discussion, $E_A = A$ and $E_B = B$.

I. <u>Composition.</u> Consists of two possibilities -- a union or intersection.

A. <u>Union of A and B</u>

If A and B are two events in a sample space, S, the union of A and B ($A \cup B$) contains all sample points in event A, B, or both.

Example 7.11: In the die toss in Example 7.10, consider the following:

If $A = E_1$, E_2 and E_3 (numbers less than 4)
and $B = E_1$, E_3 and E_5 (odd numbers),
then $A \cup B = E_1$, E_2, E_3 and E_5.

B. <u>Intersection of A and B</u>

If A and B are two events in a sample space, S, the intersection of A and B ($A \cap B$) is composed of all sample points that are in both A and B.

Example 7.12: Refer to Example 7.11. $A \cap B = E_1$ and E_3

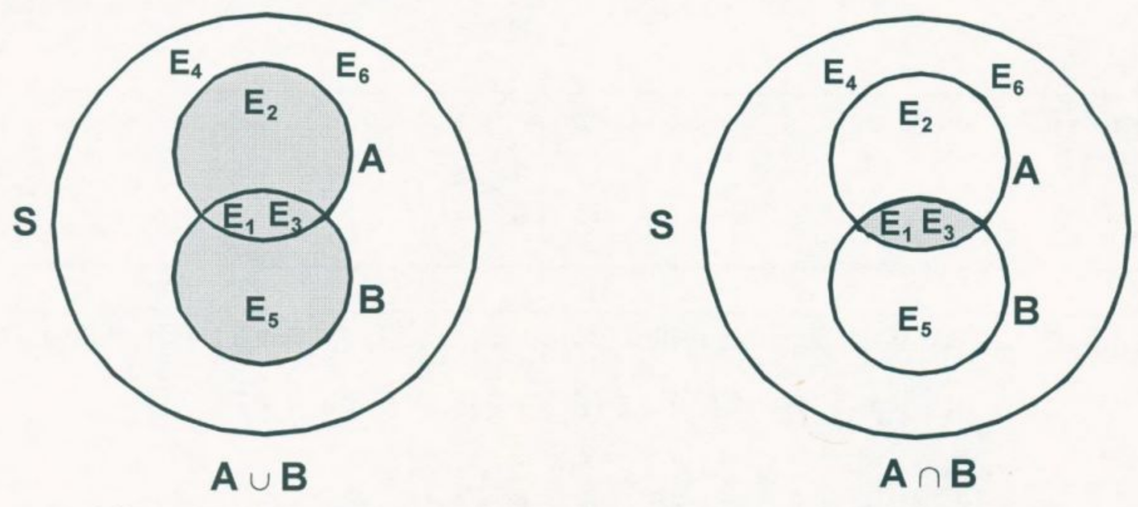


Figure 7.17 Venn Diagrams Illustrating Union and Intersection

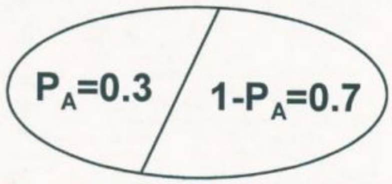
Compound Events (Continued)

II. Event Relationships. There are three relationships involved in finding the probability of an event: complementary, conditional, and mutually exclusive.

A. Complement of an Event

The complement of event A is all sample points in the sample space, S, but not in A. The complement of A is $1\text{-}P_A$.

Example 7.13: If P_A (cloudy days) is 0.3, the complement of A would be $1 - P_A = 0.7$ (clear).



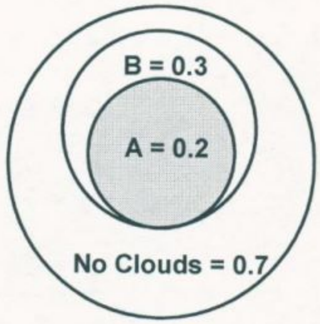
B. Conditional Probabilities

The conditional probability of event A occurring, given that event B has occurred is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{if } P(B) \neq 0$$

Example 7.14: If event A (rain) = 0.2 and event B (cloudiness) = 0.3, what is the probability of rain on a cloudy day? (Note, it will not rain without clouds.)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2}{0.3} = 0.67$$



Two events A and B are said to be independent if either:

$P(A|B) = P(A)$ or $P(B|A) = P(B)$ However,
$P(A|B) = 0.67$ and $P(A) = 0.2$= no equality, and
$P(B|A) = 1.00$ and $P(B) = 0.3$ = no equality

Therefore, the events are said to be dependent.

Compound Events (Continued)

C. Mutually Exclusive Events

If event A contains no sample points in common with event B, then they are said to be mutually exclusive.

Example 7.15: Obtaining a 3 and a 2 on the toss of a single die is a mutually exclusive event. The probability of observing both events simultaneously is zero. The probability of obtaining either a 3 or a 2 is:

$$PE_2 + PE_3 = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

D. Testing for Event Relationships

Example 7.16: Refer to Example 7.11.

Event A: E_1, E_2, E_3
Event B: E_1, E_3, E_5

Are events A and B mutually exclusive, complementary, independent, or dependent? Events A and B contain two sample points in common, so they are not mutually exclusive. They are not complementary because event B does not contain all points in S that are not in event A.

To determine if they are independent requires a check.

$$\text{Does } P(A|B) = P(A)?$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/6}{1/2} = \frac{2}{3} \qquad P(A) = \frac{1}{2}$$

$$\text{Therefore } P(A|B) \neq P(A)$$

By definition, events A and B are dependent.

The Additive Law

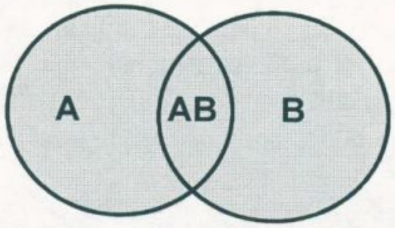
If the two events are not mutually exclusive:

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Note that $P(A \cup B)$ is shown in many texts as $P(A + B)$ and is read as the probability of A or B.

Example 7.17: If one owns two cars and the probability of each car starting on a cold morning is 0.7, what is the probability of getting to work?

$$P(A \cup B) = 0.7 + 0.7 - (0.7 \times 0.7)$$
$$= 1.4 - 0.49$$
$$= 0.91 = 91\%$$

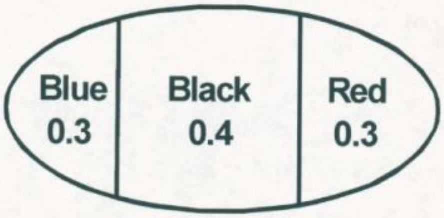


If the two events are mutually exclusive, the law reduces to:

2. $P(A \cup B) = P(A) + P(B)$ also $P(A + B) = P(A) + P(B)$

Example 7.18: If the probability of finding a black sock in a dark room is 0.4 and the probability of finding a blue sock is 0.3, what is the chance of finding a blue or black sock?

$$P(A \cup B) = 0.4 + 0.3 = 0.7 = 70\%$$



Note: The problem statements center around the word "or." Will car A or B start? Will one get a black or blue sock?

The Multiplicative Law

If events A and B are dependent, the probability of event A influences the probability of event B. This is known as conditional probability and the sample space is reduced.

For any two events, A and B, such that $P(B) \neq 0$:

1. $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ and $P(A \cap B) = P(A|B)P(B)$

Note in some texts $P(A \cap B)$ is shown as $P(A \cdot B)$ and is read as the probability of A and B. $P(B|A)$ is read as the probability of B given that A has occurred.

Example 7.19: If a shipment of 100 TV sets contains 30 defective units and two samples are obtained, what is probability of finding both defective? (Event A is the first sample and the sample space is reduced, and event B is the second sample.)

$$P(A \cap B) = \frac{30}{100} \times \frac{29}{99} = \frac{870}{9900} = 0.088$$
$$P(A \cap B) = 8.8\%$$

If events A and B are independent:

2. $P(A \cap B) = P(A) \text{ X } P(B)$

Example 7.20: One relay in an electric circuit has a probability of working equal to 0.9. Another relay in series has a chance of 0.8. What's the probability that the circuit will work?

$$P(A \cap B) = 0.9 \text{ X } 0.8 = 0.72$$
$$P(A \cap B) = 72\%$$

Note: The problem statements center around the word "and." Will TV A <u>and</u> B work? Will relay A <u>and</u> B operate?

Commonly Used Distributions

Commonly used distributions include the following:

- **Normal**
- **Binomial**
- **Poisson**
- **Chi square**
- **Student's t**
- **F distribution**

Normal Distribution

The normal distribution has numerous applications. It is useful when it is equally likely that readings will fall above or below the average.

When a sample of several random measurements are averaged, distribution of such repeated sample averages tends to be normally distributed regardless of the distribution of the measurements being averaged. Mathematically, if

$$\overline{X} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

the distribution of $\overline{X}$s becomes normal as n increases. If the set of samples being averaged have the same mean and variance, then the mean of the $\overline{X}$s is equal to the mean (μ) of the individual measurements, and the variance of the $\overline{X}$s is:

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n}$$

Where σ_x^2 is the variance of the individual variables being averaged.

The tendency of sums and averages of independent observations, from populations with finite variances, to become normally distributed as the number of variables being summed or averaged becomes large is known as the central limit theorem. For distributions with little skewness, summing or averaging as few as 3 or 4 variables will result in a normal distribution. For highly skewed distributions, more than 30 variables may have to be summed or averaged to obtain a normal distribution. The normal probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad , -\infty < x < \infty$$

Where μ is the mean and σ is the standard deviation.

Normal Distribution (Continued)

The normal probability density function is not skewed, as shown in Figure 7.18.

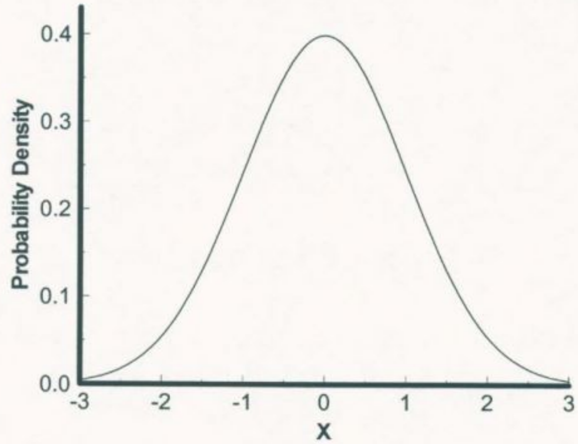


Figure 7.18 The Standard Normal Probability Density Function

The density function shown in Figure 7.18 is the standard normal probability density function. The standard normal probability density function has a mean of 0 and a standard deviation of 1. The normal probability density function cannot be integrated implicitly. Because of this, a transformation to the standard normal distribution is made, and the normal cumulative distribution function or reliability function is read from a table. The standard normal table is shown in the Appendix. If x is a normal random variable, it can be transformed to standard normal using the expression:

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

Other Z transformations include:

$$Z = \frac{X - \overline{X}}{\sigma} = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{X - \text{Std}}{\sigma} = \frac{\overline{X} - \text{Std}}{\sigma_{\overline{X}}}$$

Normal Distribution (Continued)

Example 7.21: A battery is produced with an average voltage of 60 and a standard deviation of 4 volts. If 9 batteries are selected at random, what is the probability that the total voltage of the 9 batteries is greater than 530? What is the probability that the average voltage of the 9 batteries is less than 62?

