Basic Statistics: A Review

by Allan T. Mense, Ph.D., PE, CRE

This is not a textbook on statistics. This is a refresher that presumes the reader has had some statistics background. There are some easy parts and there are some hard parts. The most useful references used to put this material together were:

1) "Engineering Statistics" by Montgomery, Runger and Hubele, 4rd Edition, John Wiley & Sons, 2007. An excellent introduction to statistics for engineers.
2) "Statistical Models in Engineering" by Hahn & Shapiro, Wiley Classics Library, 1994. (Paperback)
3) "Introduction to Error Analysis" by John R. Taylor ("Mister Wizard") University Science Books, Sausalito, CA, 1997.
4) "Quality Engineering Statistics," by Robert A. Dovich, ASQ Quality Press, 1992. (paperback). The outline of topics as well as some examples used in this refresher was taken from this book by Dovich.
5) "Practical Engineering Statistics," by Schiff & D'Agostino, Wiley Interscience, 1991. An excellent small statistics book.

Topics Covered:

Topic 1 Point Estimates

When working with data, typically a small sample from a large population of data, we wish to use this sample to estimate parameters of the overall population. The population may be finite or infinite. In describing a population we typically wish to know where the center resides, how much variation there is in the data about the central value, whether the distribution is symmetric or skewed to one side, and how peaked or flat it is. One possible set of point estimates for data would be the mean, variance, coefficient of skewness, and the coefficient of kurtosis. This is explored in the following sections.

Measures of Central Tendency:

There are three major measures of central tendency of a population; they are the mean, median and mode. We find these parameters by calculating statistical estimators for these parameters using sample data. Generally, we wish to have statistical estimators that give the best unbiased estimates of these population parameters. The population mean is

estimated from the simple arithmetic average of the sample data. If the number of data points in the sample is N the mean is calculated by

1) $$\overline{x} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i$$

where x_i is the value of the i^{th} data point in the sample of size N. This is the unbiased estimator of the population mean.

The median of the population is estimated from the median of the sample data, which is the middle data point from a data sample that is sorted from smallest to largest values (see example below). For N odd, it is the middle data point. For N even the median is the average of the middle two data points.

The mode is simply the most probable value and is determined from the sample data by plotting the data (usually as a histogram) and determining the range of the x-values that most frequently occurs; the center of that range is called the mode. There can be more than one mode, called multi-moded, for a population.

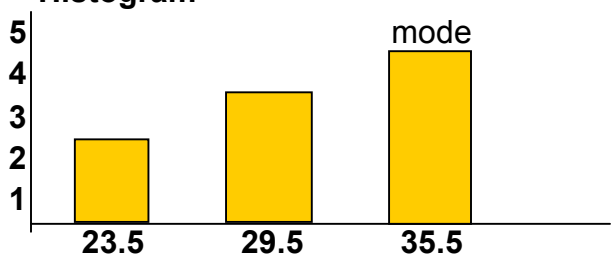
Example 1. Test data

x	value
x1	33
x2	22
x3	35
x4	27
x5	30
x6	26
x7	32
x8	34
x9	37
mean =	30.667
median =	32
mode =	35.5

Histogram of data

3 bins *Range =* 15

lower	upper	midpoint	frequency
21	26	23.5	2
27	32	29.5	3
33	38	35.5	4



The median is determined by sorting the data from smallest to largest values and counting to the middle ((N+1)/2) point. Sorting the above data produces;

22 26 27 30 32 33 34 35 37

The middle value is the (9+1)/2 =5^{th} value in the sorted order which is the number 32. Thus, the median is the point for which 50% of the numbers are bigger than the median and 50% of the numbers are less than the median. If there are an even number of data points then the median is taken to be the average of the middle two data values.

Creating a histogram of the data, as seen above, and finding the value that represents the most frequently occurring range of numbers determines the mode. For example, one easily sees there are more values that lay within the interval 33 to 38 than in the other

intervals. *We generally choose the midpoint of the interval to represent the value of the ordinate on the histogram*. In many samples of real data, there may not be just one peak value when the data is plotted in a histogram. In these multimodal cases, the mode may not be a useful or meaningful measure of central tendency.
When we deal with symmetrical distributions of data such as are represented by a Gaussian or normal or "bell-shaped" distribution the mean = median = mode.
Attribute data:
When dealing with attribute data such as what fraction of M&Ms in a package are blue, we are interested in a ratio (# blue / total # in the package). These proportions are called attribute data. Another example would be "fraction of nonconforming units" which is used when screening units from a supplier. In this instance we would record p = number of nonconforming units / number of units tested. For example if we only had 200 units in the population and all 200 units were tested and 4 failed then $\pi = 4/200 = .02$ is the population fraction defective.

Measures of Variation or Dispersion.

The four useful measures of dispersion are 1) the variance σ^2, which is estimated from the sample data by the statistic s^2, 2) the standard deviation σ, which is the square root of the variance, and is estimated by s (which is biased), 3) the range, R = largest sample value – smallest sample value, and 4) the average of the absolute value of the residuals. By far the most used measures are the variance and the standard deviation.
For a finite population of size N the variance is defined by

2) $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

and μ is the mean of the population $\mu \equiv \frac{1}{N}\sum_{i=1}^{N} x_i$. The standard deviation of the population is found by taking the square root of the population variance. We seldom have the entire population of a variable to use for calculations so we must try and infer information about the population from taking one or more samples of data from the population of interest.
When we only have a sample from the population, then the sample variance is defined by

3) $$s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where n is the size of the sample and $\bar{x}$ is the mean of the sample. The factor of n-1 is necessary to make sure that s^2 is an unbiased estimator of the population variance. More will be said of biased and unbiased estimators in Appendix A. One usually wants unbiased estimators of population parameters. The square root of s^2 is not the unbiased estimator of σ but it is usually close enough.
Excel has functions that calculate the mean and the population and sample variances and standard deviations. Mean = AVERAGE(range of cells containing data), population variance = VARP(range of cells), sample variance = VAR(range of cells), STDEVP(range of cells) = standard deviation of population, STDEV(range of cells) = standard deviation of sample (biased).
Equation 3) can be expanded into a more useful form for hand calculation.

4) $$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{N} x_i\right)^2\right)$$

The expected value of s^2, $E[s^2]$ is σ^2 and is an unbiased estimator of the population variance. The unbiased estimate of σ is not the sample statistic s = square root of the sample variance. The expected value of s, $E[s] = C_4\sigma$ where C_4 is a function of sample size and is given in Appendix B. Thus the unbiased estimate of $\sigma = s/C_4$.

Example 2: n = 25,
Sample data table

1.44	1.96	1.02	1.71	1.95
0.98	2.52	1.79	1.92	1.42
0.85	2.41	1.97	1.87	1.99
1.30	1.79	2.40	1.96	2.38
2.42	2.60	1.49	2.55	1.77

From the above data mean = 1.86, the median = 1.92,

$$\sum_{i=1}^{N} x_i = 46.47, \sum_{i=1}^{N} x_i^2 = 92.4587, \left(\sum_{i=1}^{N} x_i\right)^2 = 2159.602$$

$$s^2 = \frac{1}{24}\left(92.4587 - \frac{1}{25}2159.602\right) = 0.253, s = 0.5031$$

The unbiased estimator of σ is $s/C_4 = 0.5031/0.9896 = 0.508$. (See Appendix B for C_4)

Measure of Dispersion for Attribute Data:

When working with attribute data, such as fraction of nonconforming, the estimate of the variance is taken from the binomial distribution and is given by,

5) $$\sigma^2 = \pi(1-\pi)$$

where π is the fraction nonconforming in the population.

Example 3: If $\pi = 0.034$ then
$\sigma^2 = 0.034(1-0.034) = 0.0328, \sigma = 0.18$.

Measure of Asymmetry: Skewness

Skewness is a measure of asymmetry of data. It is called the third moment about the mean. It is usually represented by a quantity called the coefficient of skewness and is estimated from the data with the formula

6) $$\text{Coef. of Skewness} = \frac{\text{skewness}}{(\text{standard deviation})^3} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3(n-1)(n-2)}$$

If the data has a tail to the right (values > mean) then the skewness will be positive. If the data tails off to the left of the mean one has negative skewness. The data for example 2 gives Coef. Of Skewness= -0.361. We will use the abbreviation Sk for the coefficient of skewness. For the mathematically inclined the calculus definition is given by

7) $$Sk \equiv \frac{1}{\sigma^3}\int (x-\mu)^3 f(x)dx$$

and is called the third moment about the mean divided by the standard deviation cubed in order to produce a unitless measure of asymmetry.

Measure of "Peakedness": Kurtosis

A measure of the peaked nature of the data is called the kurtosis and the usual measure is called the coefficient of kurtosis and is given by the formula

8) $$\text{Coef. of kurtosis}=\frac{\text{kurtosis}}{\left(\text{standard deviation}\right)^4}=\frac{n(n+1)\sum_{i=1}^{x}\left(x_i-\bar{x}\right)^4}{s^4(n-1)(n-2)(n-3)}-3\frac{3n-5}{(n-2)(n-3)}$$

The skewness is called the fourth moment of the distribution about the mean and is given by the calculus formula as

9) $$\text{Coef. of kurtosis=Ku}\equiv\frac{1}{\sigma^4}\int\left(x-\mu\right)^4 f(x)dx$$

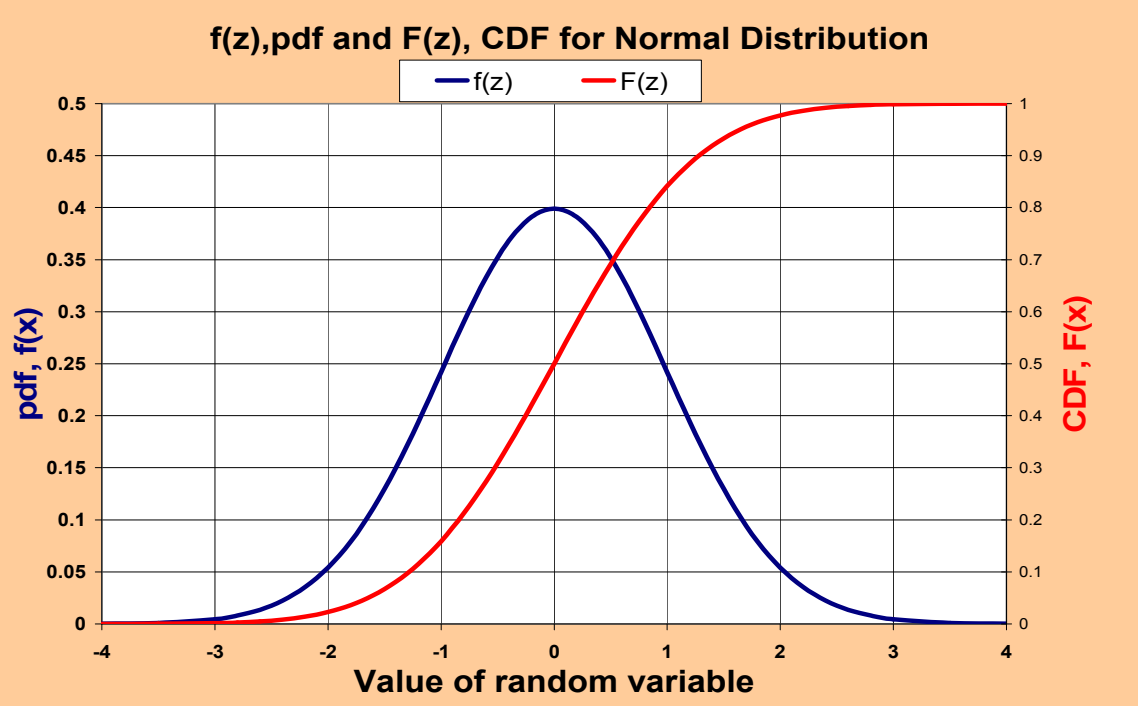
The value of Ku for a normal or Gaussian distribution is 3 and is used as a base against which peaked behavior of a distribution (or data) is measured. i.e. if one substitutes a normal distribution for f(x) into equation 9) and integrates one finds Ku=3. The data for example 2 gives Coef. of kurtosis = 2.43 so it is a bit less peaked than a normal distribution.

Topic 2: Distribution Functions:

There are discrete distributions such as the binomial distribution and there are continuous distributions such as the normal distribution. We will work mostly with continuous distributions. There are two types of distribution functions. The first is called the probability density function or pdf. The second is the cumulative distribution function or CDF. More detailed discussions are given in Appendix A. Examples of these two types for the familiar normal distribution are shown below.

Normal Distribution:

The single peaked "bell-shaped" curve (blue) is the pdf labeled f(z), and the "S-shaped" curve (red) is the CDF labeled F(z). The curves are related. At any given value of the



random variable, say x=2, the CDF is the area under the pdf up to that value e.g. for x=2 the CDF reads 0.975 or there is 97.5% probability of having an X value ≤ 2.
The formulas that describe these normal curves are

$$f_N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = N(\mu, \sigma^2)$$

$$F_N(x) = \int_{z=-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz = \Pr\{X \le x\}$$

The pdf, $f_N(x)$ represents the probability per unit of value of the random variable x. i.e. f(x)dx is the probability that the random variable X is within dx of the value x. The CDF being the area under $f_N(x)$ from X= -∞ up to X=x, is the cumulative probability that the random variable X is ≤ the value x. For as useful as the normal probability function may

be there is no closed form analytical expression for its CDF. It is given in tables in all statistics books and is available in Excel using the NORMDIST command. The point of some note is that the entire distribution function can be described with only 2 parameters (μ, σ). The coefficient of skewness = 0 and the coefficient of kurtosis = 3 for every normal curve. Estimating the values of the parameters for the data is called point estimation.
Several other distributions will be described below. A useful reference is CRC Standard Probability and Statistics Tables and Formulas, Student Edition, by Dan Zwillinger et al. Dan is a senior fellow at Raytheon.

Unit Normal Distribution.

An interesting property of the normal distribution is that one can transform the variable x using the transformation $Z=(X-\mu)/\sigma$ and this transformation produces a distribution given below and this shows a normal distribution with a zero mean and a variance = 1. Thus any normal distribution can be transformed to a unit normal distribution. In statistics books in the appendices you will find values only for unit normal distributions. The symbols ϕ and Φ have traditionally been used to signify a unit normal pdf and CDF respectively.

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \equiv \phi(z),$$
$$F(z) = F((x-\mu)/\sigma) \equiv \Phi(z)$$

Central Limit Theorem.

The reason for introducing the normal distribution so soon in this note is because we wish to introduce the most important theorem in statistics, called the central limit theorem. If one draws a sample of size n from a population whose distribution, g(x), can be fairly arbitrary in shape, and computes the average of this sample, and if one repeats this random sampling again and again, the distribution of these sample averages can be shown to approach a normal distribution. The mean of this distribution of means, is the mean of the population from which the samples were drawn and the standard deviation of the means equals the standard deviation of the population divided by the square root of the sample size. This is called the Central Limit Theorem (CLT). Note that the population distribution can be arbitrary and not look a thing like a normal distribution never-the-less the distribution of mean values from samples drawn from this arbitrary population will be distributed as a normal distribution in the limit that n is large. How large? A sample size $n > 10$ is adequate if population is unimodal w/o too much skewness and usually $n>30$ to 50 is considered a fairly good number.
The distribution of sample means, i.e. is found to be distributed as a normal distribution as the size of the sample becomes large. The distribution is given by the equation shown below.

$$f_N(\bar{x}) = \frac{1}{\sigma\sqrt{\frac{2\pi}{n}}} e^{-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2}$$

This may be the most useful theorem in applied statistics.

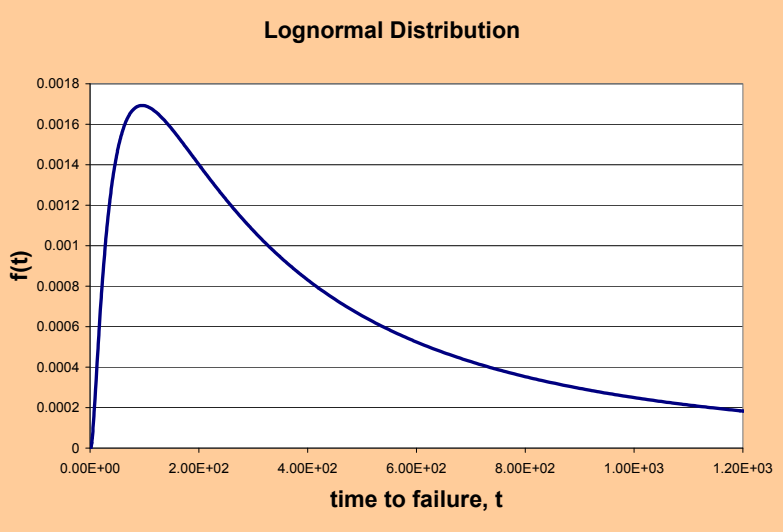
Lognormal Distribution.

The lognormal distribution is used extensively for distributions that cannot for physical reasons take on values less than zero. A typical case might be a time to failure or a time to repair a product. The lognormal distribution looks somewhat like a normal distribution.

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma y} e^{-\frac{1}{2}\left(\frac{\ln y-\theta}{\omega}\right)^2}, y > 0, \theta \equiv E\left[\ln(y)\right], \omega^2 \equiv Var\left[\ln(y)\right]$$

$$mean(y) = \mu = e^{\theta+\frac{\omega^2}{2}}, Var(y) = \sigma^2 = \mu^2\left(e^{\omega^2} - 1\right)$$

A typical graph of the pdf (θ=6,ω=1.2) is shown below.

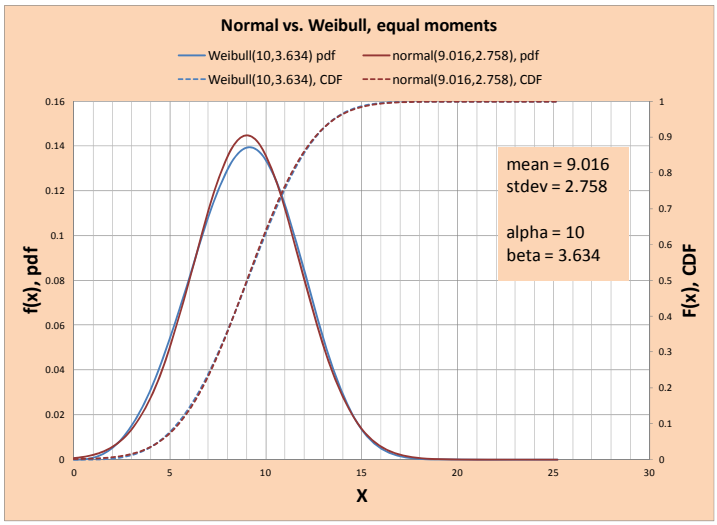


A reference to real data comes from "Lognormal Distribution for Modeling Quality Data When the Mean is Near Zero," in Journal of Quality Technology, 1990, pp 105-110.

Lognormal Distribution mean = 829, stdev = 1487.

The Weibull Distribution.

This distribution is very useful as it can take on many forms. It can have either 3 parameters or more commonly two parameters. It is used extensively in reliability work as the distribution that gives the time to failure of a part, subsystem or an entire complex system. For a reasonable range of mean values the Weibull can be made to look very much like the normal distribution as can be seen in the figure below.



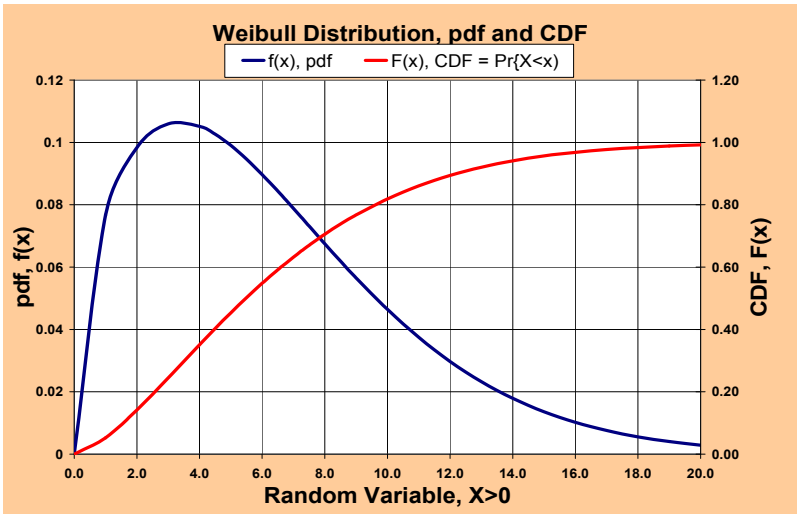
The formulas for the Weibull are given below:

Two parameter:

$$f_W(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}}$$

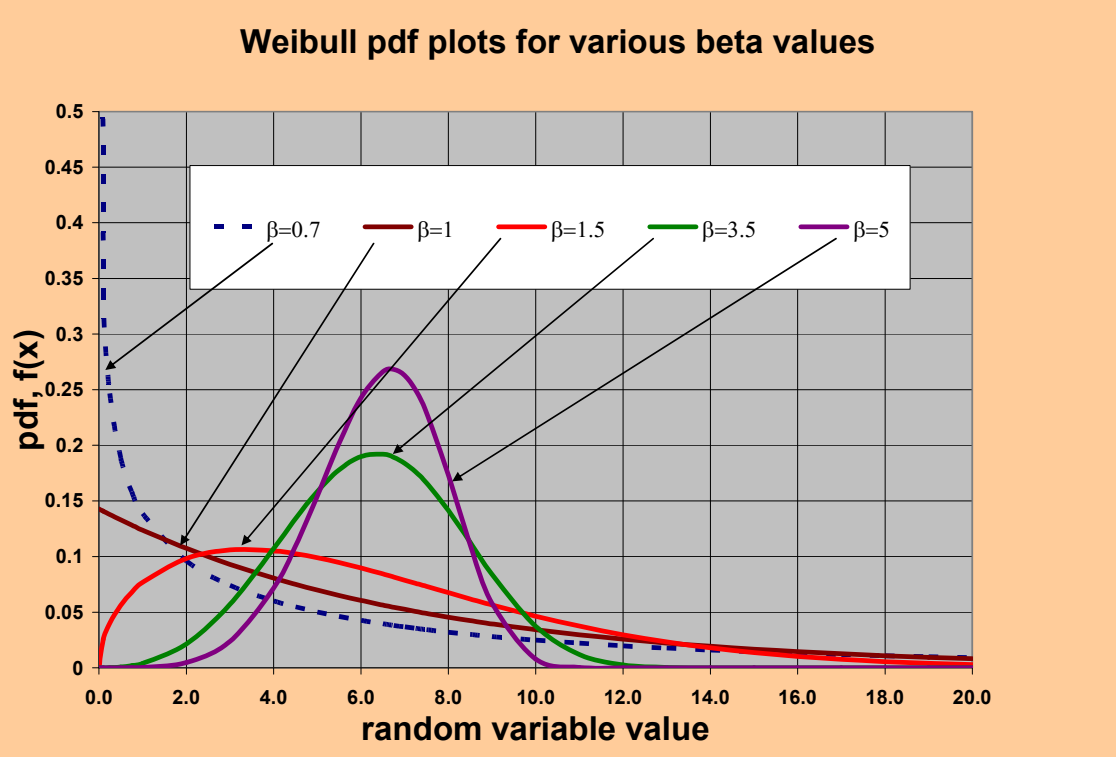
$$F_W(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\beta}} = \Pr\{T \le t\}$$

The parameter α is called the characteristic life in reliability work but more generally is called a scale parameter. The parameter β is the so-called shape parameter. In the three parameter versions one simple replaces t by t-γ, where γ = location parameter and the pdf and CDF are only defined for t ≥ γ. The shapes of these



curves are shown above. This chart shows a Weibull with $\alpha = 7$ and $\beta = 1.5$. For other parameter values look at the pdf curves below.

The special case $\beta=1$ is called the exponential distribution and is the distribution that is assumed to give the time to failure for most electronic components in a missile system. It forms the basis for the old MIL-HDBK-217F and several commercial codes. More will be made of this later.



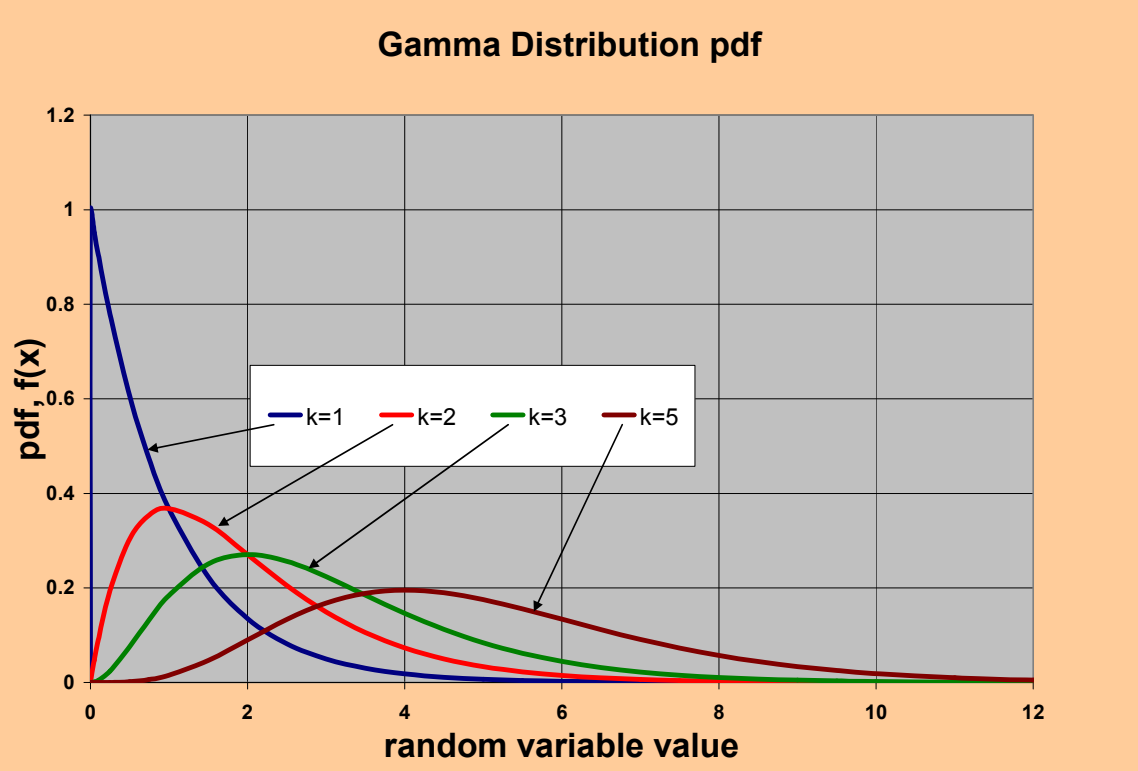
Gamma Distribution:

Another useful distribution is the gamma distribution and its pdf and CDF are given below.

$$f_{\Lambda}(x) = \frac{\lambda(\lambda x)^{k-1}}{\Gamma(k)} e^{-\lambda x}$$

$$F_{\Lambda}(x) = \int_{y=0}^{x} \frac{\lambda(\lambda y)^{k-1}}{\Gamma(k)} e^{-\lambda y} dy$$

The shapes of the pdf curves are shown below for various values of k, the shape parameter, and a fixed value of $\lambda = 1.0$. Again k=1 gives the exponential distribution.