Solution Part A: The expected total voltage for nine batteries is 540. The expected standard deviation of the voltage of the total of nine batteries is:

$$S^2_{TOTAL} = 9 \times (4)^2 = 144 \qquad S_{TOTAL} = 12$$

Transforming to standard normal: $Z = \frac{530 - 540}{12} = -0.833$

From the standard normal table, the area to the right of z is 0.7976.

Solution Part B: The expected value is 60. The standard deviation is :

$$S_X = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{9}} = 1.333 \qquad \text{Thus, } z = \frac{62 - 60}{1.333} = 1.5$$

The area to the left of z is 1 - 0.0668 = 0.9332

The probability density function of the voltage of the individual batteries and of the average of nine batteries is shown in Figure 7.19. The distribution of the averages has less variance because the standard deviation of the averages is equal to the standard deviation of the individuals divided by the square root of the sample size.

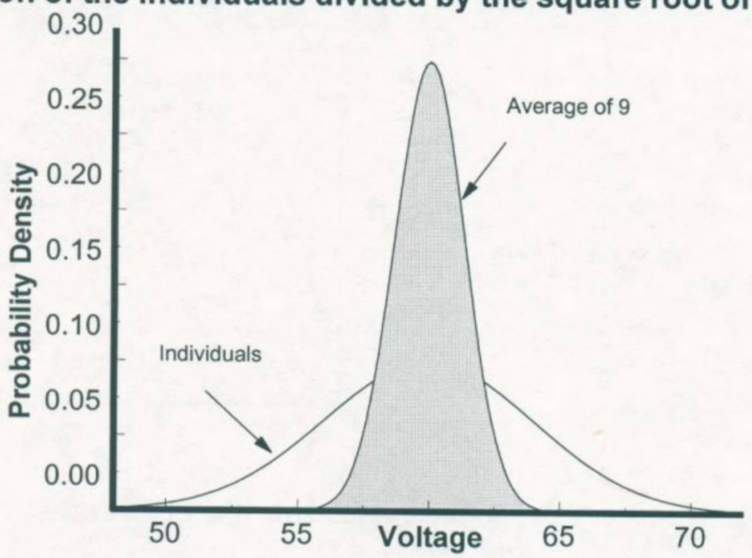


Figure 7.19 Distribution of Average Voltage

Binomial Distribution

The binomial distribution is one of several distributions used to model discrete data. Some situations call for discrete data, such as, the number of missiles required to destroy a target or the number of defectives in a lot of 1,000 items.

The binomial distribution applies when the population is large ($N > 50$) and the sample size is small compared to the population. The binomial is best applied when the sample size is less than 10% of N ($n < 0.1N$). Binomial sampling is with replacement. It is most appropriate to use when the proportion defective is equal to or greater than 0.1.

The binomial is an approximation to the hypergeometric. The normal distribution approximates the binomial when $np \geq 5$. The Poisson distribution can be used to approximate the binomial distribution when p is small (generally, less than 0.1) and n is large (generally, $n \geq 16$) by using np as the mean of the Poisson distribution.

$$P(r) = C_r^n \, p^r (1 - p)^{n-r} = \frac{n!}{r!(n - r)!} p^r (1 - p)^{n-r}$$

Where: n = sample size
r = occurrences or number of defectives
p = probability or proportion defective

There is a limited binomial probability table in the Appendix. The binomial distribution, using different p values, is shown in Figure 7.20.

$$P(r) = \frac{n!}{r!(n - r)!} p^r q^{n-r}$$

n = sample size
r = number of occurrences
p = probability
q = 1 - p

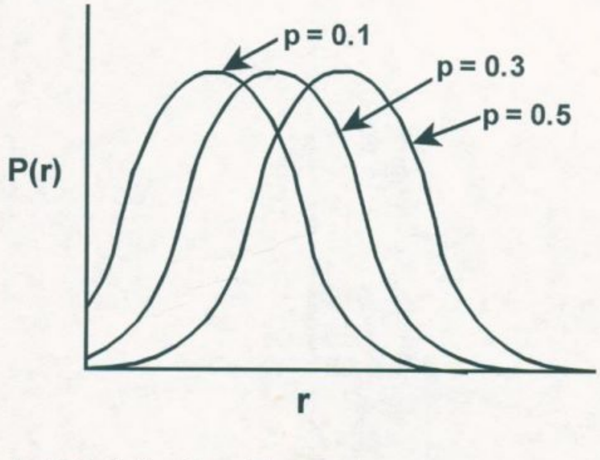


Figure 7.20 Binomial Distribution Example

Binomial Distribution (Continued)

The binomial distribution is used to model situations having only 2 possible outcomes, usually labeled as success or failure. For a random variable to follow a binomial distribution, the number of trials must be fixed, and the probability of success must be equal for all trials. The binomial probability density function is:

$$P(x,n,p) = \binom{n}{x} p^x (1-p)^{n-x} = C_x^n p^x (1-p)^{n-x}$$

Where P (x,n,p) is the probability of exactly x successes in n trials with a probability of success equal to p on each trial. Note that:

$$C_x^n = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

This notation is referred to as "n choose x," and is equal to the number of combinations of size x made from n possibilities. This function is found on most calculators.

The binomial distribution mean and standard deviation, sigma, can be obtained from the following calculations when the event of interest is the count of defined occurrences in the population, e.g., the number of defectives or effectives.

The binomial mean = $\mu = np$

The binomial sigma = $\sigma = \sqrt{np(1-p)}$

Poisson Distribution

The Poisson distribution is one of several distributions used to model discrete data and has numerous applications in industry. The Poisson distribution can be an approximation to the binomial when p is equal to or less than 0.1, and the sample size n is fairly large.

$$P(r) = \frac{\mu^r e^{-\mu}}{r!}$$

Where: $\mu = np$ = the population mean
r = number of defectives
e = 2.71828 the base of natural logarithms

Poisson Distribution (Continued)

The Poisson distribution using different p values is shown in Figure 7.21.

$$P(r) = \frac{(np)^r e^{-np}}{r!}$$

n = sample size
r = number of occurrences
p = probability
np = μ = average

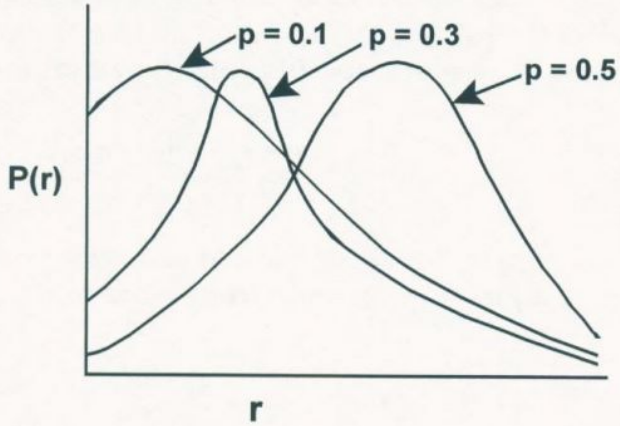


Figure 7.21 Poisson Distribution Example

The Poisson is used as a distribution for defect counts and can be used as an approximation to the binomial. For np<5, the binomial is better approximated by the Poisson than the normal distribution. When the normalized Poisson is used to model defects, the sample size should be large enough for the Poisson mean to have a value of at least 4 or 5. However, whatever the mean value, the cumulative Poisson distribution provides both the individual and cumulative terms.

The Poisson distribution is used to model rates, such as rabbits per acre, defects per unit, or arrivals per hour. The Poisson distribution is closely related to the exponential distribution. If x is a Poisson distributed random variable, then 1/x is an exponential random variable. If x is an exponential random variable, then 1/x is a Poisson random variable. For a random variable to be Poisson distributed, the probability of an occurrence in an interval must be proportional to the length of the interval, and the number of occurrences per interval must be independent.

The Poisson distribution average and standard deviation can be obtained from the following calculations:

$$\text{The Poisson average} = \mu = n\bar{p} = \bar{c}^*$$

$$\text{The Poisson standard deviation} = \sqrt{\mu} = \sqrt{n\bar{p}} = \sqrt{\bar{c}}^*$$

* From the attribute c chart.

Chi Square Distribution

The chi square, t, and F distributions are formed from combinations of random variables. Because of this, they are generally not used to model physical phenomena, like time to fail, but are used to make decisions and construct confidence intervals. These three distributions are considered sampling distributions. The student should be advised that there are numerous applications of these distributions in Section VIII.

The chi square distribution is formed by summing the squares of standard normal random variables. For example, if z is a standard normal random variable, then:

$$y = z_1^2 + z_2^2 + z_3^2 + \ldots + z_n^2$$

is a chi square random variable (statistic) with n degrees of freedom. A chi square statistic is also created by summing two or more chi square statistics and dividing by the sum of the degrees of freedom. A distribution having this property is regenerative. The chi square distribution is a special case of the gamma distribution with a failure rate of 2, and degrees of freedom equal to 2 divided by the number of degrees of freedom for the corresponding chi square distribution. The chi square probability density function is:

$$f(x) = \frac{x^{(\nu/2-1)}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}, x > 0$$

where ν is the degrees of freedom, and $\Gamma(x)$ is the gamma function. The chi square probability density function is shown in Figure 7.22.

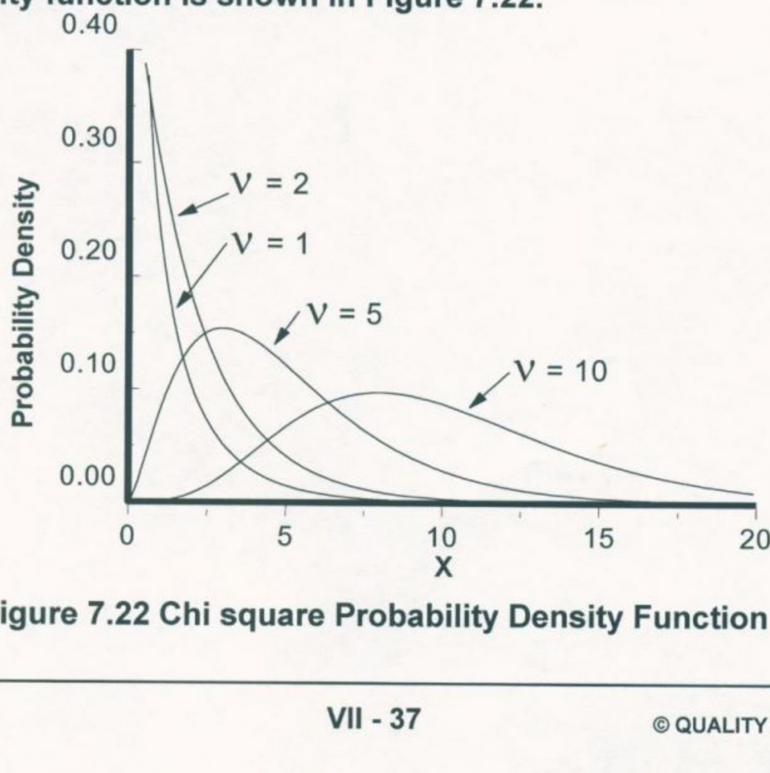


Figure 7.22 Chi square Probability Density Function

Chi Square Distribution (Continued)

The critical values of the chi square distribution are given in the Appendix.

Example 7.22: A chi square random variable has 7 degrees of freedom, what is the critical value if 5% of the area under the chi square probability density is desired in the right tail?

Solution : When hypothesis testing, this is commonly referred to as the critical value with 5% significance, or $\alpha = 0.05$. From the chi square table in the Appendix, this value is 14.067.