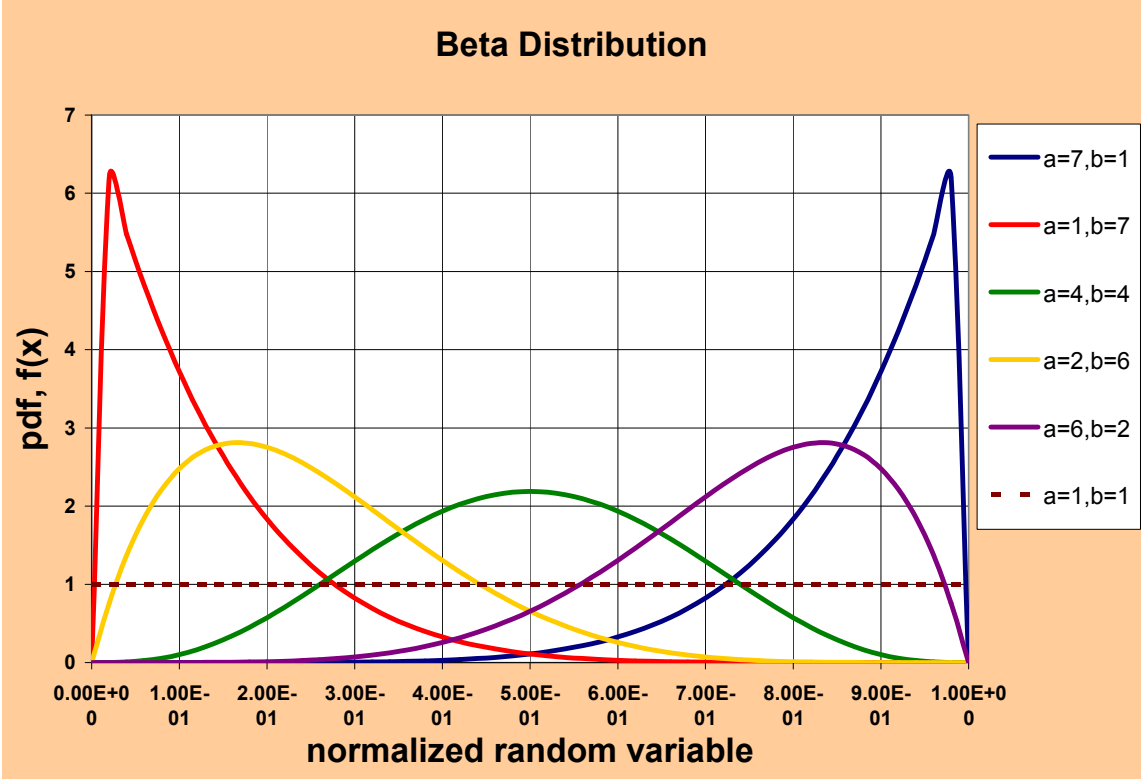
The distributions defined so far are called unbounded distributions as the random variable can take on values that can extend to infinity in one or more directions. A useful bounded distribution is called the beta distribution and is given below

Beta Distribution (0≤x≤1)

$$f_\beta(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, B(\alpha,\beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$F_\beta(x) = \int_{y=0}^{x} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} dy$$

If the interval of interest is a < z <b one simply transforms using x = (z-a)/(b-a) and the limits become (0,1) instead of (a,b).



When $\alpha = \beta$ one has a symmetric distribution. $\alpha=\beta=1$ produces a uniform distribution

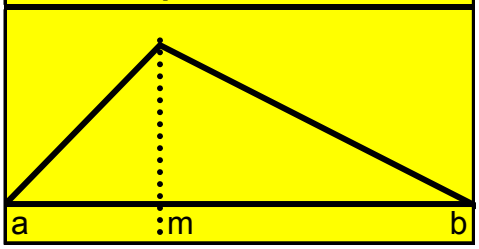
The question now becomes "are there 3 parameter distributions other than adding a location parameter to the two parameter distributions such as the gamma or Weibull distributions?" The answer is yes. The most widely used is the triangle distribution where one specifies the minimum value, the maximum value and the value at which the triangle peaks. There are also 4 parameter and 5-parameter distributions. The Johnson family of distributions has four parameters and one procedure for determining those four parameters is to match the analytical values for the 4 moments to the values of the four moments determined from the data itself. At RMS is done using a tool called JFit. Other 4-parameter distributions include the generalized lambda and generalized kappa distributions.

Triangle Distribution.

$$f_{\Delta}(x)=\begin{cases}\dfrac{2}{b-a}\left(\dfrac{x-a}{m-a}\right), a =< X =< m\\ \dfrac{2}{b-a}\left(\dfrac{b-x}{b-m}\right), b >= X >= m\end{cases}$$

$$F_{\Delta}(x)=\begin{cases}0, x<a\\ \dfrac{(x-a)^2}{(b-a)(m-a)}, a\le x\le m\\ 1-\left(\dfrac{(b-x)^2}{(b-a)(b-m)}\right), m\le x\le b\\ 1, x>b\end{cases}$$

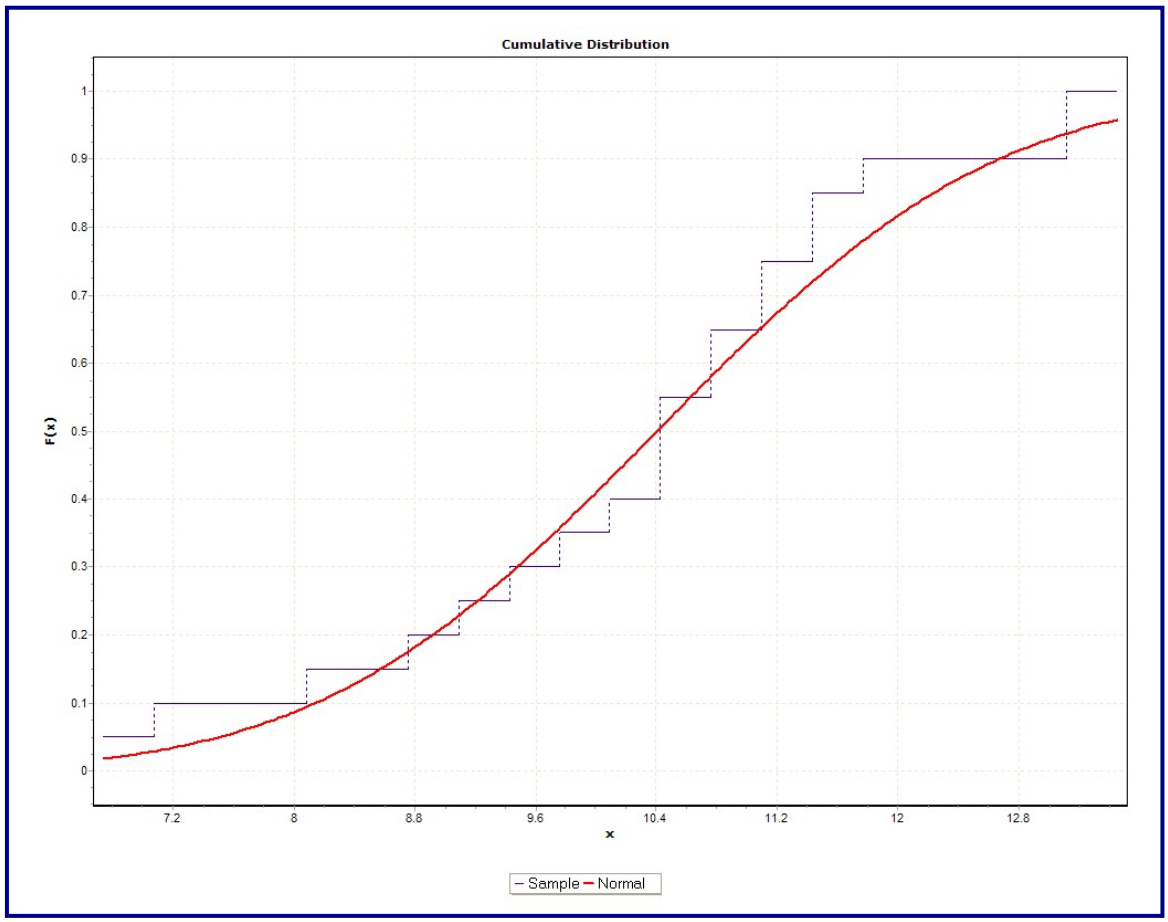
a= lower limit
b = upper limit
m = most probable or mode



In many simulation situations one can use a triangle distribution if there is no data available to construct a more suitable distribution.

Empirical Distribution (EDF).

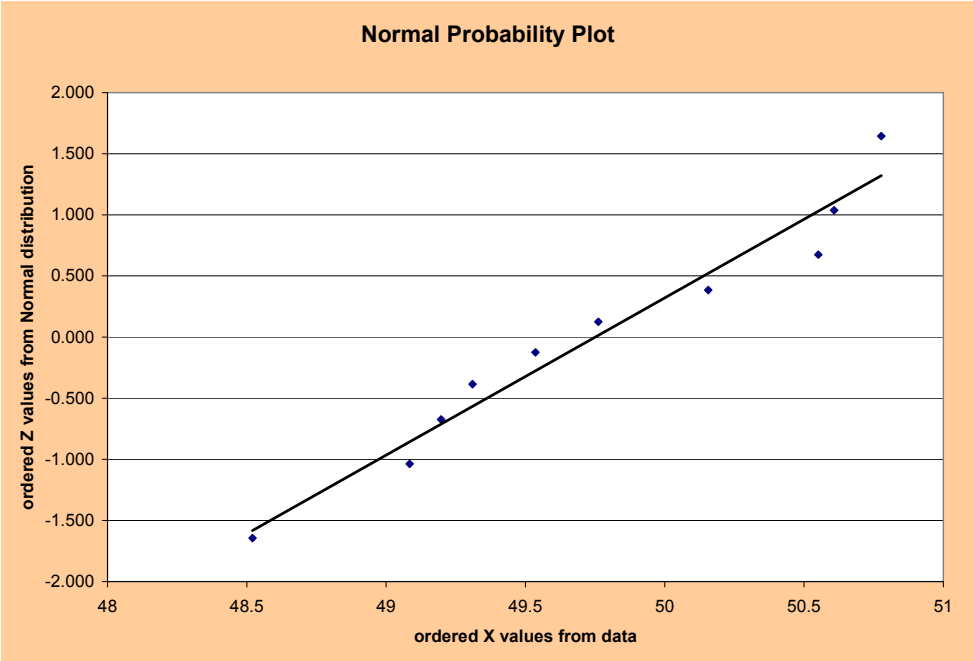
Many times we wish to construct a non-parametric distribution from the data. Such a distribution is called an empirical distribution. Consider the case in which one has n data values $\{x_1, x_2, \ldots, x_n\}$.. The EDF is produced by Plotting $F(x_j) = j/n$ vs. x_j. See the plot below.



The set of step functions represents the EDF and the red continuous curve is the CDF for a normal distribution that has been matched (with some measure of goodness-of-fit) to the data. There are 20 data points. The program EasyFit™ was used to generate this plot.

j	x(j)=data	(j-0.5)/n	$z(j)=F^{-1}((j-0.5)/n)$
1	48.5203	0.05	-1.645
2	49.08449	0.15	-1.036
3	49.19733	0.25	-0.674
4	49.31017	0.35	-0.385
5	49.53585	0.45	-0.126
6	49.76152	0.55	0.126
7	50.15645	0.65	0.385
8	50.55139	0.75	0.674
9	50.60781	0.85	1.036
10	50.77706	0.95	1.645

To generate the plot list the data values from smallest to largest and set up the following table, The plot column 4 vs column 2. The result is shown on the next graph.



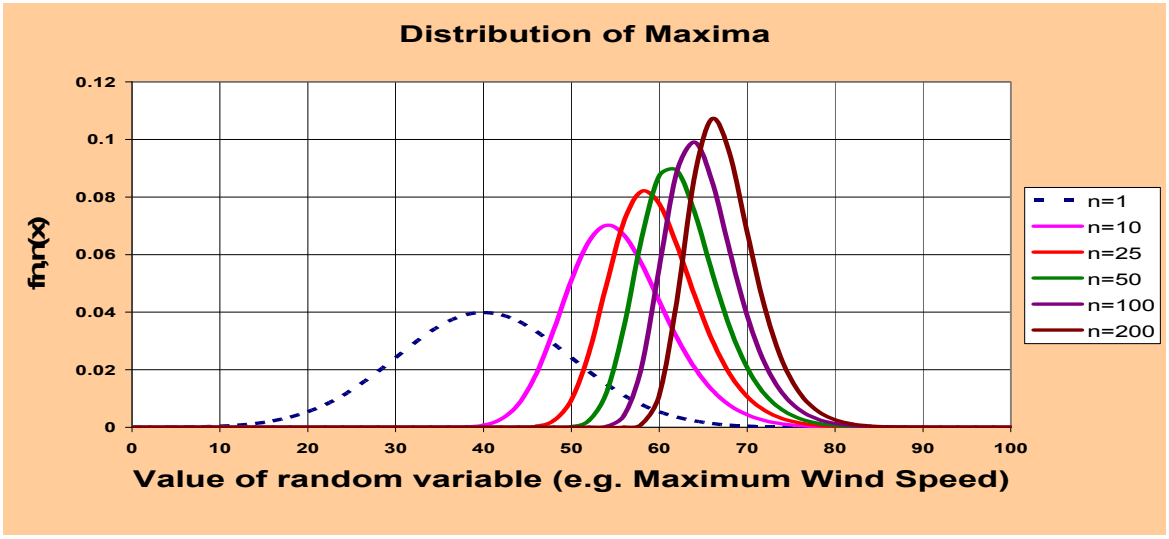
Probability Plot assuming a normal distribution for the data. Blue dots are actual data points.