F Distribution

If X is a chi square random variable with ν_1 degrees of freedom, and Y is a chi square random variable with ν_2 degrees of freedom, and if X and Y are independent, then:

$$F = \frac{X / \nu_1}{Y / \nu_2}$$

is an F distribution with ν_1 and ν_2 degrees of freedom. The F distribution is used extensively to test for equality of variances from two normal populations.

The F probability density function is:

$$f(x) = \left(\frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \right) \left(\frac{x^{\nu_1/2-1}}{\left(1+\frac{\nu_1 x}{\nu_2}\right)^{(\nu_1+\nu_2)/2}} \right), x > 0$$

F Distribution (Continued)

The F probability density function is shown in Figure 7.23.

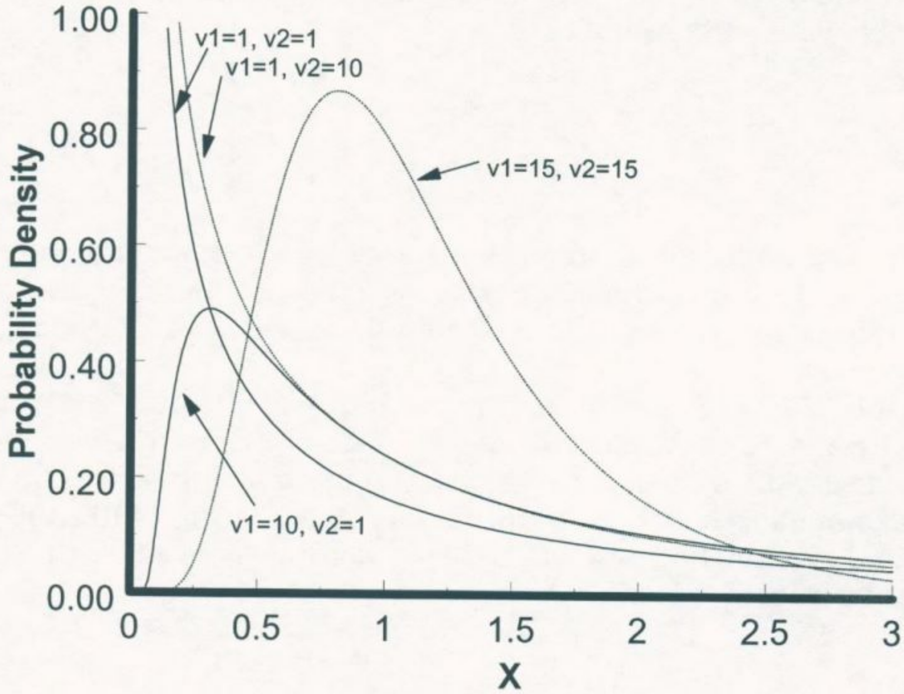


Figure 7.23 The F Probability Density Function

The F cumulative distribution function is given in Tables VII and VIII in the Appendix. Both the lower and upper tails are listed, but most texts only give one tail, and require the other tail to be computed using the expression:

$$F_{\alpha,\, n1,\, n2} = \frac{1}{F_{1-\alpha,\, n2,\, n1}}$$

Example 7.23: Given that $F_{0.05}$ with $\nu_1 = 8$ and $\nu_2 = 10$ is 3.07, find the value of $F_{0.95}$ with $\nu_1 = 10$ and $\nu_2 = 8$.

Answer:

$$F_{0.95,10,8} = \frac{1}{F_{0.05,8,10}} = \frac{1}{3.07} = 0.326$$

Student's t Distribution

The student's t distribution is formed by combining a standard normal random variable and a chi square random variable. If z is a standard normal random variable, and X^2 is a chi square random variable with ν degrees of freedom, then a random variable with a t distribution is:

$$t = \frac{z}{\sqrt{\frac{X^2}{\nu}}}$$

Like the normal distribution, when random variables are averaged, the distribution of the average tends to be normal, regardless of the distribution of the individuals. The t distribution is equivalent to the F distribution with 1 and ν degrees of freedom.

The t distribution is commonly used for hypothesis testing and constructing confidence intervals for means. It is used in place of the normal distribution when the standard deviation is unknown. The t distribution compensates for the error in the estimated standard deviation. If the sample size is large, n>100, the error in the estimated standard deviation is small, and the t distribution is approximately normal. The t probability density function is:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{\frac{-(\nu+1)}{2}} \qquad \text{for } -\infty < x < \infty$$

Where ν is the degrees of freedom. The t probability density function is shown in Figure 7.24.

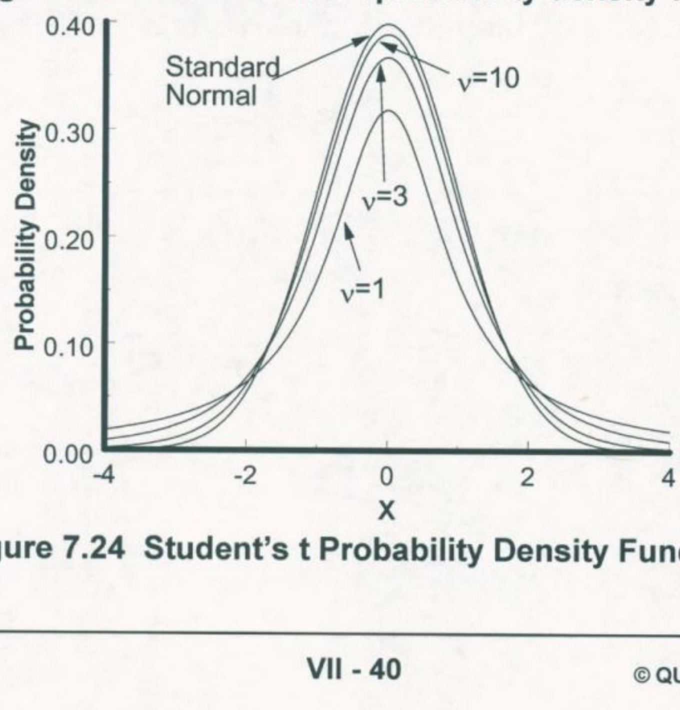


Figure 7.24 Student's t Probability Density Function

Student's t Distribution (Continued)

The mean and variance of the t distribution are:

$$\mu = 0 \qquad \sigma^2 = \frac{\nu}{\nu - 2},\ \nu \geq 3$$

From a random sample of n items, the probability that:

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

falls between any two specified values is equal to the area under the t probability density function between the corresponding values on the x-axis with n-1 degrees of freedom.

Example 7.24: The burst strength of 15 randomly selected seals is given below. What is the probability that the burst strength of the population is greater than 500?

480	489	491	508	501
500	486	499	479	496
499	504	501	496	498

Solution: The mean of these 15 data points is 495.13. The sample standard deviation of these 15 data points is 8.467. The probability that the population mean is greater than 500 is equal to the area under the t probability density function, with 14 degrees of freedom, to the left of:

$$t = \frac{495.13 - 500}{8.467/\sqrt{15}} = 2.227$$

From the t table in the Appendix, the area under the t probability density function, with 14 degrees of freedom, to the left of -2.227 is 0.0214. This value must be interpolated (2.227 falls between the 0.025 value of 2.145 and the 0.010 value of 2.624) but can be computed directly using electronic spreadsheets, or calculators. Simply stated, making an inference from the sample of 15 data points, there is a 2.14% possibility that the true population mean is greater than 500.

Other Distributions

Other less commonly used distributions include the following:

- **Hypergeometric**
- **Bivariate**
- **Exponential**
- **Lognormal**
- **Weibull**

Hypergeometric Distribution

The hypergeometric distribution is used to model discrete data. The hypergeometric distribution applies when the population size, N, is small compared to the sample size, or stated another way, when the sample, n, is a relatively large proportion of the population (n >0.1N). Sampling is done without replacement. The hypergeometric distribution is a complex combination calculation and is used when the defined occurrences are known or can be calculated.

The number of successes, r, in the sample follows the hypergeometric function:

$$P(r) = \frac{C_r^d \; C_{n-r}^{N-d}}{C_n^N}$$

Where:

N = population size
n = sample size
d = number of occurrences in the population
N - d = number of non occurrences in the population
r = number of occurrences in the sample

The term x is used instead of r in many texts.

The hypergeometric distribution is similar to the binomial distribution. Both are used to model the number of successes given a fixed number of trials and two possible outcomes on each trial. The difference is that the binomial distribution requires the probability of success to be the same for all trials, while the hypergeometric distribution does not.

Hypergeometric Distribution (Continued)

Example 7.25: From a group of 20 products, 10 are selected at random for testing. What is the probability that the 10 selected contain the 5 best units?

$$N = 20,\ n = 10,\ d = 5,\ (N\text{-}d) = 15 \text{ and } r = 5$$

$$P(r) = \frac{C_5^5 C_5^{15}}{C_{10}^{20}} \quad \left(\text{note that } C_r^n = \frac{n!}{r!(n-r)!}\right)$$

$$P(r) = \frac{\left(\frac{5!}{5!0!}\right)\left(\frac{15!}{5!10!}\right)}{\left(\frac{20!}{10!10!}\right)} = \left(\frac{15!}{5!10!}\right)\left(\frac{10!10!}{20!}\right) = 0.0163 = 1.63\%$$

The hypergeometric distribution using different r values is shown in Figure 7.25.

$$P(r) = \binom{d}{r}\frac{\binom{N-d}{n-r}}{\binom{N}{n}}$$

n = sample size
r = number of occurrences
d = occurrences in population
N = population size

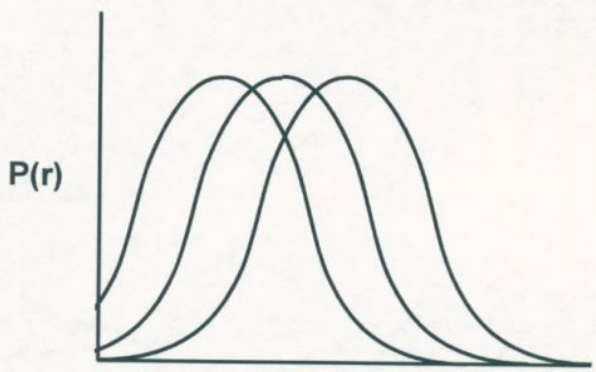


Figure 7.25 Hypergeometric Distribution Example

The mean and the variance of the hypergeometric distribution are:

$$\mu = \frac{nm}{N} \qquad \sigma^2 = \left(\frac{nm}{N}\right)\left(1 - \frac{m}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Note that m = d in this version of the equation.

Choosing the Correct Discrete Distribution

To determine the correct discrete distribution, ask the following questions.

1. Is a rate being modeled, such as defects per car, and is there no upper bound on the number of possible occurrences? If the answer is yes, the Poisson distribution is probably the appropriate distribution. If the answer is no, go to question 2.

2. Is there a fixed number of trials? If yes, go to question 3. If no, is there a fixed number of successes with the number of trials being the random variable? If the answer is yes, use either the geometric or negative binomial distributions.

3. Is the probability of success the same on all trials? If yes, use the binomial distribution, if no, use the hypergeometric distribution. The hypergeometric is sampling without replacement. The probability of success may not be the same for all trials.

These questions are summarized in the flow chart shown in Figure 7.26.

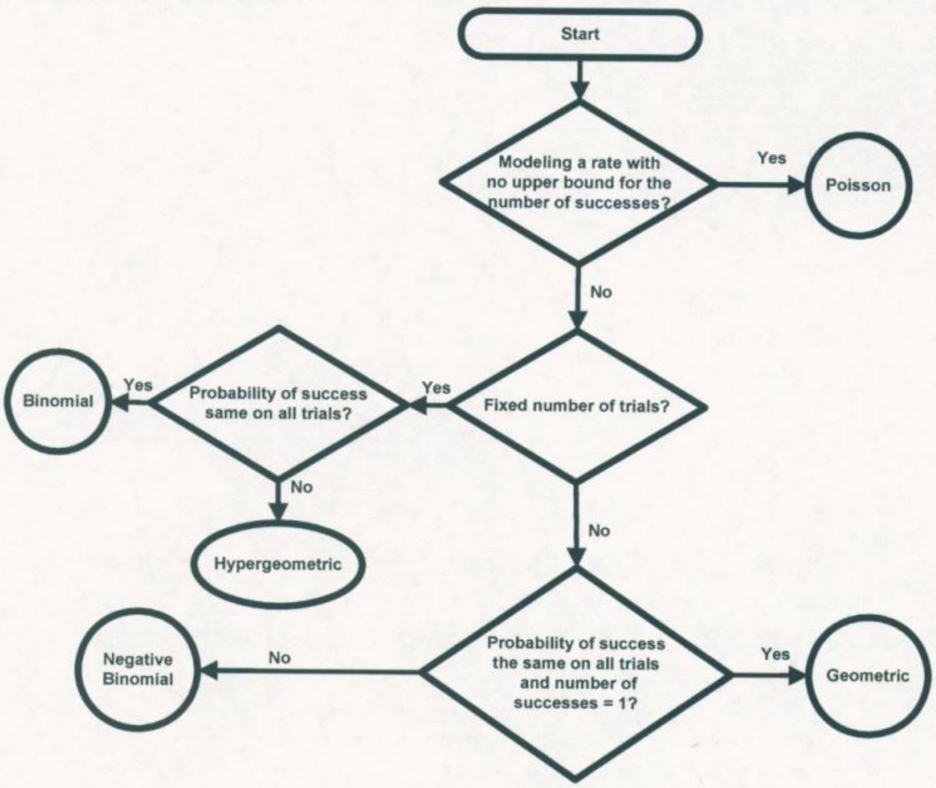


Figure 7.26 Discrete Distribution Flow Chart

The geometric and negative binomial are not summarized.

Bivariate Normal Distribution

The joint distribution of two variables is called a bivariate distribution. Bivariate distributions may be discrete or continuous. There may be total independence of the two independent variables, or there may be a covariance between them.

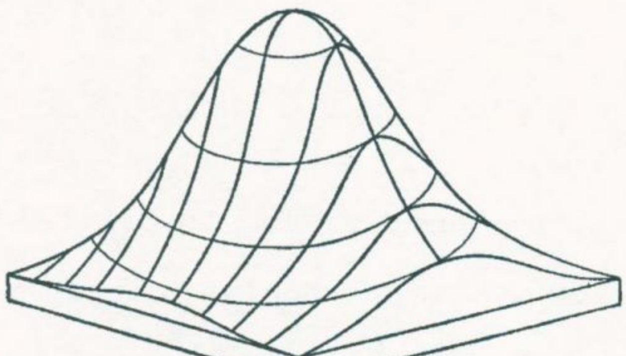
The graphical representation of a bivariate distribution is a three dimensional plot, with the x and y-axis representing the independent variables and the z-axis representing the frequency for discrete data or the probability for continuous data.

A special case of the bivariate distribution is the bivariate normal distribution, in which there are two random variables. For this case, the bivariate normal density is given by Freund (1962)[7] as:

$$f(x_1, x_2) = \frac{\exp\left[-\dfrac{\dfrac{(x_1-\mu_1)}{\sigma_1^2} - \dfrac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \dfrac{(x_2-\mu_2)}{\sigma_2^2}}{2(1-\rho^2)}\right]}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

Where: μ_1 and μ_2 are the two means
σ_1 and σ_2 are the two variances and are each > 0
ρ is the correlation coefficient of the two random variables

The bivariate normal distribution surface is shown in Figure 7.27. Note that the maximum occurs at $x_1 = \mu_1$ and $x_2 = \mu_2$.



(Freund, 1962)[7]

Figure 7.27 Bivariate Normal Distribution Surface

Additional information on bivariate distributions may be found in Duncan (1986)[6].

Exponential Distribution

The exponential distribution applies to the useful life cycle of many products. The exponential distribution is used to model items with a constant failure rate.

The exponential distribution is closely related to the Poisson distribution. If a random variable, x, is exponentially distributed, then the reciprocal of x, y = 1/x follows a Poisson distribution. Likewise, if x is Poisson distributed, then y = 1/x is exponentially distributed. Because of this behavior, the exponential distribution is usually used to model the mean time between occurrences, such as arrivals or failures, and the Poisson distribution is used to model occurrences per interval, such as arrivals, failures, or defects. The exponential probability density function is:

$$f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}} = \lambda e^{-\lambda x} \quad , x \geq 0$$

Where: λ is the failure rate and θ is the mean

From the equation above, it can be seen that $\lambda = 1/\theta$. The exponential probability density function is shown in Figure 7.28.

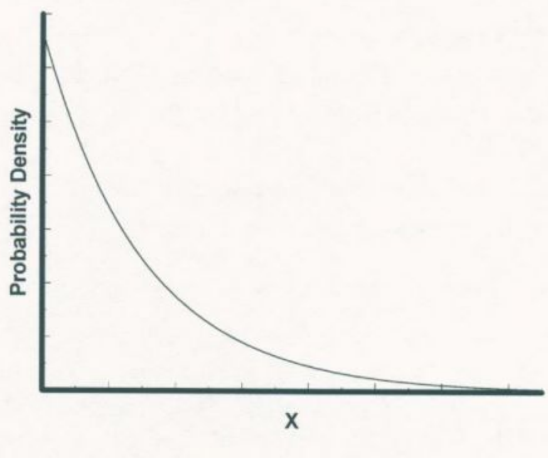


Figure 7.28 Exponential Probability Density Function

The variance of the exponential distribution is equal to the mean squared.

$$\sigma^2 = \theta^2 = \frac{1}{\lambda^2} \quad \text{hence} \quad \sigma = \theta = \frac{1}{\lambda}$$

The exponential distribution is characterized by its hazard function which is constant. Because of this, the exponential distribution exhibits a lack of memory. That is, the probability of survival for a time interval, given survival to the beginning of the interval, is dependent only on the length of the interval.

Lognormal Distribution

If a data set is known to follow a lognormal distribution, transforming the data by taking a logarithm yields a data set that is approximately normally distributed. This is shown in Table 7.29.

Original Data	Normalized Data	ln
12	ln(12)	2.48
28	ln(28)	3.33
87	ln(87)	4.47
143	ln(143)	4.96

Table 7.29 Transformation of Lognormal Data

The most common transformation is made by taking the natural logarithm, but any base logarithm, also yields an approximate normal distribution. The remaining discussion will use the natural logarithm denoted as "ln".

When random variables are summed, as the sample size increases, the distribution of the sum becomes a normal distribution, regardless of the distribution of the individuals. Since lognormal random variables are transformed to normal random variables by taking the logarithm, when random variables are multiplied, as the sample size increases, the distribution of the product becomes a lognormal distribution regardless of the distribution of the individuals. This is because the logarithm of the product of several variables is equal to the sum of the logarithms of the individuals. This is shown below:

$$y = x_1 \, x_2 \, x_3$$
$$\ln y = \ln x_1 + \ln x_2 + \ln x_3$$

The standard lognormal probability density function is:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2} \text{ , } x > 0$$

Where: μ is the location parameter or mean of the natural logarithms of the individual values

σ is the scale parameter or standard deviation of natural logarithms of the individual values. Some references show σ as the shape parameter.

Lognormal Distribution (Continued)

The location parameter is the mean of the data set after transformation by taking the logarithm, and the scale (or shape) parameter is the standard deviation of the data set after transformation.

The lognormal distribution takes on several shapes depending on the value of the shape parameter. The lognormal distribution is skewed right, and the skewness increases as the value of σ increases. This is shown in Figure 7.30.

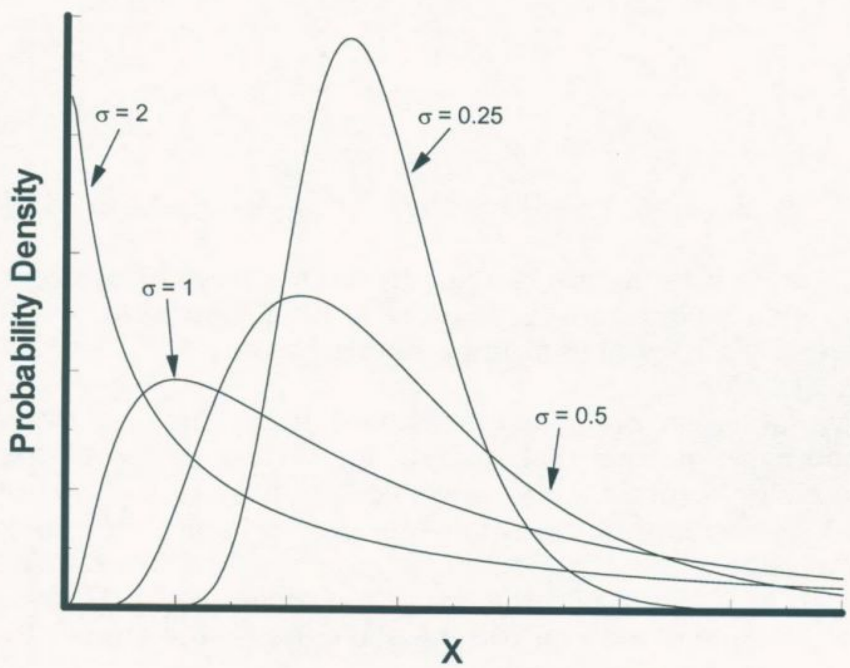


Figure 7.30 Lognormal Probability Density Function

The mean of the lognormal distribution can be computed from its parameters:

$$\text{mean} = e^{\left(\mu + \sigma^2/2\right)}$$

The variance of the lognormal distribution is:

$$\text{variance} = \left(e^{2\mu + \sigma^2}\right)\left(e^{\sigma^2} - 1\right)$$

Where μ and σ^2 are the mean and variance of natural log values.

(Dovich, 2009)[5]

Weibull Distribution

The Weibull distribution is one of the most widely used distributions in reliability and statistical applications. It is commonly used to model time to fail, time to repair, and material strength. There are two common versions of the Weibull distribution, the two parameter Weibull and the three parameter Weibull. The difference is the three parameter Weibull distribution has a location parameter when there is some non-zero time to first failure.

The three parameter Weibull probability density function is:

$$f(x) = \frac{\beta}{\theta}\left(\frac{x - \delta}{\theta}\right)^{\beta - 1} \exp -\left(\frac{x - \delta}{\theta}\right)^{\beta} \quad \text{, for } x \geq \delta$$

Where: β is the shape parameter
θ is the scale parameter
δ is the location parameter

The three parameter Weibull distribution can also be expressed as:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t - \gamma}{\eta}\right)^{\beta - 1} \exp -\left(\frac{t - \gamma}{\eta}\right)^{\beta} \quad \text{, for } t \geq \gamma$$

Where: β is the shape parameter
η is the scale parameter (determines the width of the distribution)
γ is the non-zero location parameter (the point below which there are no failures)

Note: The Weibull discussion on the following pages will use θ for the scale parameter and δ for the location parameter.