Distribution of Maxima. (Largest Order Statistic) (advanced topic)

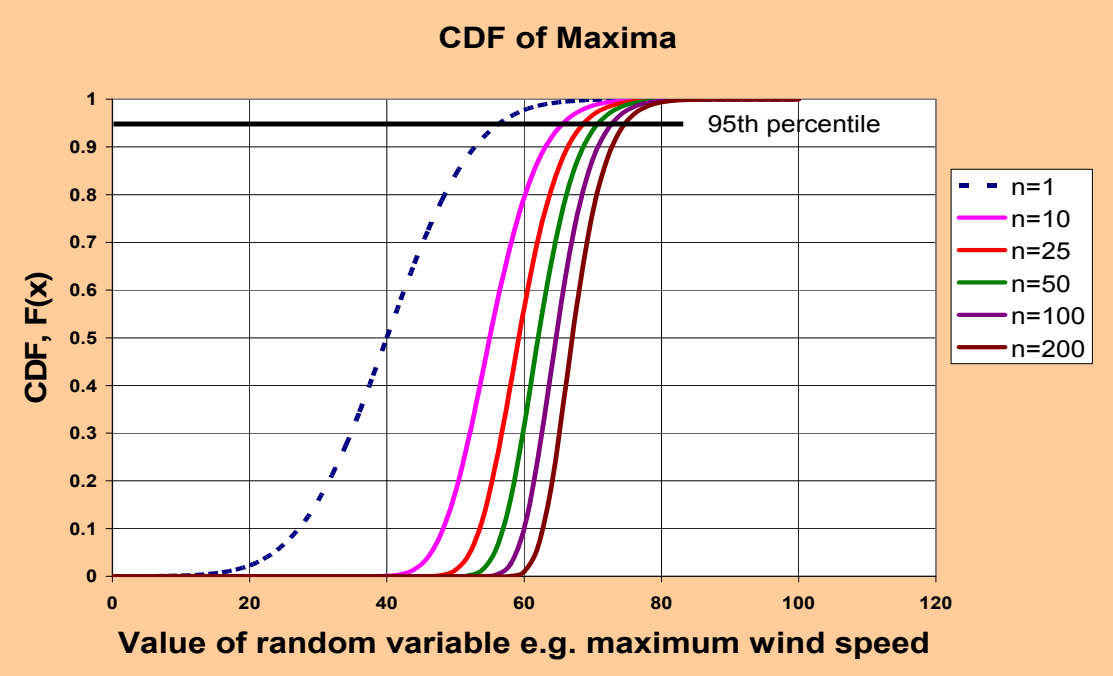
The maximum value of a given random variable, say the maximum wind speed in any given year over a period of n years, has a distribution just as one might plot the distribution of the mean wind speed due to all storms in a season for n seasons. The CDF and pdf for the maximum value is determined by

$$F_{n,n}(X) = \left[F(X)\right]^n$$

$$f_{n,n}(X) = n\left[F(X)\right]^{n-1} f(x)$$



Where f(x) is the distribution of wind speeds, X, for a season and is presumed to be the same type of distribution based on historical data. If there were 30 years of historical data then this data would be used to determine f(x) and F(x). Suppose we are interested in the distribution of maximum wind speeds over the next 25 years. Then n=25 would be used in the above formulas. The plot of the pdf is shown below for the case of f(x) being a normal distribution with mean=40 and stdev=10. A plot of the CDF is shown below along with a line showing the 95th percentile values for X for various values of n.



Distribution of Minima. (Smallest Order Statistic)(advanced topic)

Similarly one might be interested in the distribution of the minima over n years where historical data has shown that X is distributed as f(x) with a CDF, F(x). The distribution of the lowest (minimum) value of X is given by,

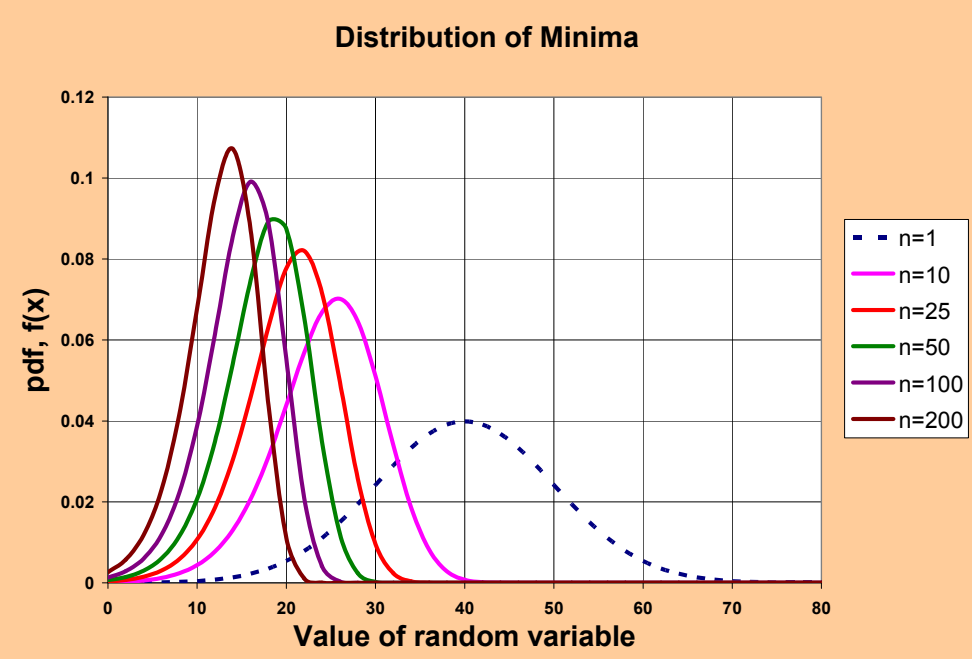
$$F_{1,n}(X) = 1 - \left[1 - F(X)\right]^n$$

$$f_{1,n}(X) = n\left[1 - F(X)\right]^{n-1} f(x)$$

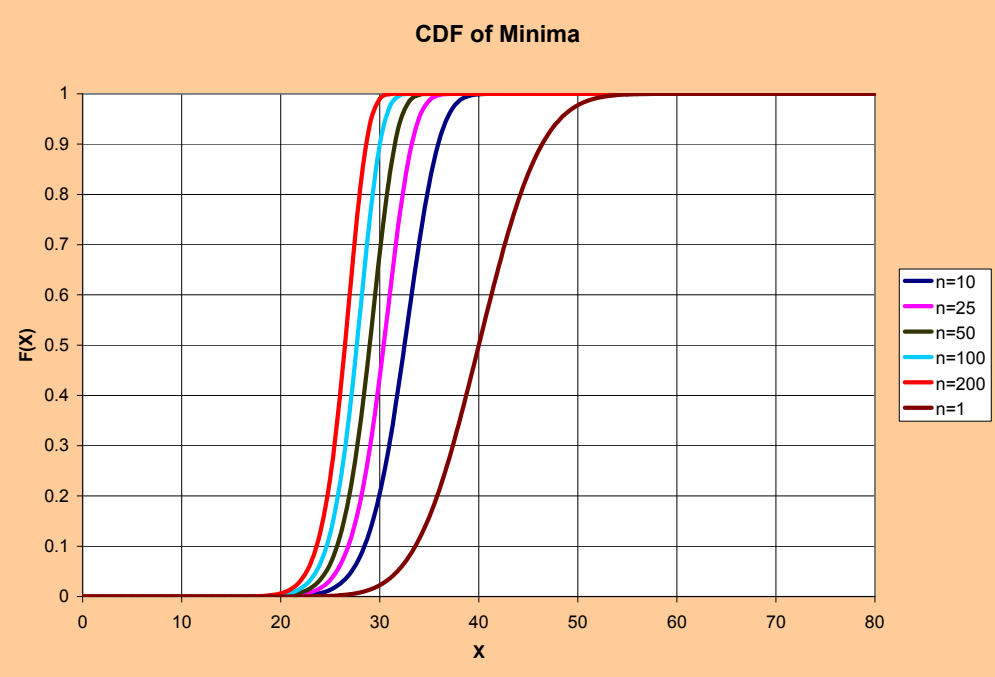
A plot of the pdf is shown below. This distribution is used to predict the reliability of a series of components Let X represent time. R(x)= reliability =1-F(x) therefore, f(x) = -dR/dx and

$$R_{1,n}(X) = \left[R(X)\right]^n$$

$$f_{1,n}(X) = n\left[R(X)\right]^{n-1} f(x)$$



Similarly the CDF for the minima is plotted below.



Topic 3: Confidence, Tolerance & Prediction Intervals

An interval estimate for a population parameter (e.g. the population mean = μ) is called a **confidence interval.** We cannot be certain that the interval contains the true, unknown population parameter—we only use a sample from the full population to compute the point estimate, e.g. $\bar{x}$, and the interval. However, the confidence interval is constructed so that we have high confidence that it does contain the unknown population parameter. Confidence intervals are widely used in engineering and the sciences.

A **tolerance interval** is another important type of interval estimate. For example, the chemical product viscosity data might be assumed to be normally distributed. We might like to calculate limits that bound 95% of the possible viscosity values. For a normal distribution, we know that 95% of the distribution is in the interval $\mu-1.96\sigma, \mu+1.96\sigma$. However, this is not a useful tolerance interval because the parameters μ and σ are unknown. Point estimates such as $\bar{x}$ and s can be used in the interval equation for μ and σ. However, we need to account for the potential error in each point estimate. The result is an interval of the form $\bar{x}$ -ks, $\bar{x}$ +ks, where k is an appropriate constant (that is larger than 1.96 to account for the estimation error). As for a confidence interval, it is not certain that this equation bounds 95% of the distribution, but the interval is constructed so that we have high confidence that it does. Tolerance intervals are widely used and, as we will subsequently see, they are easy to calculate for normal distributions.

Confidence and tolerance intervals bound unknown elements of a distribution. It is important to learn and appreciate the value of these intervals

A **prediction interval** provides bounds on one (or more) future observations from the population. For example, a prediction interval could be used to bound a single, new measurement of viscosity—another useful interval. With a large sample size, the prediction interval for normally distributed data tends to the tolerance interval in equation 2, but for more modest sample sizes the prediction and tolerance intervals are different. Keep the purpose of the three types of interval estimates clear:

- A confidence interval bounds population or distribution parameters (such as the mean viscosity).
- A tolerance interval bounds a selected proportion of a distribution.
- A prediction interval bounds future observations from the population or distribution.

Confidence Intervals.

In the previous section, point-estimates were calculated from the data drawn from a population of interest. If another sample were to be taken from the same population the point estimates would most likely turn out with different values. To compensate for sampling variations we use the concept of confidence intervals. A typical form for such an interval is given by,

P{lower limit ≤true value of population parameter ≤ upper limit} ≥ confidence level

Example: $\mathbf{P\{x - Z_{\alpha/2}\sigma/n^{1/2} < \mu < x + Z_{\alpha/2}\sigma/n^{1/2}\}=1-\alpha}$

This is read "the probability of the population parameter, μ, being between the upper and lower limit is greater than or equal to the confidence level (1-α), a fraction <1. This is a two sided (or two-tailed) confidence interval since we are looking for limits on both sides of the parameter. The procedure that will be used to generate the upper limit and lower limit will in some sense guarantee that the interval so calculated using a sample of data (size n) will contain the true but unknown population parameter for a percentage, (1-α)100%, of the samples chosen, as long as the sample was drawn using a random sampling technique. These percentages are called confidence levels and they are typically 0.90, 0.95, or higher when necessary.

A 95% confidence level (i.e. confidence level = 0.95) would imply the following. If I sample data from the population of interest then the prescription I use to calculate the upper and lower limits of the confidence interval will produce an interval that will include the true (unknown) population parameter in 95% of the samples that are possible to draw. Therefore 5% of the possible samples would not include the parameter. This may sound a bit obtuse. There is a temptation to simply say that a 95% confidence interval, once calculated, has only a 5% chance of not including the true (unknown) parameter of interest. These two statements sound the same but are not. This may be a subtlety not worth emphasizing the first time one explores statistics but it has been included to provide the "real" interpretation for those who are stickler's for the truth.

Note: The second statement, the one we would like to have be correct, IS correct if we use Bayesian methods to create what is called a credibility interval. This Bayesian stuff will be addressed later.

Let's try an example. Remember once a sample of data is taken and the interval calculated then either the true population parameter is inside that interval or it is not! How are intervals (upper& lower limits) calculated?

Confidence Interval for the Mean.

To calculate the confidence interval one needs to know the distribution of the sampling statistic. In the case of the mean we know the sampling distribution for the mean is the normal distribution when n is large. This is due to the central limit theorem. If we were to estimate some other population parameter such as the variance we need to find the sampling distribution for s^2.

When a sample size is large (n>30) and the population standard deviation is known, we make use of the Central Limit Theorem which says that $\bar{X}$ is distributed approximately as a normal distribution. Remember if we do not know how $\bar{X}$ is distributed then we cannot proceed further. The discussion of this is well outlined in Reference [1] by Montgomery et al.

The interval is given by $\left(\bar{X} - Z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + Z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right)$ however we many times write this interval in a more meaningful way as a probability statement.

6) $$P\left\{\bar{X} - Z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu_X \le \bar{X} + Z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right\} \ge 1-\alpha$$

This is read as "the probability that the population mean (unknown & unknowable) is between the lower and upper limits shown above is greater than or equal to 1-α" where the confidence level is defined as (1-α)100%. The reason for defining confidence level as 1-α, where α is called the significance level, is mostly historical but it is the convention used in all statistics texts. An example for a 95% confidence level α = 0.05. The value α represents the probability that the mean of the population would lie outside the calculated upper and lower limits. $Z_{1-\alpha/2}$ is the number of standard deviations from the mean that we are adding to or subtracting from the mean to have a confidence level of 1-α, half above the mean and half below the mean, and $\sigma/\sqrt{n}$ is the standard deviation of the mean value, $\bar{X}$. This ratio is called the standard error of the mean, SEM or $\sigma_{\bar{X}}$. The random variable $\bar{X}$ is the sample mean calculated from the sample of size n. If the CL is 95% then $Z_{0.975}$=1.96 (almost 2 standard deviations) and we would be 95% confident that the population mean would lie between the sample mean plus 1.96 times $\sigma_{\bar{X}}$ and the sample mean minus 1.96 times $\sigma_{\bar{X}}$. Some typical Z-values are shown below for the usual array of α values. One can use the excel function NORMSINV(probability) to find these values. For a two-sided confidence interval we look for Z-values evaluated at 1−α/2 probability.

α	0.01	0.02	0.05	0.10	0.20
$Z_{1-\alpha/2}$	2.576	2.326	1.960	1.645	1.282

Example 4: Again taking the data from the table in Example 2 and assuming the standard deviation of the population = 0.8, we obtain $\bar{X} = 2.15, \sigma = 0.8, n = 25$ and calculate the 95% confidence interval to be $\left(2.15 - 1.96\frac{0.8}{\sqrt{25}}, 2.15 + 1.96\frac{0.8}{\sqrt{25}}\right) = (1.836, 2.464)$. The value 1.96 can be found using the Excel™ function NORMSINV(0.975). What does one obtain using NORMSINV(0.025)?

If we do not know the standard deviation of the population but instead have to use a sample standard deviation calculated from a small sample, then we use a t-value from the student t-distribution instead of the standard normal Z-values. The t-distribution [Ref 1] looks much like the normal distribution but has "fatter tails than a normal distribution and also depends on the sample size n. For a sample size of n=25 the values of t for various alpha values are shown below.

α	0.01	0.02	0.05	0.10	0.20
$t_{\alpha/2,24}$	2.797	2.492	2.064	1.711	1.318
$Z_{1-\alpha/2}$	2.576	2.326	1.960	1.645	1.281

Note that we use t-values evaluated at n-1=25-1=24 degrees of freedom.

Example 5: Using the same information as in Example 4 but instead using the sample standard deviation s = 0.8205 on obtains the 95% confidence interval to be, $\left(2.15-2.064\frac{0.82}{\sqrt{25}}, 2.15+2.064\frac{0.82}{\sqrt{25}}\right)=(1.812, 2.488)$. The value 2.15 can be found using the Excel™ function TINV (0.05, 24). A peculiarity of Excel™ is that for the t-distribution only, the inverse values are given assuming a two tailed test from the start. i.e. internal to Excel it calculates the 0.025 value and gives that answer for TINV. The t-distribution is parameterized by the so-called degrees of freedom (dof) which for this interval calculation is given by the sample size – 1. The one is subtracted to account for the fact that one degree of freedom was used to calculate the sample variance itself.

Note that *due to the uncertainty in the standard deviation the confidence interval is larger* for the same confidence level. When the sample size n >30 the t-distribution is well approximated by the normal distribution. For n>30 one can still use the Z-values which relieves some complications. One assumption behind the t-distribution is that the sampling distribution of mean values is normally distributed, which by the central limit theorem is true in the limit that n is large.

Confidence Interval for the Variance.

Unlike the confidence interval for the mean, the confidence interval for the variance is not symmetrical about the point estimate. The variance is distributed as a chi-square random variable.

The formula for calculating the confidence interval for the variance is given by the probability statement,

7) $$P\left\{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right\} \ge 1-\alpha$$

The chi-square distribution is given in tables or can be computed in excel using the CHIINV(probability,dof) function. This CI assumes that the variable itself (i.e. x) is distributed normally or at least comes from a bell-shaped distribution. The dof = n-1 to account for calculation of sample variance.

Example 6: Using data from Example 2 where we found that s^2=0.673, the CI calculation proceeds as follows; $\left(\frac{(24)0.673}{36.42} \le \sigma^2 \le \frac{(24)0.673}{13.85}\right)=\left(0.443 \le \sigma^2 \le 1.166\right)$ for a 90% confidence interval for the population variance. The Excel™ function CHIINV(0.05,24)=36.42 and CHIINV(.95,24)=13.85.

Note this is a broad range of values. It is not immediately apparent that an increased sample size helps. In fact, if the sample standard deviation is the same then having that value using a larger sample makes the confidence interval even broader. If however, in the real case, where increasing the sample size decreases the sample variance then the reduced sample variance outweighs the (n-1) in chi-square inverse and the confidence interval becomes smaller.

Confidence Interval for Fraction Nonconforming (normal distribution).

Any estimate of fraction nonconforming is distributed as a random variable with associated confidence limits. If the sample size is large enough (n>30) then one can use the normal distribution. Let Z=(x-m)/s where x is any value of interest, such as a specification limit.

Example 7: Suppose the UCL for a process is 90, the process mean is 88.5 and the standard deviation is 1.23. Both of these values were estimated from a relatively large sample n=150. Using the formula gives Z = (90-88.5)/1.23 = 1.22. A value of 1.22 implies a probability P(Z>1.22)=0.111 which says the nonconforming fraction is 11.1%.

In actuality the estimates of the mean and standard deviation are point estimates taken from the sample data and using the z formula may seem somewhat dissatisfying as there is uncertainty in the parameters of the distribution itself. Fortunately there are some approximations that can lend more accuracy to the solution to this dilemma.

Weingarten [J.Qual.Tech., 14, #4 (Oct.,1982),pp207-210] has developed formulas that are more accurate e.g.

8)
$$Z_{UCL} = \left(\frac{x-\bar{x}}{s}\right) - Z_{1-\alpha/2}\sqrt{\frac{1}{n}+\frac{1}{2n}\left(\frac{x-\bar{x}}{s}\right)^2}$$
$$Z_{LCL} = \left(\frac{x-\bar{x}}{s}\right) + Z_{1-\alpha/2}\sqrt{\frac{1}{n}+\frac{1}{2n}\left(\frac{x-\bar{x}}{s}\right)^2}$$

Again specifying an interest in an upper spec limit of x=90 and using the process mean value of 88.5 and sample standard deviation of 1.23 we find $\left(\frac{x-\bar{x}}{s}\right) = 1.22$. Now defining Z_{UCL}=Z-value for upper confidence limit, Z_{LCL} = Z-value for lower confidence limit. Desiring a 95% confidence level, one finds Z_{UCL}=1.009, Z_{LCL}=1.431. For a moment, these values may seem incorrect or reversed from what they should be. The formulas are correct. If I use Z=1.009 and go to the standard normal tables I find P(Z>1.009) = .156 or the fraction of nonconforming is 15.6%. This is an upper confidence limit about the spec limit x=90. Using the lower confidence limit of Z=1.431, I find P(Z>1.431) = 0.0762 or 7.62% of the units are nonconforming. This is a lower confidence limit about the spec limit x=90. Thus (.0762 ≤ P(x>90) ≤.156) with a confidence of 95%. This is a subtle point but is very important. Obviously if the sample size is very large then little error occurs in using the point estimate (e.g. P(x>90)=P(Z>1.22)=0.111) for fraction nonconforming.

Confidence Interval for Proportion:

The point estimate for a proportion was given before as p = number of occurrences / total sample size (n) = x/n. When the sample size, n, is large and n*p >5 and n*(1-p)>5 we

can use the normal distribution to calculate confidence intervals. The formula for the interval is shown below.

9) $$\left(\overline{p}-Z_{1-\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\le\pi\le\overline{p}+Z_{1-\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right)$$

where $\overline{p}$ is the proportion calculated from the data. Note: The expected value for the sample variance, $\overline{p}(1-\overline{p})/n$.

Example 8: p = proportion, n = sample size, Z = normal distribution z-value dependent of confidence level. Using values from Example 3 where p = 16/700 =0.023 and n=700, for a 90% two sided confidence interval $Z_{0.95}$=1.645 and the confidence interval for π becomes,

$$\left(.023-1.645\sqrt{\frac{.023(.977)}{700}}\le\pi\le.023+1.645\sqrt{\frac{.023(.977)}{700}}\right)=(0.014\le\pi\le.032)$$

which is the 90% confidence interval for the population proportion π.

Small Sample Size Confidence Intervals.

If the sample size is small to moderate there is more accuracy gained in using the following formulas for confidence limits on proportions or fraction nonconforming.

10) $$p_L\equiv\frac{\left(p-\frac{1}{2n}+\frac{Z_{1-\alpha/2}^2}{2n}\right)-Z_{1-\alpha/2}\sqrt{\frac{1}{n}\left(p-\frac{1}{2n}\right)\left(1-p+\frac{1}{2n}\right)+\left(\frac{Z_{1-\alpha/2}}{2n}\right)^2}}{\left(1+\frac{Z_{1-\alpha/2}^2}{n}\right)}$$

11) $$p_U\equiv\frac{\left(p+\frac{1}{2n}+\frac{Z_{1-\alpha/2}^2}{2n}\right)+Z_{1-\alpha/2}\sqrt{\frac{1}{n}\left(p+\frac{1}{2n}\right)\left(1-p-\frac{1}{2n}\right)+\left(\frac{Z_{1-\alpha/2}}{2n}\right)^2}}{\left(1+\frac{Z_{1-\alpha/2}^2}{n}\right)}$$

where r = number of occurrences, n = sample size, p ≡ r/n, Z = calculated from unit normal distribution

Example 9: take n=50, r=7, $Z_{1-\alpha/2}$=1.96 (i.e. 95% two sided confidence interval),
$\Phi_L=0.063,\Phi_U=0.274$
This leads to the interval $(0.063\le\pi\le0.274)$ at a 95% confidence level and the point estimate for π ~ p = r/n = 7/50 = 0.14.

Continuity Correction Factor

In actuality the distribution of the proportion estimator, $\overline{p}$, is the distribution of X, the number of events of interest which is then scaled by dividing by n, the sample size. X is distributed as a series Bernoulli variable experiments (X=1 is success, X=0 is failure) and is therefore governed by the binomial distribution. This is a discrete distribution and for larger sample sizes (n>30) it is approximated by the normal distribution which is

continuous. It has been found over the years that one can obtain better numerical results for approximate confidence intervals by using the following simple rules.

1) for cases $P\{X \le x\}$ use $P\{X \le x + 0.5\} = P\{Z \le (x+0.5-np)/(np(1-p))^{1/2}\}$
2) for cases $P\{X \ge x\}$ use $P\{X \ge x - 0.5\} = P\{Z \ge (x-0.5-np)/(np(1-p))^{1/2}\}$

The confidence Interval then becomes a somewhat simplified version of equations 10) and 11) shown above.

$$\left(\overline{p} - \frac{1}{2n} - Z_{1-\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \le \pi \le \overline{p} + \frac{1}{2n} + Z_{1-\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \right)$$

This confidence interval is said to have the continuity corrections applied.

Confidence Interval for Poisson Distributed Data.

When dealing with counting types of data such as the number of defective parts in a sample one typically is dealing with a Poisson process or Poisson distributed data. To calculate a confidence interval for the average number of defects present in a population that has been sampled with a sample size n, one can use the chi square distribution as follows: To calculate the upper confidence limit first calculate the number of degrees of freedom dof = 2*(r+1). For the lower limit dof=2*r. When calculating the 90% two-sided confidence interval use the 0.10/2 and the 1-0.10/2 probability values to find the chi square values. The confidence interval is found to be

12) $$\left(\frac{\chi^2_{1-\alpha/2,2r}}{2} \le \overline{r} \le \frac{\chi^2_{\alpha/2,2(r+1)}}{2} \right)$$

Example 10: Suppose in a complex assembly one found r=13 nonconformances. Find the 90% confidence interval for the number of nonconformances.

$$UCL = \frac{\chi^2_{1-\alpha/2,2(r+1)}}{2} = \frac{\chi^2_{.05,28}}{2} = 41.337/2 = 20.67$$

$$LCL = \frac{\chi^2_{1-\alpha/2,2r}}{2} = \frac{\chi^2_{.95,26}}{2} = 15.379/2 = 7.69$$

The confidence interval becomes $\left(7.69 \le \overline{r} \le 20.67\right)$

Nonparametric Confidence Intervals:

One can calculate confidence intervals for fraction nonconforming units without having to presuppose a normal or any other distribution. The advantage of this nonparametric approach is that it relieves us of trying to substantiate a population distribution. The penalty is that the resultant interval is larger than the parametric estimates.

If there are n tests and r units are found to fail then p = r/n is an estimate of the failure probability and 1-r/n is an estimate of the reliability at the end of the test. Since failure of an operating unit is considered to be a binary event (reliability taking into account degradation is the topic of another paper) and thus can be represented by a Bernoulli trial for each of the n units under test, one can ask the following questions. What is the upper

limit on the population proportion of failures, p_U, that will allow for r or fewer failures with a $\alpha/2$ probability. In equational form we have

$$\sum_{i=0}^{r} {}_nC_r p_U^i \left(1-p_U\right)^{n-i} = \alpha/2 .$$

Similarly, if we ask what is the lowest value for the proportion, p_L, which would allow for there to be r or greater failures to occur with probability $\alpha/2$, one would obtain

$$\sum_{i=r}^{n} {}_nC_r p_L^i \left(1-p_L\right)^{n-i} = \alpha/2, or, \sum_{i=0}^{r-1} {}_nC_r p_L^i \left(1-p_L\right)^{n-i} = 1-\alpha/2 .$$

These equations can be solved by iteration in Excel.

The results of such iteration are shown below.

PL 0.24	**Use Solver**	sum 0.975	1-a/2 0.975
r =	**33**	**α =**	**0.05**
n =	**100**	**1–α/2 =**	**0.975**
PU 0.43		sum 0.025	a/2 0.025

The results can also be determined using the F-distribution or the incomplete Beta distribution (see below).

Exact (non parametric) Confidence Interval for proportions.

Two-sided confidence interval the exact interval that is independent of any distribution is given by the following formulas.

$p_L = 1 - \text{Beta}(1-\alpha/2, n-r, r+1)$

$p_U = 1 - \text{Beta}(\alpha/2, n-r+1, r)$

n = number of tests, r = number of failures, $1-\alpha$ = confidence level

Topic #4: Hypothesis Testing

Testing a sample mean versus a hypothesized mean when □ is known.

When the standard deviation is known (or can be assumed), the distribution of sample averages drawn from a population will be distributed normally with a standard deviation of $\sigma/\sqrt{n}$. This is many times called the standard error of the mean or SEM. This result comes from the Central Limit Theorem.

Using this information we can develop and test hypotheses to determine the location of the mean of a population μ with respect to a hypothesized mean value μ_0. The hypotheses that can be tested are;

Two sided test.
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

The null hypothesis, H_0, is that the population mean is equal to the hypothesized mean. The alternative hypothesis H_1 is that the two means are not equal.

REMEMBER THE FOLLOWING!

Rule 1: The *statement we wish to prove is always placed in the alternate hypothesis*, that is, you want to see if the data you have collected from the population of interest will allow you to reject the null hypothesis in favor of the alternate hypothesis. The *hypotheses statements are always stated in terms of the parameters of the population*, the hypotheses are not statement about statistics (such as comparing sample mean values or sample standard deviations).

One-sided test.
$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$
or
$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

Rule 2: The *equality sign is always placed in the statement of the null hypothesis* (by convention not necessity). This may seem confusing at first but simply follow these two rules and you cannot go wrong. (Note: Textbooks differ in the manner in which they set up hypothesis testing and some do not follow this equal sign placement convention, so read carefully.)