The shape parameter is what gives the Weibull distribution its flexibility. By changing the value of the shape parameter, the Weibull distribution can model a wide variety of data. If $\beta = 1$ the Weibull distribution is identical to the exponential distribution, if $\beta = 2$, the Weibull distribution is identical to the Rayleigh distribution; if β is between 3 and 4, the Weibull distribution approximates the normal distribution.

The Weibull distribution approximates the lognormal distribution for several values of β. For most populations, more than fifty samples are required to differentiate between the Weibull and lognormal distributions. The effect of the shape parameter on the Weibull distribution is shown in Figure 7.31.

Weibull Distribution (Continued)

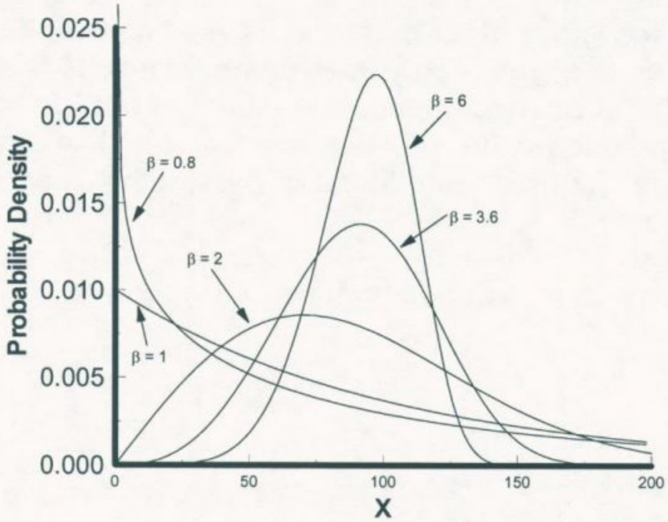


Figure 7.31 Effect of the Weibull Shape Parameter, β (with θ = 100 and δ = 0).

The scale parameter determines the range of the distribution. The scale parameter is also known as the *characteristic life* if the location parameter is equal to zero. If δ does not equal zero, the characteristic life is equal to θ+δ; 63.2% of all values fall below the characteristic life regardless of the value of the shape parameter.

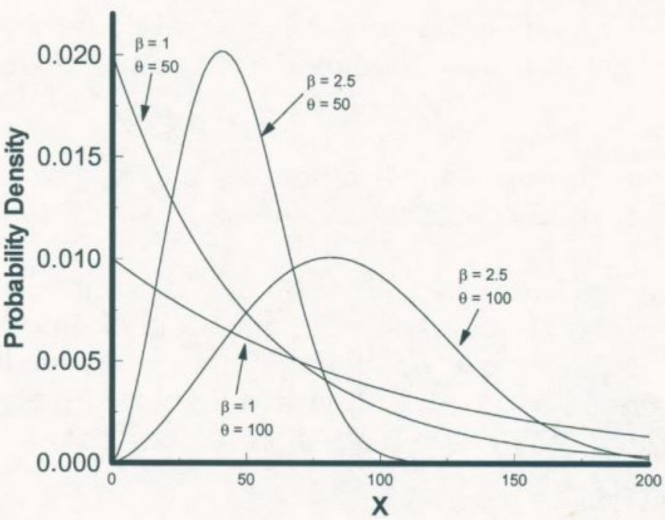


Figure 7.32 Effect of the Weibull Scale Parameter

Weibull Distribution (Continued)

The location parameter is used to define a failure-free zone. The probability of failure when x is less than δ is zero. When $\delta>0$, there is a period when no failures can occur. When $\delta<0$, failures have occurred before time equals 0. At first this seems illogical, but a negative location parameter is caused by shipping failed units, failures during transportation, and shelf-life failures. Generally, the location parameter, δ, is assumed to be zero. The effect of the location parameter is shown in Figure 7.33.

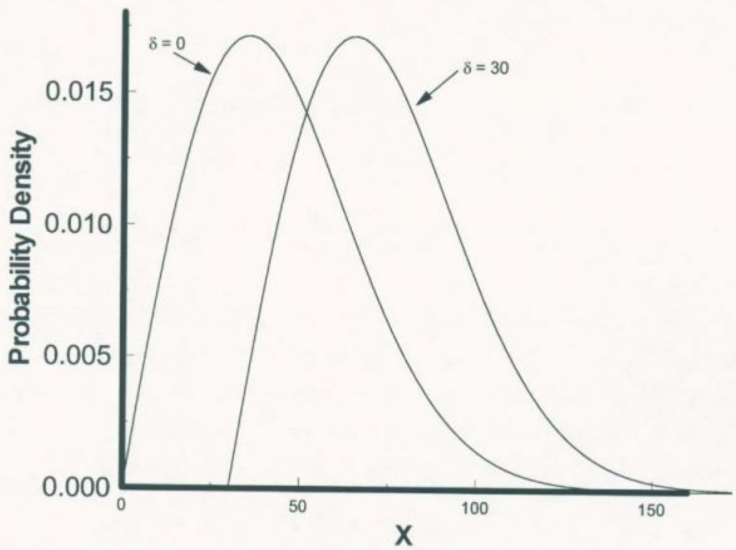


Figure 7.33 Effect of the Weibull Location Parameter

The mean and variance of the Weibull distribution are computed using the gamma distribution. The mean of the Weibull distribution is equal to the characteristic life if the shape parameter is equal to one.

The mean of the Weibull distribution is:

$$\mu = \theta\Gamma\left(1 + \frac{1}{\beta}\right)$$

The variance of the Weibull distribution is:

$$\sigma^2 = \theta^2\left[\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)\right]$$

The variance of the Weibull distribution decreases as the value of the shape parameter increases.

Process Capability

Process Capability is presented in the following topic areas:

- Capability studies
- Capability indices
- Performance indices
- Short-term vs. long-term
- Non-normal data
- Attributes data
- Performance metrics

The above topics are presented in a slightly different order than the ASQ BOK.

Process Capability Studies

The determination of process capability requires a predictable pattern of statistically stable behavior (most frequently a bell-shaped curve) where the chance causes of variation are compared to the engineering specifications. A capable process is a process whose spread on the bell-shaped curve is narrower than the tolerance range or specification limits. USL is the upper specification limit and LSL is the lower specification limit.

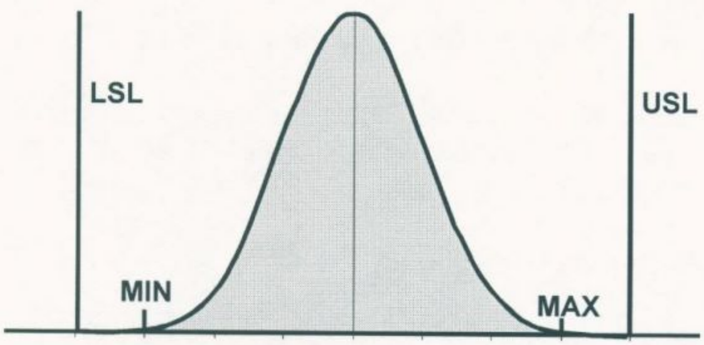


Figure 7.34 A Comparison of Process Spread to Tolerance Range

It is often necessary to compare the process variation with the engineering or specification tolerances to judge the suitability of the process. Process capability analysis addresses this issue. A process capability study includes three steps:

- Planning for data collection
- Collecting data
- Plotting and analyzing the results

Process Capability Studies (Continued)

The objective of process quality control is to establish a state of control over the manufacturing process and then maintain that state of control through time. Actions that change or adjust the process are frequently the result of some form of capability study. When the natural process limits are compared with the specification range, any of the following possible courses of action may result:

- Do nothing. If the process limits fall well within the specification limits, no action may be required.

- Change the specifications. The specification limits may be unrealistic. In some cases, specifications may be set tighter than necessary. Discuss the situation with the final customer to see if the specifications may be relaxed or modified.

- Center the process. When the process spread is approximately the same as the specification spread, an adjustment to the centering of the process may bring the bulk of the product within specifications.

- Reduce variability. This is often the most difficult option to achieve. It may be possible to partition the variation (stream-to-stream, within piece, batch-to-batch, etc.) and work on the largest offender first. For a complicated process, an experimental design may be used to identify the leading source of variation.

- Accept the losses. In some cases, management must be content with a high loss rate (at least temporarily). Some centering and reduction in variation may be possible, but the principal emphasis is on handling the scrap and rework efficiently.

Other capability applications:

- Providing a basis for setting up a variables control chart
- Evaluating new equipment
- Reviewing tolerances based on the inherent variability of a process
- Assigning more capable equipment to tougher jobs
- Performing routine process performance audits
- Determining the effects of adjustments during processing

Modified from Juran (1999)[8]

Identifying Characteristics

The identification of characteristics to be measured in a process capability study should meet the following requirements:

- The characteristic should be indicative of a key factor in the quality of the product or process
- It should be possible to adjust the value of the characteristic
- The operating conditions that affect the measured characteristic should be defined and controlled

If a part has fourteen different dimensions, process capability would not normally be performed for all of these dimensions. Selecting one, or possibly two, key dimensions provides a more manageable method of evaluating the process capability. For example in the case of a machined part, the overall length or the diameter of a hole might be the critical dimension. The characteristic selected may also be determined by the history of the part and the parameter that has been the most difficult to control or has created problems in the next higher level of assembly.

Customer purchase order requirements or industry standards may also determine the characteristics that are required to be measured. In the automotive industry, the *Production Part Approval Process (PPAP)* (AIAG, 2006)[1] states "An acceptable level of preliminary process capability must be determined prior to submission for all characteristics designated by the customer or supplier as safety, key, critical, or significant, that can be evaluated using variables (measured) data." Chrysler, Ford and General Motors use symbols to designate safety and/or government regulated characteristics and important performance, fit, or appearance characteristics.

(AIAG, 2006)[1]

Identifying Specifications/Tolerances

The process specifications or tolerances are determined either by customer requirements, industry standards, or the organization's engineering department. Various process capability indices are described later in this Section.

The process capability study is used to demonstrate that the process is centered within the specification limits and that the process variation predicts the process is capable of producing parts within the tolerance requirements.

When the process capability study indicates the process is not capable, the information is used to evaluate and improve the process in order to meet the tolerance requirements. There may be situations where the specifications or tolerances are set too tight in relation to the achievable process capability. In these circumstances, the specification must be reevaluated. If the specification cannot be opened, then the action plan is to perform 100% inspection of the process, unless inspection testing is destructive.

Developing Sampling Plans

The appropriate sampling plan for conducting process capability studies depends upon the purpose and whether there are customer or standards requirements for the study. Ford and General Motors specify that process capability studies for PPAP submissions be based on data taken from a significant production run of a minimum of 300 consecutive pieces. (AIAG, 2006)[1]

If the process is currently running and is in control, control chart data may be used to calculate the process capability indices. If the process fits a normal distribution and is in statistical control, then the standard deviation can be estimated from:

$$\sigma_R \approx \frac{\bar{R}}{d_2}$$

For new processes, for example for a project proposal, a pilot run may be used to estimate the process capability. The disadvantage of using a pilot run is that the estimated process variability is most likely less than the process variability expected from an ongoing process.

Process capabilities conducted for the purpose of improving the process may be performed using a design of experiments (DOE) approach in which the optimum values of the process variables which yield the lowest process variation is the objective.

Verifying Stability and Normality

If only common causes of variation are present in a process, then the output of the process forms a distribution that is stable over time and is predictable. If special causes of variation are present, the process output is not stable over time.

(AIAG, 2008)[2]

Figure 7.35 depicts an unstable process with both process average and variation out-of-control. Note, the process may also be unstable if either the process average or variation is out-of-control.