How confident do you want to be that H_0 is consistent with the data? Say you wish to be 95% confident, then we need to set up some kind of statistical test that will give us this level of confidence.

Let $Z^* \equiv \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ where we are using the mean of the sample data, $\bar{X}$, as an estimator of the population mean, μ_0. The standard deviation is assumed known and n is the sample size. The variable Z is called a statistic because it contains only data and known constants. The mean of the population, μ_0, is not a statistic; it is a known parameter

(hypothesized) that characterizes a population. We want to know if data comes from a population whose mean is μ_0.
The test will be formulated by establishing a critical value for the random variable being tested, i.e. Z. In this problem, the critical Z-value, labeled Z_{crit} is presumed to come from a unit normal distribution, N(0,1). If we did not know how Z was distributed we could not perform this test!
How is the value for Z_{crit} determined? Remember from previous examples that if Z=1.96 this indicates that P(Z<1.96) = 0.975, so for the sake of argument pick 1.96 as the critical Z-value. Then calculate a test statistic Z* and if that test statistic is greater than 1.96 then the data from which I calculated the test statistic is from a population who's mean is far away (at least 1.96 standard deviations of the mean) from the hypothesized mean, μ_0. It is so far away that the probability of the data coming from a population whose mean is μ_0 is only 5%. Simply put, this data it is not likely to have come from a population whose mean was equal to the hypothesized mean if Z* is greater than Z_{crit}.

Example 11: To ensure a new production line has the capacity to meet demand, it is necessary to maintain the holding time in the bulk tanks to an average of 180 minutes. A similar process is line is being used at another location. Using the original process as a reference, test whether or not the holding tank in the new line averages 180 minutes at the 5% level of risk. The sample size is 10. Assume the process in the new line has a standard deviation of 3 minutes.

The null and alternate hypotheses are $\begin{matrix} H_0 : \mu = 180 \\ H_1 : \mu \neq 180 \end{matrix}$. This is a two-sided test in that we wish to see if the new line is either above or below the mean of the known reference line. The critical Z value is $Z_{0.975}$=1.96 since we wish to take only a 5% risk i.e. have a 95% confidence level. Assume the data from the new line is given below

	185	187	187	185	184	185	184	181	180	188
mean =	184.6									
test Z =	4.85									

It can clearly be seen that Z=(184.6-180)/3 and |Z| = 4.85 which is > Z_{crit} = 1.96. Therefore we must reject the null hypothesis and say that the new production line does not have the same mean as the reference line.
Technically we cannot say that the new line has a greater average because we performed a two-sided test. We need to perform a one sided test. If we did a one-sided test with $H_0 : \mu \leq 180, H_1 : \mu > 180$ then $Z_{crit} = Z_{0.950}$ = 1.645 and since Z=4.85 > 1.645 we would reject H_0 at this 95% confidence level and then we can technically say that the new line has a mean that is greater than the mean of the reference line.

Testing Sample Mean Versus Hypothesized Mean When Standard Deviation Is Estimated From Sample Data.

When σ is unknown and must be estimated from the sample data we would expect there to be more uncertainty than when s is known. This uncertainty will manifest itself in our thinking and tests. Thus, we will use the t-distribution (sometimes called the student-t distribution) to perform our hypothesis tests instead of the normal distribution. The t

distribution assumes that the random variable being examined comes from a normal distribution but the test is still reasonably accurate if the population, from which the sample is drawn, is symmetric and somewhat "bell-shaped." The t statistics is defined for this problem by

$$t \equiv \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

The t distribution looks much like the normal distribution but has fatter tails. In fact, the t-distribution goes over to the normal distribution for large sample sizes.

Example 12: Using the data from the previous example let us test the null hypothesis that the new line has a mean holding time of 183 minutes and still require a 95% level of confidence. The problem sets up as follows;

$H_0 : \mu = 183, H_1 : \mu \neq 183$

$1 - \alpha = .95, \alpha / 2 = 0.025$

$dof = n - 1 = 10 - 1 = 9$

In using the t-distribution we need to use another parameter called the degrees of freedom (dof) and for this problem the dof=n-1.

The critical value of the t-statistic, t_{crit}, for a two-sided test at α=0.05 and dof=9 is found from the tables (or using excel function TINV(a,dof)) to be $t_{0.025,9} = 2.262$. From the sample data, we find $\bar{X} = 184.6, s = 2.547$ which gives a test statistic

$$t \equiv \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{184.6 - 183}{2.547/\sqrt{10}} = 1.987 .$$

Since |t|=1.987 < 2.262 = t_{crit} we cannot reject the null hypothesis so we cannot reject the possibility that the new line has a mean holding time of 183 minutes on the basis of the data we have taken.

We are not accepting that 183 minutes is the population mean of the population from which we have sampled, we simply cannot reject that possibility. This is s subtle point but is important. We could not have rejected, on this basis of this same data, a null hypothesis that μ_0 = 184 or 185. There are an infinite number of hypothesis values that would not have been rejected on the basis of this data. The power in this test is when you can reject the null hypothesis.

Testing for Difference Between Two Populations Means - σ is Known.

The test statistic is given by

$$Z \equiv \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

where the subscripts denote the sample averages from populations 1 and 2 respectively and similarly for the variances from the populations under test except that we are assuming the variances are known.

Example 13: Using the data below test the null hypothesis that the mean of process one is less than or equal to the mean of process two, versus the alternative hypothesis that the mean of process one is greater than the mean of process two. Test at the 99% confidence level (α=0.01) and assume variance of process one is 4.3 and of process two is 2.5.

$H_0 : \mu_1 \le \mu_2$

$H_1 : \mu_1 > \mu_2$

Data

process 1	process 2
85.2	89
87.3	89.4
92.5	90.8
80.8	84.3
84.8	88.2
	88.1

Calculating the relevant information we obtain,

$\bar{X}_1 = 86.12, \bar{X}_2 = 88.3$

$n_1 = 5, n_2 = 6$

$\sigma_1^2 = 4.3, \sigma_2^2 = 2.5$

and the test statistic becomes $Z = \dfrac{88.3 - 86.12}{\sqrt{4.3/5 + 2.5/6}} = 1.93$

The critical value of Z from the normal distribution for a one-sided test at a=0.01 is found to be from tables (or Excel NORMSINV(0.99)) $Z_{crit}=Z_{0.99}=2.326$.

Since $Z=1.93 < 2.326=Z_{crit}$ we cannot reject the null hypothesis that the mean of process 1 is less than or equal to the mean of processes 2.

Testing for Difference in Two Population Means whose Standard Deviations are Unknown but Presumed Equal.

Here we will use the t-distribution as the variances are unknown and calculated from data. The test statistic is given by,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{1/n_1 + 1/n_2}}, s_p^2 \equiv \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

The variable s_p is called the pooled standard deviation and it is just a weighted average of the calculated variances from each sample. One can pool in this manner if one believes the data should come from processes that have the same variance. The appropriate dof = n_1+n_2-2 for this test.

Example 14: Using data from Example 14, we will test the hypothesis of equal means.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

at a confidence level of 90% (α=0.10). the two-sided critical t value is found to be

$t_{crit} = t_{0.05,9} = +/- 1.833$, using $\bar{X}_1 = 86.12, \bar{X}_2 = 88.3, n_1 = 5, n_2 = 6$,

and calculating the variances of both samples one obtains

$s_1^2 = 18.247, s_2^2 = 4.8$

that produces a pooled variance

$$s_p^2 = \frac{(4)(18.247)+(5)(4.8)}{5+6-2} = 10.77 .$$

Using this to evaluate the test statistic

$$t = \frac{86.12-88.30}{3.283\sqrt{1/5+1/6}} = -1.097$$

one notes that –1.833 < -1.097 < +1.833 so one cannot reject the null hypothesis and thus the two mean values cannot be claimed to be unequal based upon the given test data.

Testing for Difference Between Two Proportions:

When one uses large sample sizes such that n*p>5 and (1-p)*n>5 then one can use normal distribution theory.

Let $Z = \frac{p_1 - p_2}{s_{p_1-p_2}}, s_{p_1-p_2}^2 = \hat{p}(1-\hat{p})(1/n_1 + 1/n_2)$

And the value

$\hat{p} \equiv \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ is a weighted proportion

Example 15:

Assume one has two technologies to produce the same product. We will test the proportion nonconforming of the product from both processes. Determine if the fraction nonconforming for each process is the same at a 95% confidence level.

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

$\alpha = 0.05, \alpha/2 = 0.025$

$Z_{crit} = Z_{0.975} = \pm 1.96$

The data for each process is stated to be as follows:

Process 1: number under test n1 = 300, number of defects d1 = 12, p1 = d1/n1 = 12/300 = 0.04.

Process 2: number under test n2 = 350, number of defects d2 = 21, p2 = d2/n2 = 21/350 = 0.06.

Using this data one finds $\hat{p} = \frac{(300)(.04)+(350)(.06)}{300+350} = 0.051$ and the standard deviation

$s_{p_1-p_2} = \sqrt{.051(.949)(1/300+1/350)} = 0.017$ and the test statistics becomes

$$Z = \frac{.04-.06}{0.017} = -1.176$$

Since –1.96 < -1.176 < 1.96 we cannot reject the null hypothesis at the 95% confidence level. Note: Had we chosen the 70% confidence level then Z_{crit}= +/- 1.036 and we could reject the null hypothesis at this lower level confidence (i.e. we are willing to accept a 30% chance of being wrong).

Testing for Differences in Count Data.

Suppose we have Y_1 nonconformances in sample 1 and Y_2 nonconformances in sample 2 We wish to know if the number of nonconformances can be considered essentially equal or stated another way is the difference in number of defects between the two samples statistical significant?

When the sample sizes are unequal and large we can use the normal approximation as the Poisson distribution can be approximated by the normal for large sample sizes and small probabilities.

The test statistic is $Z \equiv \frac{n_2 Y_1 - n_1 Y_2}{\sqrt{n_1 n_2}\sqrt{Y_1 + Y_2}}$ (note the transposition of the sample sizes in the numerator). One finds a critical value from the standard normal distribution.

Example 16:

A sample of 40 units found there to be 67 nonconformances of a particular type. After a design change to aid assembly a sample of 23 units found only 31 nonconformances. Determine if the engineering change resulted in a decrease (one-sided test) at a confidence level of 90%.

$H_0 : \mu_1 \le \mu_2$

$H_1 : \mu_1 > \mu_2$

$\alpha = 0.10$

$Z_{crit} = Z_{0.90} = +1.282$

Calculating the test statistic gives $Z \equiv \frac{(32)(67)-(40)(31)}{\sqrt{(40)(32)}\sqrt{67+31}} = 2.552$.

Since Z = 2.552 > 1.282 = Z_{crit} one can reject the null hypothesis and conclude that the design process did indeed reduce the number of nonconformances. The second process comes from a population that has a lower number of nonconformances per unit.

Topic #5: Testing for Differences in Variances

There are three common tests for variances: 1) testing to determine if the variance of a population as estimated by a sample equals a hypothesized or known variance, 2) testing to determine if two variances estimated from two samples could come from the same population having equal variances, and 3) testing for the equality of 3 or more variances as normally done in an Analysis of Variance (ANOVA) procedure.

Testing Variance calculated from a Sample Compared to a Hypothesized Variance.

The null hypothesis becomes $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$ for a two sided test or $H_0 : \sigma^2 \leq \sigma_0^2, H_1 : \sigma^2 > \sigma_0^2$ or $H_0 : \sigma^2 \geq \sigma_0^2, H_1 : \sigma^2 < \sigma_0^2$ for one-sided test.

The test statistics in this case is the Chi-square distribution, so this is called a χ^2 test. The formula for the test statistic is;

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

where n is the sample size and s^2 is the calculated sample variance and σ_0^2 is the hypothesized or known variance.

Example 17:

Process variance is known to be 0.068. A new method to reduce process time will be considered if it does not increase the variance. (Seems reasonable!) The new method resulted in a variance of 0.087. Using a 5% level of risk (α), test the null hypothesis of there being no difference in the variances versus the new variance is greater than the current process variance. Assume a sample size of 12 for the new process.

$H_0 : \sigma^2 \leq \sigma_0^2, H_1 : \sigma^2 > \sigma_0^2$ and $\alpha = 0.05$ the chi-square statistic also has a dof parameter which for this problem is equal to n-1.

Dof = n-1=12-1=11 and the critical value $\chi^2_{crit} = \chi^2_{.05,11} = 19.675$. (Excel CHIINV(.05,11))

The test statistics is computed to be;

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(12-1)(0.087)}{.068} = 14.074.$$

Since 14.074 < 19.675 we cannot reject the null hypothesis. There is not sufficient evidence that the variance of the new hypothesis is greater than the current process variance.

Testing for Difference in Two Observed Variances Using Sample Data.

When comparing variances of two populations using sample data, the F distribution is the appropriate distribution if both populations from which the samples are drawn are normal. Some authors indicate that off-normal populations can significantly affect the accuracy of the results, others (including Montgomery) indicate that having normal distributions is not as critical as once supposed. The other assumption is that the two samples are independent of one another, e.g. they are not correlated.

The test statistic is given by;

$$F \equiv \frac{s_1^2}{s_2^2}, s_1^2 > s_2^2$$

$$dof_1 = n_1 - 1, dof_2 = n_2 - 1$$

To find critical F-values one can use tables or Excel. Tables of F distribution values are cumbersome to read. Use the Excel function FINV(α,dof1,dof2) instead of tables.

Example 18:

Two competing machine tool companies present equipment proposals to a customer. After proposals have been reviewed, a test is established to determine whether the variances of the products produced by the two machines are equivalent. A sample of 25 units from machine 1 produced a variance of 0.1211 and 30 units from machine 2 produced a variance of 0.0701. At a 90% confidence level, test the hypothesis that the variances are equal. The hypotheses for this two-sided test are,

$H_0 : \sigma_1^2 = \sigma_2^2, H_1 : \sigma_1^2 \neq \sigma_2^2$

and n1=25, dof1 = 25-1=24, n2=30, dof2=30-1=29. The critical F value $F_{crit} = F_{.05,24,29}$= 1.90. Using $s_1^2 = 0.1211, s_2^2 = 0.0701 (s_1^2 > s_2^2)$ one calculates the test statistic,

$$F \equiv \frac{0.1211}{.0701} = 1.73$$

Since F=1.73 < 1.90 = F_{crit} we cannot reject the null hypothesis and therefore we conclude there is not enough evidence to say one tool has better variance than the other tool.