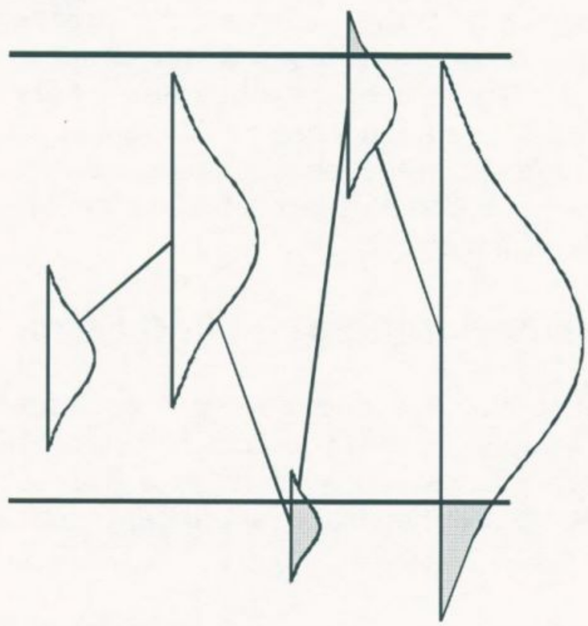


Figure 7.35 Unstable Process with Average and Variation Out-of-control

Common causes of variation refer to the many sources of variation within a process that has a stable and repeatable distribution over time. This is called a state of statistical control and the output of the process is predictable. Special causes refer to any factors causing variation that are not always acting on the process. If special causes of variation are present, the process distribution changes and the process output is not stable over time. (AIAG, 2008)[2]

When plotting a process on a control chart, lack of process stability can be shown by several types of patterns including: points outside the control limits, trends, points on one side of the center line, cycles, etc.

Verifying Stability and Normality (Continued)

The validity of the normality assumption may be tested using the chi square hypothesis test. To perform this test, the data is partitioned into data ranges. The number of data points in each range is then compared with the number predicted from a normal distribution. Using the hypothesis test with a selected confidence level, a conclusion can be made as to whether the data follows a normal distribution. The chi square hypothesis test is:

H_0: The data follows a specified distribution
H_1: The data does not follow a specified distribution

and is tested using the following test statistic:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

(NIST, 2001)[11]

Refer to Section VIII for the definition of terms in the above equation and additional chi square information.

Continuous data may be tested using the Kolmogorov-Smirnov goodness-of-fit test. It has the same hypothesis test as the chi square test, and the test statistic is given by (NIST, 2001)[11]:

$$D = \max_{1 \le i \le N} \left| F(Y_i) - \frac{i}{N} \right|$$

Where D is the test statistic and F is the theoretical cumulative distribution of the continuous distribution being tested. An attractive feature of this test is that the distribution of the test statistic does not depend on the underlying cumulative distribution function being tested. Limitations of this test are that it only applies to continuous distributions and that the distribution must be fully specified. The location, scale, and shape parameters must be specified and not estimated from the data.

The Anderson-Darling test is a modification of the Kolmogorov-Smirnov test and gives more weight to the tails of the distribution. See the (NIST, 2001)[11] reference at the end of this Section for further discussion of distribution tests.

If the data does not fit a normal distribution, the chi square hypothesis test may also be used to test the fit to other distributions such as the exponential or binomial distributions. Refer to non-normal data transformations discussed later in this Section.

The Normal Distribution

When all special causes of variation are eliminated, many variable data processes, when sampled and plotted, produce a bell-shaped distribution. If the base of the histogram is divided into six (6) equal lengths (three on each side of the average), the amount of data in each interval exhibits the following percentages:

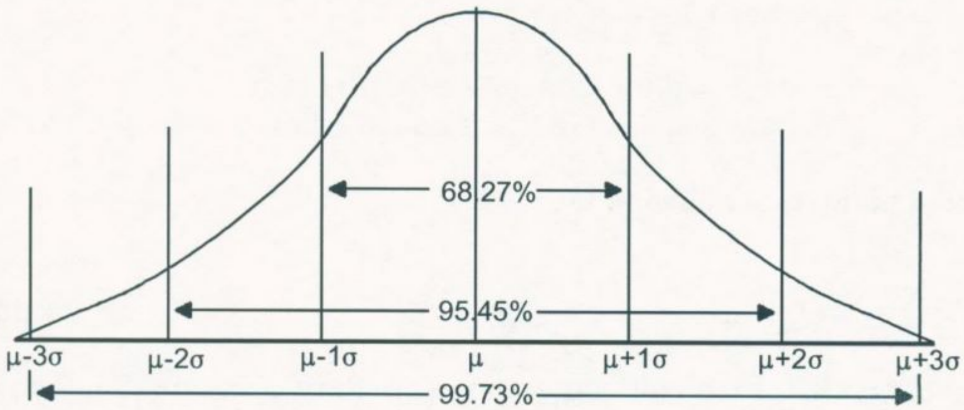


Figure 7.36 The Normal Distribution

The Z Value

The area outside of specification for a normal curve can be determined by a Z value.

$$Z_{LOWER} = \frac{\overline{X} - LSL}{S} \qquad Z_{UPPER} = \frac{USL - \overline{X}}{S}$$

The Z transformation formula is:

$$Z = \frac{X - \mu}{\sigma}$$

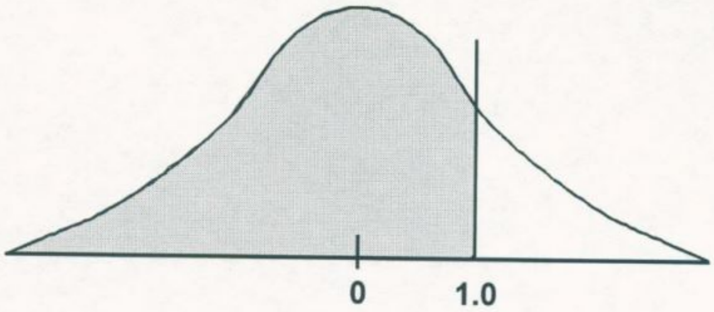
Where: x **= data value (the value of concern)**
μ **= mean**
σ **= standard deviation**

This transformation will convert the original values to the number of standard deviations away from the mean. The result allows one to use a single standard normal table to describe areas under the curve (probability of occurrence).

Z Value (Continued)

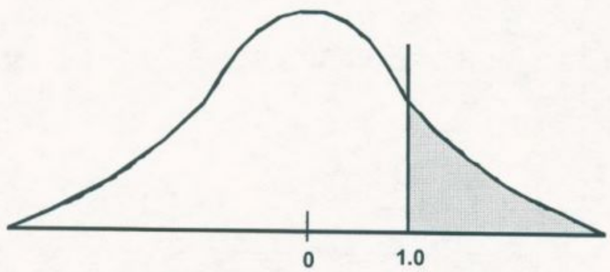
There are several ways to display the normal (standardized) distribution:

1. **As a number under the curve up to the z value:**



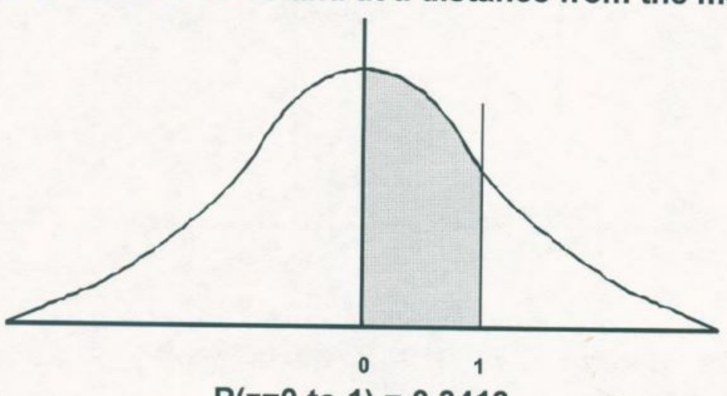
P(z = - ∞ to 1) = 0.8413

2. **As a number beyond the z value:**



P(z =1 to +∞) = 0.1587

3. **As a number under the curve and at a distance from the mean:**

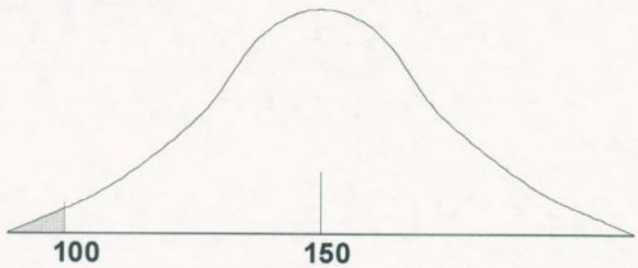


P(z=0 to 1) = 0.3413

The standard normal table in this Primer uses the second method of calculating the probability of occurrence.

Z Value Examples

Example 7.26: Tenth grade students weights follow a normal distribution with a mean μ = 150 lb and a standard deviation of 20 lb. What is the probability of a student weighing less than 100 lb?



$\mu = 150$

$X = 100$

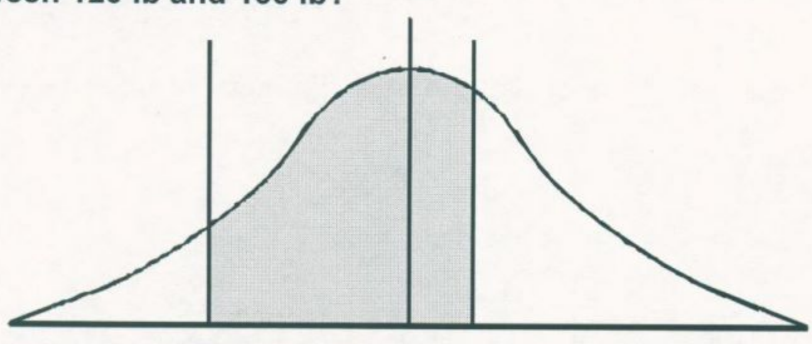
$\sigma = 20$

$$Z = \frac{X - \mu}{\sigma} = \frac{100 - 150}{20} = -\frac{50}{20} = -2.5$$

Since the normal table has values about the mean, a Z value of - 2.5 can be treated as 2.5.

P(Z = - ∞ to -2.5) = 0.0062. That is, 0.62% of the students will weigh less than 100 lb.

Example 7.27: Using the data from Example 7.26, what is the probability of a student weighing between 120 lb and 160 lb?



The best technique to solve this problem, using the standard normal table in this Primer, would be to determine the tail area values, and to subtract them from the total probability of 1.

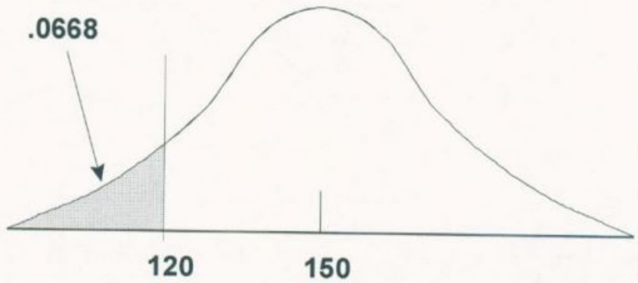
Z Value Examples (Continued)

Example 7.27 (continued):

First, determine the Z value and probability below 120 lb.

$$Z = \frac{x - \mu}{\sigma} = \frac{120 - 150}{20} = -\frac{30}{20} = -1.5$$

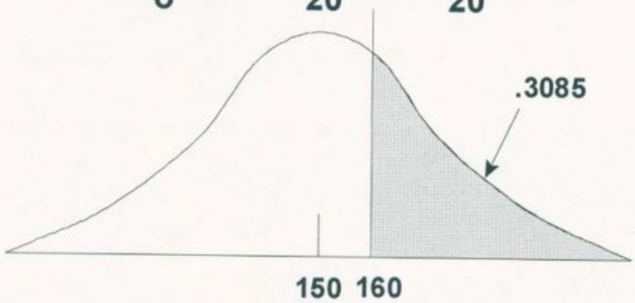
$P(Z = -\infty \text{ to } -1.5) = 0.0668$



Second, determine the Z value and probability above 160 lb.