One may wonder isn't there a lower F value that would be a lower confidence limit for this two-sided test and the answer is yes. However, since we insisted that the variance in the numerator be the larger of the two variances there is only one critical F value that we need concern ourselves with and that is the one computed above.

If sample sizes are large (n>100) then one can use a normal distribution approximation to test the equality of two variances. The test statistic is $Z = (s_1 - s_2) \Big/ \sqrt{\left(s_1^2 \big/ 2(n_2 - 1)\right) + \left(s_2^2 \big/ 2(n_1 - 1)\right)}$ and the critical values are taken from the standard normal tables. E.g. $Z_{0.975}$=+/- 1.96. If Z falls outside this range then you can reject the hypothesis that the variances are equal. Note that the standard deviations are used in the numerator of this test statistic.

Testing for Differences in Several Observed Variances.

This test is discussed in length by Montgomery and is called Bartlett's test. It uses the chi-square distribution with (k-1) degrees of freedom where there are k samples whose variances are being tested. It is a one-sided test.

The null hypothesis is

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \text{L} = \sigma_k^2, H_1$: at least one variance is unequal.

The test statistics is

$$\chi^2 = 2.3026\frac{q}{c}, dof = k-1$$

$$q \equiv (N-k)\log_{10}\left(s_p^2\right) - \sum_{i=1}^{k}(n_i - 1)\log_{10}\left(s_i^2\right)$$

$$c \equiv 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}(n_i - 1)^{-1} - (N-k)^{-1}\right]$$

and $N=n_1+n_2+\ldots+n_k$, and $s_p^2 \equiv \frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{N-k}$ is the pooled variance.

Example 19:

We draw four samples of six parts each (one sample from each treatment) and calculate the variance of each sample.

$$s_1^2 = 10.96, s_2^2 = 9.45, s_3^2 = 7.11, s_4^2 = 8.10$$

$$n_1 = n_2 = n_3 = n_4 = 6$$

Test for equality at $\alpha=0.05$ and dof=k-1=4-1=3. $\chi^2_{crit} = \chi^2_{0.05,3} = 7.815$

$$s_p^2 == \frac{5(10.96)+5(9.45)+5(7.11)+5(8.10)}{24-4} = 8.905$$

$$q \equiv (24-4)\log_{10}(8.905) - \sum_{i=1}^{4}(n_i-1)\log_{10}\left(s_i^2\right) = 0.115$$

$$c \equiv 1 + \frac{1}{3(4-1)}\left[\frac{4}{5} - \frac{1}{20}\right] = 1.083$$

The test statistics then evaluates to $\chi^2 = 2.3026\frac{0.115}{1.083} = 0.245$.

Since 0.245 <7.815 we cannot reject the null hypothesis so we conclude that the variances are homogeneous based upon the data extant.

Topic #6: Decision Errors, Producer's and Consumer's risks.

Type I (producer's risk) and Type II (consumer's risk) errors

When testing hypotheses we usually work with small samples from large populations. Because of the uncertainties of dealing with sample statistics, decision errors are possible. There are two types of decision errors: Type I errors and Type II errors.

A Type I error is committed when you reject the null hypothesis when it is true. The probability of committing a Type I error is labeled α, and is the α variable we have been working with in the entire paper. It is the risk that or the probability that the test statistic will lie outside the constructed confidence interval even though the population mean is within the confidence interval.

A Type II error occurs when we do not reject (accept) the null hypothesis when the null hypothesis is in fact false. This risk is labeled β and we have not dealt with this in previous sections of this refresher. Both risks always exist and they are traded off one against another.

If the null hypothesis is that the product is good then, the risk of a Type I error, α, is called the producer's risk since the producer has produced a good product but the consumer rejects it as bad (rejects the null hypothesis). The risk of a Type II error, β, is called the consumer's risk since the consumer accepts a product that is indeed bad (accepts the null hypothesis when the alternate hypothesis is true). To calculate β one must say something specific about the alternate hypothesis. The β risk is different for every alternate hypothesis.

RISK	True State of Affairs	
Decisions	H0 is true	H1 is true
Reject H0	Type I error	correct decision
Accept H0	correct decision	Type II error

To get a better picture of this risk measurement process, look at the standard normal distribution in Figure 1 below. The line at z=1.645 is the value of z for which the probability of exceeding 1.645 is 5%.

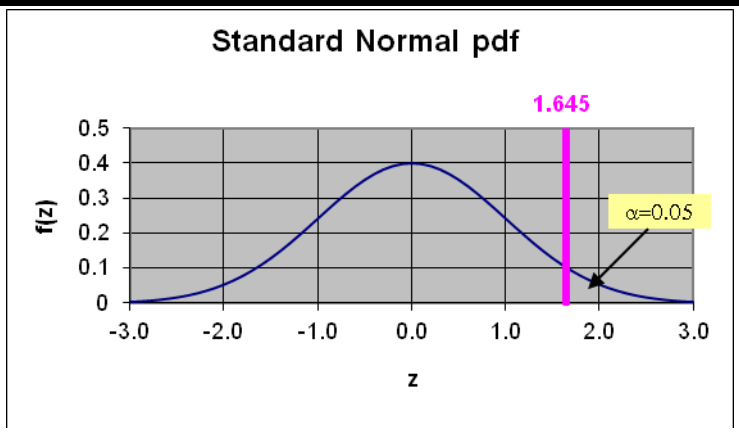


Figure 1. Standard Normal Distribution with z=1.645 as demarcation line for alpha risk of 5%.

The hypotheses for a one-sided test are $H_0 : \mu_1 \le \mu_2, H_1 : \mu_1 > \mu_2$. The null hypothesis can be rejected if μ_1 is sufficiently greater than μ_2. If the standard deviation is known, the test statistic will be calculated as a Z-value, e.g. $Z = (\mu_1 - \mu_2)\sqrt{n}/\sigma$. If the rejection of the null hypothesis only occurs for values of μ_1 sufficiently greater than μ_2, the null hypothesis will only be rejected for Z values that exceed a critical Z value called Zcrit. He value of Zcrit depends only on the risk you are willing to take (α risk). For $\alpha = 0.05$ the one-sided critical Z value is shown in Figure 1, and has a value Zcrit=1.645. If the calculated value of the test statistic, Z, exceeds Zcrit, we will reject the null hypothesis, i.e. μ_1 is more than Zcrit standard deviations larger than μ_2 , so there is a 5% probability or less that $\mu_1 \le \mu_2$. This seems reasonable.

What if instead we wanted to test $H_0 : \mu_1 \ge \mu_2, H_1 : \mu_1 < \mu_2$? The null hypothesis can be rejected if μ_2 is sufficiently greater than μ_1. Again, if the standard deviation is known, the test statistic will be calculated as a Z-value, e.g. $Z^* = (\mu_1 - \mu_2)\sqrt{n}/\sigma$. If the rejection of the null hypothesis only occurs for values of μ_2 sufficiently greater than μ_1, the null hypothesis will only be rejected for Z values that are less than a critical Z value called Zcrit. The value of Zcrit depends only on the risk you are willing to take (α risk). For $\alpha = 0.05$ the one-sided critical Z value is shown in Figure 2, and has a value Zcrit= -1.645.

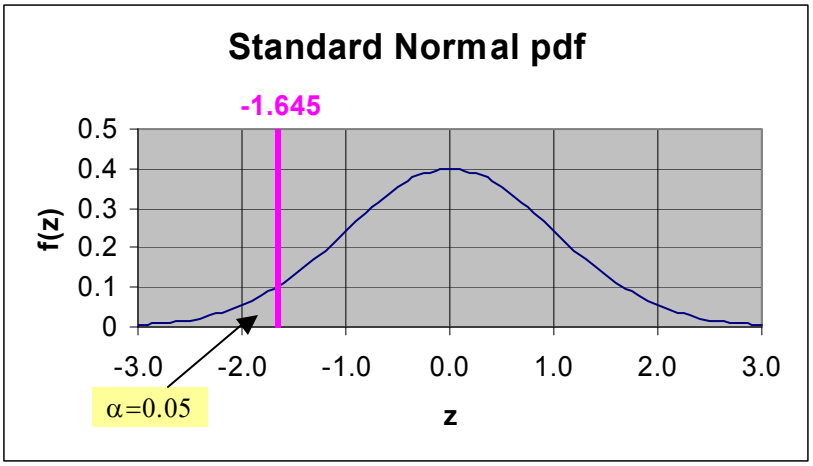


Figure 2. Standard Normal Distribution with z=-1.645 as demarcation line for alpha risk of 5%.

Now we are performing a one-sided (lower tail) test if we are testing if μ_1 is less than μ_2, then Z* will be negative. If Z* is sufficiently negative then we will reject the null hypothesis ($\mu_1 \ge \mu_2$). This makes sense.

Another case of interest is $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \ne \mu_2$. Now we wish to reject the null hypothesis if μ_1 is a good bit larger than μ_2 or if μ_2 is a good bit larger than μ_1. Again defining $Z^* = (\mu_1 - \mu_2)\sqrt{n}/\sigma$. There are two critical Z values, one positive and one negative. For α=0.05 we look for the critical Z values corresponding to having 2.5% probability to the right of the positive value and 2.5% probability to the left of the negative value. See figure 3 below. These values turn out to be $Z_{crit} = +1.960$ and $Z_{crit} = -$

1.960. If the Z* value calculated is either greater than +1.96 or less than –1.96 one would to reject $H_0 : \mu_1 = \mu_2$. This makes sense.

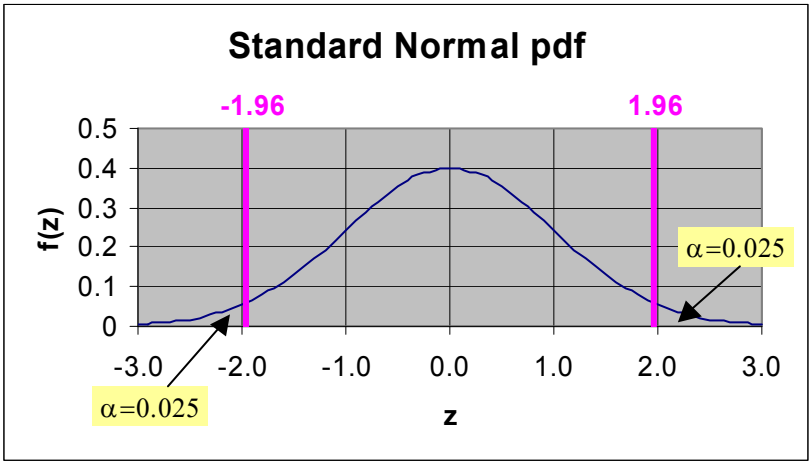


Figure 3. Standard Normal Distribution with z=+1.96 and z=-1.96 as demarcation lines for alpha risk of 5%, i.e. 2.5% in each tail.

To demonstrate the β risk (called the consumer risk) one has to quantitatively state the alternate hypothesis e.g. $H_0 : \mu_1 = 0, H_1 : \mu_1 = 4$. This plot is shown below in Figure 4, where It has been assumed the mean of the alternate distribution is Z=4 and the standard deviations of both distributions is 1. If I pick a producer's risk α=0.025 and if the alternative distribution has a mean at z=4 then the area to the left of the Z=1.96 for the alternative distribution is calculated to be 0.021 and this is the consumer's risk β. i.e. the real mean was z=4 but the unit was accepted as having a z=0 mean.

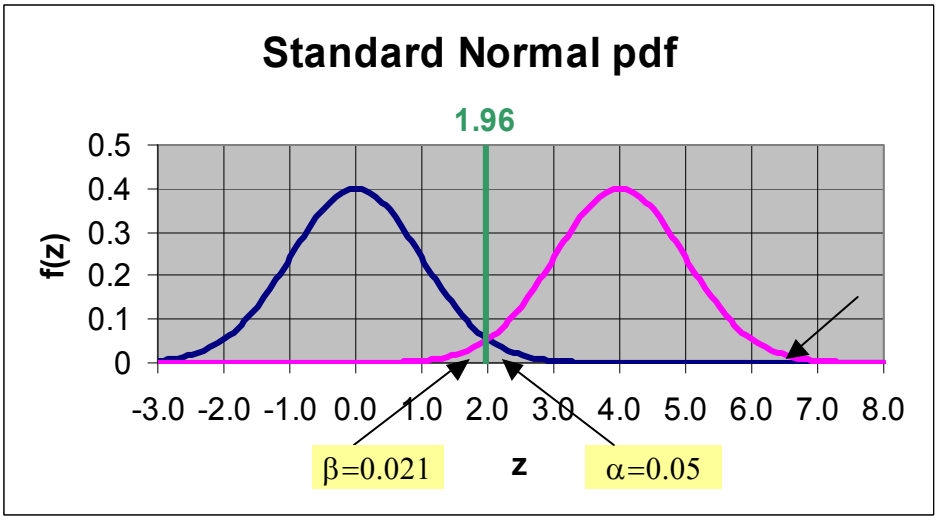


Figure 4. Two Normal Distributions. At Z=+1.96, α = 0.025 and represents the probability of a Type I error if the true mean =0. Postulating an alternative distribution whose mean=4 we see that the Z=1.96 line means that the alternative distribution may have values below 1.96 only 2.1% of the time. This is the consumer's risk β.

Obviously if one postulates a different alternative distribution or mean value or standard deviation there will be different β values for the same α value. This makes the problem a little complicated in general. However, there is another way to view this problem of determining β.

If for example the hypotheses were $H_0 : \mu = 196, H_1 : \mu \neq 196$ and if the true mean = 197 we should reject the null hypothesis. To calculate the probability of acceptance (risk) we calculate the probability of failing to reject the null hypothesis for a specific alternative (e.g. mean = 197).

Try viewing this problem from a confidence interval approach.