$P(Z = 0.5 \text{ to } +\infty) = 0.3085$

$$Z = \frac{x - \mu}{\sigma} = \frac{160 - 150}{20} = \frac{10}{20} = 0.5$$



Third, the total probability - below - above = probability between 120 and 160 lb.

$$1 - 0.0668 - 0.3085 = 0.6247$$

Thus, 62.47% of the students will weigh more than 120 lb but less than 160 lb.

Capability Index Failure Rates

There is a direct link between the calculated C_p (and P_p values) with the standard normal (Z value) table. A C_p of 1.0 is the loss suffered at a Z value of 3.0 (doubled, since the table is one sided). Refer to Table 7.37 below.

C_p	Z value	ppm
0.33	1.00	317,311
0.67	2.00	45,500
1.00	3.00	2,700
1.10	3.30	967
1.20	3.60	318
1.30	3.90	96
1.33	4.00	63
1.40	4.20	27
1.50	4.50	6.8
1.60	4.80	1.6
1.67	5.00	0.57
1.80	5.40	0.067
2.00	6.00	0.002

Table 7.37 Failure Rates for C_p and Z Values

In Table 7.37, ppm equals parts per million of nonconformance (or failure) when the process:

- Is centered on $\overline{X}$
- Has a two-tailed specification
- Is normally distributed
- Has no significant shifts in average or dispersion

When the C_p, C_{pk}, P_p, and P_{pk} values are 1.0 or less, Z values and the standard normal table can be used to determine failure rates. With the drive for increasingly dependable products, there is a need for failure rates in the C_p range of 1.5 to 2.0.

Process Capability Indices

To determine process capability, an estimation of sigma is necessary:

$$\sigma_R = \frac{\bar{R}}{d_2}$$

σ_R is an estimate of process capability sigma and comes from a control chart.

The capability index is defined as:

$$C_P = \frac{(USL - LSL)}{6\sigma_R}$$

As a rule of thumb:

$C_P > 1.33$ Capable
$C_P = 1.00$ to 1.33 Capable with tight control
$C_P < 1.00$ Incapable

The capability ratio is defined as:

$$C_R = \frac{6\sigma_R}{(USL - LSL)}$$

As a rule of thumb:

$C_R < 0.75$ Capable
$C_R = 0.75$ to 1.00 Capable with tight control
$C_R > 1.00$ Incapable

Note, this rule of thumb logic is somewhat out of step with the six sigma assumption of a ±1.5 sigma shift. The above formulas only apply if the process is centered, stays centered within the specifications, and $C_P = C_{pk}$.

Process Capability Indices (Continued)

C_{pk} is the ratio giving the smallest answer between:

$$C_{pk} = \frac{USL - \overline{X}}{3\sigma_R} \text{ or } \frac{\overline{X} - LSL}{3\sigma_R}$$

Process Capability Exercise

Example 7.28: For a process with $\overline{X} = 12$, $\sigma_R = 2$ an USL = 16 and LSL = 4, determine C_p and $C_{pk\ min}$:

$$C_p = \frac{USL - LSL}{6\sigma_R} = \frac{16 - 4}{6(2)} = \frac{12}{12} = 1$$

$$C_{pk\ upper} = \frac{USL - \overline{X}}{3\sigma_R} = \frac{16 - 12}{3(2)} = \frac{4}{6} = 0.667$$

$$C_{pk\ lower} = \frac{\overline{X} - LSL}{3\sigma_R} = \frac{12 - 4}{3(2)} = \frac{8}{6} = 1.333$$

$$C_{pk\ min} = C_{pk\ upper} = 0.667$$

C_{pm} Index

The C_{pm} index is defined as:

$$C_{pm} = \frac{USL - LSL}{6\sqrt{(\mu - T)^2 + \sigma^2}}$$

Where: USL = upper specification limit
LSL = lower specification limit
μ = process mean
T = target value
σ = process standard deviation

C_{pm} is based on the Taguchi index, which places more emphasis on process centering on the target.

(Breyfogle, 1999)[4]

C_{pm} Index Exercise

Example 7.29: For a process with μ = 12, σ = 2, T = 10, USL = 16 and LSL = 4, determine C_{pm}:

$$C_{pm} = \frac{USL - LSL}{6\sqrt{(\mu - T)^2 + \sigma^2}}$$

$$C_{pm} = \frac{16 - 4}{6\sqrt{(12 - 10)^2 + 2^2}}$$

$$C_{pm} = 0.707$$

Process Performance Indices

To determine process performance, an estimation of sigma is necessary:

$$\sigma_i = \sqrt{\frac{\sum (X - \bar{X})^2}{(n - 1)}}$$

σ_i is a measure of total data sigma and generally comes from a calculator or computer.

The performance index is defined as:

$$P_P = \frac{(USL - LSL)}{6\sigma_i}$$

The performance ratio is defined as:

$$P_R = \frac{6\sigma_i}{(USL - LSL)}$$

P_{pk} is the ratio giving the smallest answer between:

$$P_{pk} = \frac{(USL - \bar{X})}{3\sigma_i} \quad \text{or} \quad \frac{(\bar{X} - LSL)}{3\sigma_i}$$

Short-Term and Long-Term Capability

Up to this point, process capability has been discussed in terms of stable processes, with assignable causes removed. In fact, the process average and spread are dependent upon the number of units measured or the duration over which the process is measured.

When a process capability is determined using one operator on one shift, with one piece of equipment, and a homogeneous supply of materials, the process variation is relatively small. As factors for time, multiple operators, various lots of material, environmental changes, etc. are added, each of these contributes to increasing the process variation. Control limits based on a short-term process evaluation are closer together than control limits based on the long-term process.

Smith (2001)[15] describes a short run with respect to time and a small run, where there is a small number of pieces produced. When a small amount of data is available, there is generally less variation than is found with a larger amount of data. Control limits based on the smaller number of samples will be narrower than they should be, and control charts will produce false out-of-control patterns.

Smith suggests a modified $\bar{X}$ and R chart for short runs, running an initial 3 to 10 pieces without adjustment. A calculated value is compared with a critical value and either the process is adjusted or an initial number of subgroups is run. Inflated D_4 and A_2 values are used to establish control limits. Control limits are recalculated after additional groups are run.

For small runs, with a limited amount of data, Smith recommends the use of the X and MR chart. The X represents individual data values, not an average, and the MR is the moving range, a measure of piece-to-piece variability. Process capability or C_{pk} values determined from either of these methods must be considered preliminary information. As the number of data points increases, the calculated process capability will approach the true capability.

When comparing attribute with variable data, variable data generally provides more information about the process, for a given number of data points. Using variables data, a reasonable estimate of the process mean and variation can be made with 25 to 30 groups of five samples each. Whereas a comparable estimate using attribute data may require 25 groups of 50 samples each. Using variables data is preferable to using attribute data for estimating process capability.

Information on rational subgrouping and breakdown of variation is given in Section X.

Process Capability for Non-normal Data (Continued)

A probability plot can also be used to display the non-normal data. The data points are clustered to the left with some extreme points to the right. Since this is a non-normal distribution, a traditional process capability index is meaningless. Refer to Figure 7.40.

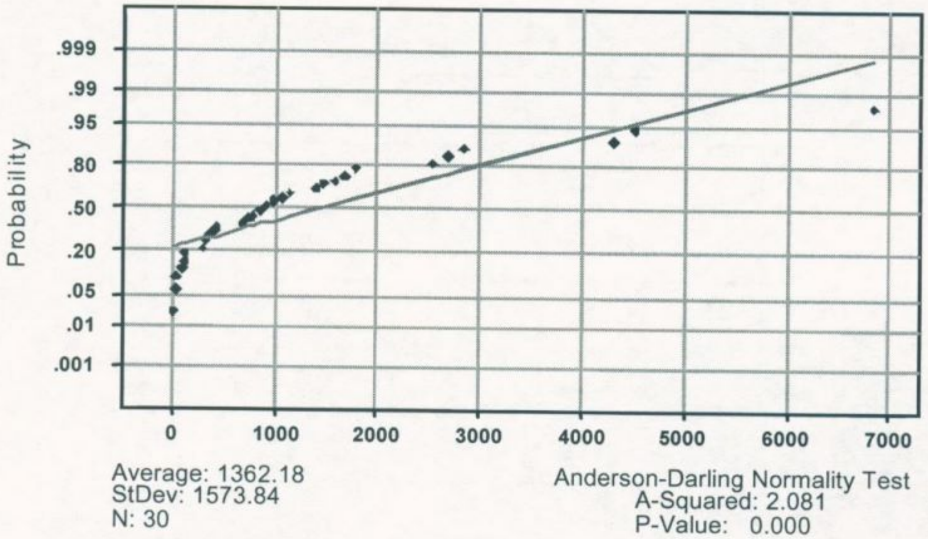


Figure 7.40 Probability Plot for Raw Data

If the investigator has some awareness of the history of the data, and knows it to follow a Poisson distribution, then a square root transformation is a possibility. The standard deviation is the square root of the mean. Some typical data transformations include:

- **Log transformation (log x)**
- **Square root or power transformation (x^y)**
- **Exponential (e^x)**
- **Reciprocal (1/x)**

In order to find the right transformation, some exploratory data analysis may be required. Among the useful power transformation techniques is the Box-Cox procedure. The applicable formula is:

$$y' = y^{\lambda}$$

Where lambda, λ, is the power or parameter that must be determined to transform the data. For λ= 2, the data is squared. For λ = 0.5, a square root is needed.

Process Capability for Non-normal Data (Continued)

As an example, take the square root of the data from the first column in Table 7.38.

The equation will be: $y' = y^{0.5}$

Original data	Transformed data
1.46	1.208
118.59	10.89
410.06	20.25
915.06	30.25
1477.63	38.44
2687.39	51.840

Table 7.41 Table of Transformed Data Using $y^{0.5}$

One can also use Excel or Minitab to handle the data calculations and to draw the normal probability plot. With the use of Minitab, an investigator can let the Box-Cox tool automatically find a suitable power transform. In this example, a power transform of 0.337 is indicated. All 30 transformed data points from Table 7.38, using $y' = y^{0.337}$, are shown in Table 7.42 below.

1.137	2.719	2.895	4.597	4.798
5.002	6.702	6.924	7.147	7.372
7.599	8.996	9.234	9.474	9.716
9.960	10.204	10.451	10.699	11.452
11.706	11.961	12.218	12.476	14.053
14.320	14.589	16.785	17.065	19.638

Table 7.42 Box-Cox Transformed Data Using y to the Power of 0.337

Process Capability for Non-normal Data (Continued)

A probability plot of the newly transformed data will show a near normal distribution. See Figure 7.43 below. (Box, 2005)[3], (Neter, 1996)[10]

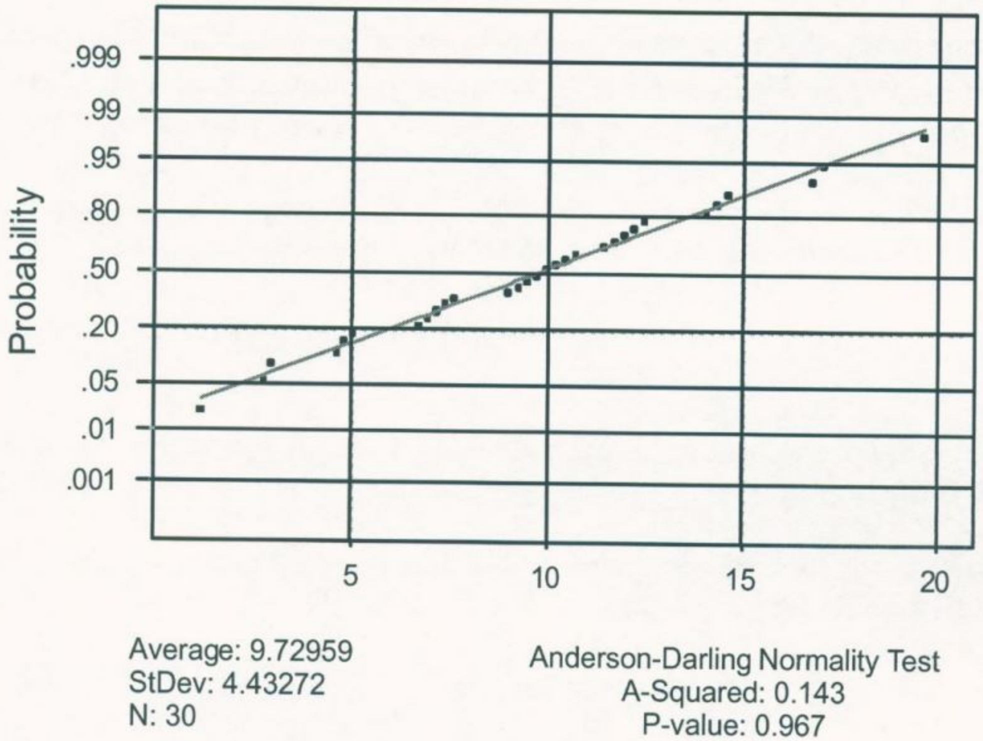


Figure 7.43 Transformed Normal Probability Plot

Now, a process capability index can be determined for the data. However, the investigator must remember to also transform the specifications. If the original specifications were 1 and 10,000, the new limits would be 1 and 22.28.

Process Capability for Attribute Data

The control chart represents the process capability, once special causes have been identified and removed from the process. For attribute charts, capability is defined as the average proportion or rate of nonconforming product. (AIAG, 2008)[2]

- For p charts, the process capability is the process average nonconforming, $\overline{p}$ and is preferably based on 25 or more in-control periods. If desired, the proportion conforming to specification, $1 - \overline{p}$, may be used.

- For np charts, the process capability is the process average nonconforming, $\overline{p}$, and is preferably based on 25 or more in-control periods.

- For c charts, the process capability is the average number of nonconformities, $\overline{c}$, in a sample of fixed size n.

- For u charts, the process capability is the average number of nonconformities per reporting unit, $\overline{u}$.

The average proportion of nonconformities may be reported on a defects per million opportunities scale by multiplying $\overline{p}$ times 1,000,000.

(AIAG, 2008)[2]

Process Performance Metrics

Breyfogle (2003)[4] defines a large number of six sigma measurements with the suggestion that some are controversial and an organization need not use them all. The authors of this Primer have presented only those that are widely used:

Widely Used Symbols

- Defects = D
- Units = U
- Opportunities (for a defect) = O
- Yield = Y

Defect Relationships

- Total opportunities: TO = TOP = U x O
- Defects per unit: $DPU = \frac{D}{U}$ also $= -\ln(Y)$ See yield
- Defects per normalized unit: $= -\ln(Y_{norm})$ See yield
- Defects per unit opportunity: $DPO = \frac{DPU}{O} = \frac{D}{U \times O}$
- Defects per million opportunities: $DPMO = DPO \times 10^6$

Example 7.30: A matrix chart indicates the following information for 100 production units. Determine DPU:

Defects/unit	0	1	2	3	4	5
Number of units	70	20	5	4	0	1

$$DPU = \frac{D}{U} = \frac{1(20) + 2(5) + 3(4) + 5(1)}{100} = \frac{47}{100} = 0.47$$

One would expect to find an average of 0.47 defects per unit.

Example 7.31: Assume that each unit in Example 7.30 had 6 opportunities for a defect (i.e. characteristics A, B, C, D, E, and F). Determine DPO and DPMO.

$$DPO = \frac{DPU}{O} = \frac{0.47}{6} = 0.078333 \qquad DPMO = DPO \times 10^6 = 78,333$$

Process Performance Metrics (Continued)

Yield Relationships

Note, the Poisson equation is normally used to model defect occurrences. If there is a historic defect per unit (DPU) level for a process, the probability that an item contains X flaws (P_x) is described mathematically by the equation:

$$P_{(x)} = \frac{e^{-DPU}\, DPU^{x}}{X!}$$

Where: X is an integer greater or equal to 0
DPU is greater than 0

Note that 0! (zero factorial) = 1 by definition.

If one is interested in the probability of having a defect free unit (as most of us are), then X = 0 in the Poisson formula and the math is simplified:

$$P_{(0)} = e^{-DPU}$$

Therefore, the following common yield formulas follow:

Yield or first pass yield: $Y = FPY = e^{-DPU}$

Defects per unit: $DPU = -\ln(Y)$ **(ln means natural logarithm)**

Rolled throughput yield: $Y_{rt} = RTY = \prod_{i=1}^{n} Y_i$

Normalized yield: $Y_{norm} = \sqrt[n]{RTY}$ **(Where n = # steps)**

Total defects per unit: $TDPU = -\ln(Y_{rt})$

Example 7.32: Assume that a process has a DPU of 0.47. Determine the yield.

$$Y = e^{-DPU} = e^{-0.47} = 0.625 = 62.5\%$$

Example 7.33: For a process with a first pass yield of 0.625 determine the DPU.

$$DPU = -\ln(Y) = -\ln 0.625 = 0.47$$

Process Performance Metrics (Continued)

Yield Relationships (Continued)

Example 7.34: A process consists of 4 sequential steps: 1, 2, 3, and 4. The yield of each step is as follows: $Y_1 = 99\%$, $Y_2 = 98\%$, $Y_3 = 97\%$, $Y_4 = 96\%$. Determine the rolled throughput yield and the total defects per unit.

$$Y_{rt} = \prod_{i=1}^{n} Y_i = (0.99)(0.98)(0.97)(0.96) = 0.90345 = 90.345\%$$

$$TDPU = -\ln(RTY) = -\ln 0.90345 = 0.1015$$

Rolled Throughput Yield (RTY)

Rolled throughput yield is defined as the cumulative calculation of yield or defects through multiple process steps. The determination of the rolled throughput yield (RTY) can help a team focus on serious improvements. Rath (2000)[13] provides the following steps for determination of RTY and how it can influence the project scope:

- Calculate the yield for each step and the resulting RTY
- The RTY for a process will be the baseline metric
- Revisit the project scope
- Significant yield differences can suggest improvement opportunities

Table 7.44 is an example of a RTY calculation and analysis for 5 process steps.

Weld 1	➡	Weld 2	➡	Fab 1	➡	Fab 2	➡	Assembly
yield: 90%		yield: 86%		yield: 92%		yield: 87%		yield: 65%

Table 7.44 Example of RTY Calculation and Analysis

$$RTY = 0.90 \times 0.86 \times 0.92 \times 0.87 \times 0.65 = 0.403$$

The process RTY is only 40.3%. A very significant drop in yield occurs in the assembly process with only 65% yield. This indicates that the assembly process could warrant an initial improvement project.

Process Performance Metrics (Continued)

Sigma Relationships

Probability of a defect = P(d)

$$P(d) = 1 - Y \quad \text{or} \quad 1 - FPY$$

also $P(d) = 1 - Y_{RT}$ (for a series of operations)

P(d) can be looked up in a Z table (using the table in reverse to determine Z).

Example 7.35: The first pass yield for a single operation is 95%. What is the probability of a defect and what is the Z value?

$$P(d) = 1 - 0.95 = 0.05$$

Using the Z table for 0.05 approximates 1.645 sigma.

Schmidt & Launsby

The Z value determined in Example 7.35 is called Z long-term or Z equivalent.

Z short-term is defined as: $Z_{ST} = Z_{LT} + 1.5$ shift

Example 7.36: If Z long-term = 1.645, what is Z short-term?

$$Z_{ST} = Z_{LT} + 1.5 = 1.645 + 1.5 = 3.145$$

Schmidt and Launsby (1997)[14] report that the 6 sigma quality level (with the 1.5 sigma shift) can be approximated by:

$$6 \text{ Sigma Quality Level} = 0.8406 + \sqrt{29.37 - 2.221 \times \ln(ppm)}$$

Example 7.37: If a process were producing 80 defectives/million, what would be the 6 sigma quality level?

$$6\sigma = 0.8406 + \sqrt{29.37 - 2.221 \times \ln(80)}$$
$$= 0.8406 + \sqrt{29.37 - 2.221\,(4.3820)}$$
$$= 0.8406 + 4.4314 = 5.272 \quad \text{(about 5.3)}$$

This answer can be also looked up in Appendix Table II, which (by interpolation) appears to be about 5.3.

References

1. AIAG. (2006). *Production Part Approval Process (PPAP)*. Chrysler Corporation, Ford Motor Company, and General Motors Corporation.

2. AIAG. (2008). *Statistical Process Control (SPC) Reference Manual*. Chrysler Corporation, Ford Motor Company, and General Motors Corporation.

3. Box, G., Hunter, J., & Hunter, W. (2005). *Statistics for Experimenters*, 2nd ed. Hoboken, NJ: John Wiley & Sons.

4. Breyfogle, F. W., III (2003). *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*. New York: John Wiley & Sons.

5. Dovich, R. A., & Wortman, B. L. (2009). *CRE Primer*. Terre Haute, IN: Quality Council of Indiana.

6. Duncan, A. J. (1986). *Quality Control and Industrial Statistics*, 5th ed. Homewood, IL: Richard D. Irwin, Inc.

7. Freund, J. E. (1962). *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice-Hall Inc.

8. Juran, J. M. (1999). *Juran's Quality Handbook*, 5th ed. New York: McGraw-Hill.

9. Lapin, L. L. (1982). *Statistics for Modern Business Decisions*, 3rd ed. New York: Harcourt.

10. Neter, J., Kutner, M., Nachtshem, C., & Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th ed. Boston: McGraw-Hill.

11. NIST, Information Technology Laboratory, Statistical Engineering Division. (2001). *Engineering Statistics Handbook*. Retrieved September 21, 2001 from http://www.itl.nist.gov/div898/handbook/.

12. Omdahl, T. (2009). *Quality Dictionary*. Terre Haute, IN: Quality Council of Indiana.

13. Rath & Strong. (2000). *Rath & Strong's Six Sigma Pocket Guide*. Lexington, MA: Rath & Strong Management Consultants.

14. Schmidt, S. R. & Launsby, R. G. (1997). *Understanding Industrial Designed Experiments*, 4th ed. Colorado Springs, CO: Air Academy Press.

15. Smith, G. M. (2001). *Statistical Process Control and Quality Improvement*. Upper Saddle River, NJ: Prentice-Hall.

References (Continued)

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

17. Wortman, B. L., & Carlson, D. R. (2012). *CQE Primer*. Terre Haute, IN: Quality Council of Indiana.