If the α risk were set at 5%, and σ=6 and n=9, the confidence interval for the true mean is given by

$(196 - (1.96)6/\sqrt{9} \leq \text{true mean} \leq 196 + (1.96)6/\sqrt{9}) =$ (192.08,199.2). Thus if the true mean were actually 197 the β risk is the probability that the average of a sample of nine values will fall in the range 192.08 to 199.2. How is this calculated?

If the mean=197 then the upper tail probability of rejection is the probability of getting a value greater than 199.2. Z = (199.2 –197)/(6/3)=1.10, P(Z>1.10)=1-P(Z<1.10)=1-.864=.136.

If mean = 1.97 the lower tail probability of rejection is the probability of obtaining a value less than 192.08. This is calculated as Z=(192.08-197)/(6/3)=-2.46 and P(Z<-2.46) = 0.007.

The probability of rejection is the sum of these two probabilities 0.136+0.007=0.143. The probability of acceptance is therefore 1-0.143 = 0.857 and this is the β risk. The β risk is 85.7%. Simply stated, given above data and H_0: μ = 196 the probability of accepting the null hypothesis if the alternate hypothesis were true is the β risk and for μ=197 that equals 0.857. Usually we specify acceptable α and β risks and use them to calculate what sample size we need to assure those risk levels.

Appendix A: Probability Distributions

The following are quick synopses of various continuous and discrete distributions.

Normal Distribution.

Most commonly known distribution called the Gaussian distribution by the person who may have first formulated it. The probability density function, pdf, is given by;

$$f_N(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

where μ = mean and σ = standard deviation from the mean. The cumulative distribution function, CDF is given by;

$$F_N(y) = P(Y \le y) = \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

The pdf is symmetric about the mean, the skewness is zero and the coefficient of kurtosis is 3. The area from $-\infty$ to $+\infty$ equals 1. The mean and standard deviation are independent of one another. This is a necessary and sufficient indicator of a normal distribution. (note σ cannot be a function of μ, and this is not true of many other distributions). Properties: Area under the pdf from $\mu-1\sigma$ to $\mu+1\sigma$ = .6827, area under pdf from $\mu-2\sigma$ to $\mu+2\sigma$ = 0.9545, and area under pdf from $\mu-3\sigma$ to $\mu+3\sigma$ = .9973.

Predictions using the normal distribution.

Transform random variable of interest, x, into the z variable by the transformation;

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

If the mean and the standard deviation are approximated by sample statistics then use,

$$Z = \frac{x-\bar{X}}{s}$$

Then one uses the "Standard Normal Distribution" that is shown in the tables of most statistics books.

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$F(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

Weibull Distribution.

Very versatile distribution as it can assume many shapes. Used extensively in reliability modeling. The pdf is given by.

$$f_W(y) = \frac{\beta}{\eta}\left(\frac{y}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\mu}\right)^{\beta}}.$$

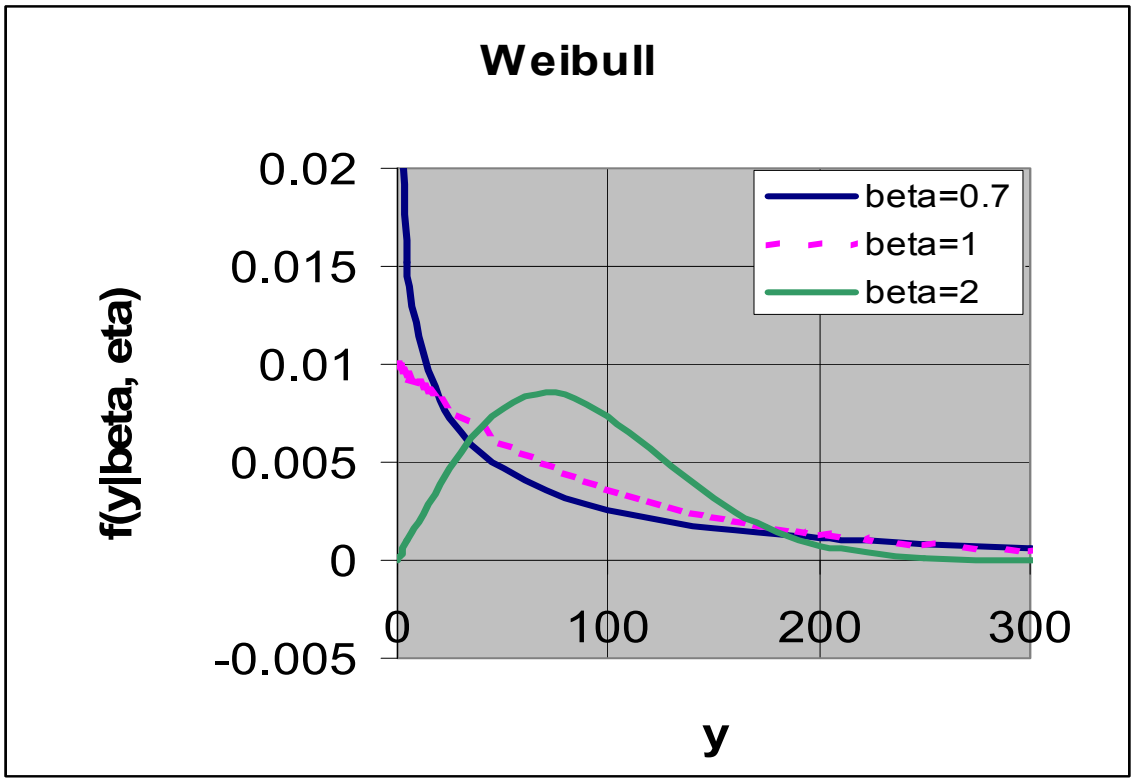
The CDF is found to be,

$$F_W(y) = P(Y \le y) = \int_0^y \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}} dt = 1 - e^{-\left(\frac{y}{\eta}\right)^{\beta}}$$

the $mean = \eta\Gamma\left(1 + 1/\beta\right)$, variance= $\eta^2\left[\Gamma\left(1 + 2/\beta\right) - \left(\Gamma\left(1 + 1/\beta\right)\right)^2\right]$.

β is the shape parameter and determines the distributions shape.
η is the scale parameter, 63.21% of the y values fall below the value.
Γ(x) is the complete gamma function and if x = integer Γ(x) = (x-1)!

The shape parameter alters the distribution shape quite a bit as can be seen in the chart below.



Note the significant change in shape with the parameter β. β=1 is the exponential distribution.

Lognormal Distribution.

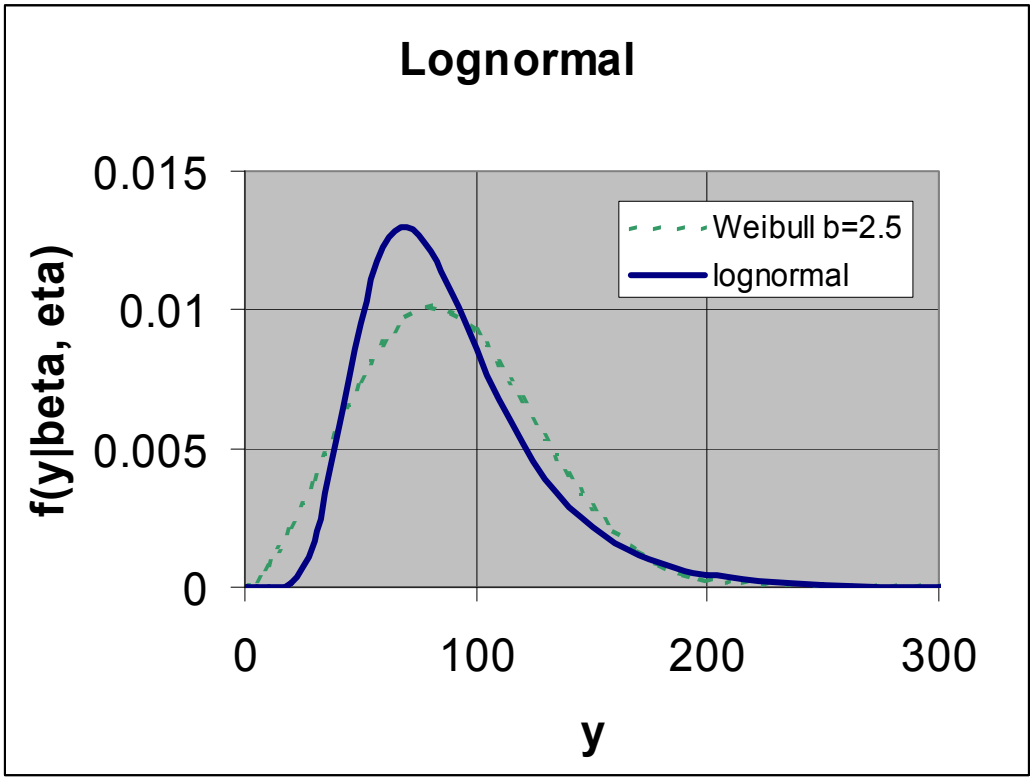
This distribution is defined for x greater than or equal zero and has a positive skewness. It is used to model repair time distributions in determining the availability of a system. The pdf is given by,

$$f_{\log}(y) = \frac{1}{\sigma y\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln y - \mu}{\sigma}\right)^2}$$

where μ = mean of the ln(y) (not the mean of y) and σ^2 is the variance of the ln(y) (not the variance of y). The CDF is given by,

$$F_{\log}(y) = P(Y \le y) = \int_0^y \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2} dx$$

The mean of y itself is given by,
Mean=E[Y]= $\exp(\mu + \sigma^2/2)$, Variance = $(\text{Mean})^2(\exp(\sigma^2)-1)$



Note the somewhat similar shape to a Weibull of shape factor 2.5.

Binomial Distribution.
This distribution is appropriate when one has a series of trials whose outcomes can be only one of two possibilities. This distribution is used in acceptance testing of small samples. The assumptions are that the probability of an occurrence, x, is a constant throughout all the trials (say n trials). Each trial is independent of all previous trials.

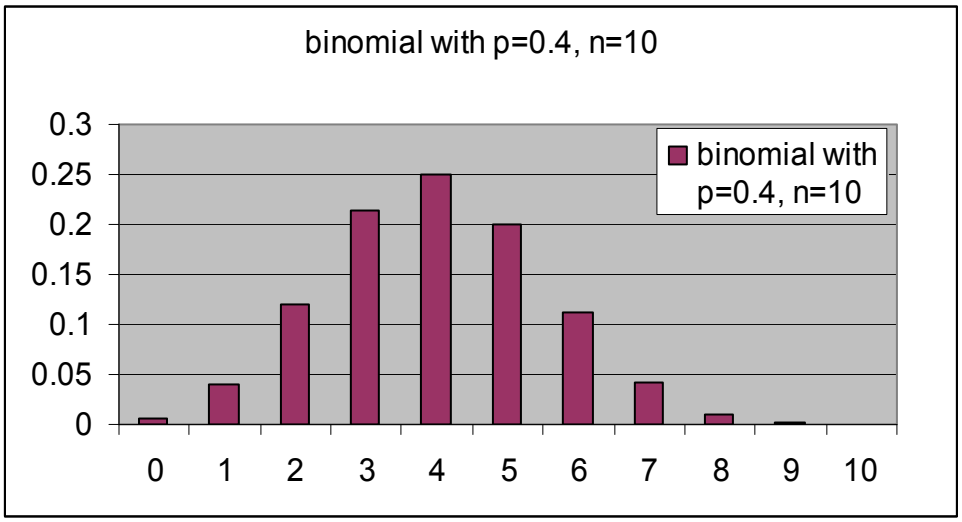
If I let p = probability of the occurrence of interest in any given trial, then 1-p is the probability of nonoccurrence in the same trial.
The probability mass function (as it is called) and the cumulative probability function are given by,

$$\Pr(x \mid n) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} = b(x \mid n, p)$$

$$P(X \le x \mid n) = \sum_{k=0}^{x} \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} = B(x \mid n, p)$$

The mean of this distribution = n*p, the variance = n*p*(1-p).

The ratio of the factorials is merely the number of combinations of x occurrences in n trials i.e. how many arrangements on can have of x occurrences in n trials.

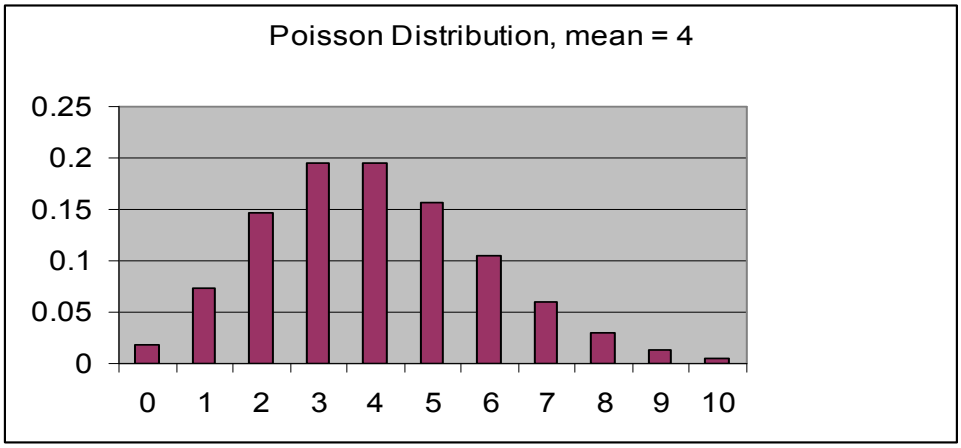


The mean value is equal to n*p = 10*0.4=4.

Poisson Distribution.
The Poisson distribution results from counting types of problems e.g. in one hour how what is the probability of receiving x calls into a call center given that the average incoming call rate is 10 calls in one hour. Each incoming call must be independent of all others. The potential number on incoming calls must be much larger than the expected number of calls. The probability mass function is given by,

$$p_{POI}(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

The mean = λ and the variance = λ.
The figure below shows the probability mass function for the specific case of mean = 4.



Hypergeometric Distribution.
When the population from which sample are drawn is finite and when the sampling is without replacement then the probability of an occurrence does not remain constant but changes depending on the outcome of previous draws. (The gambling game called KINO works using these statistics). Thus the binomial distribution cannot be used.
The pmf for the hypergeometric distribution is given by,

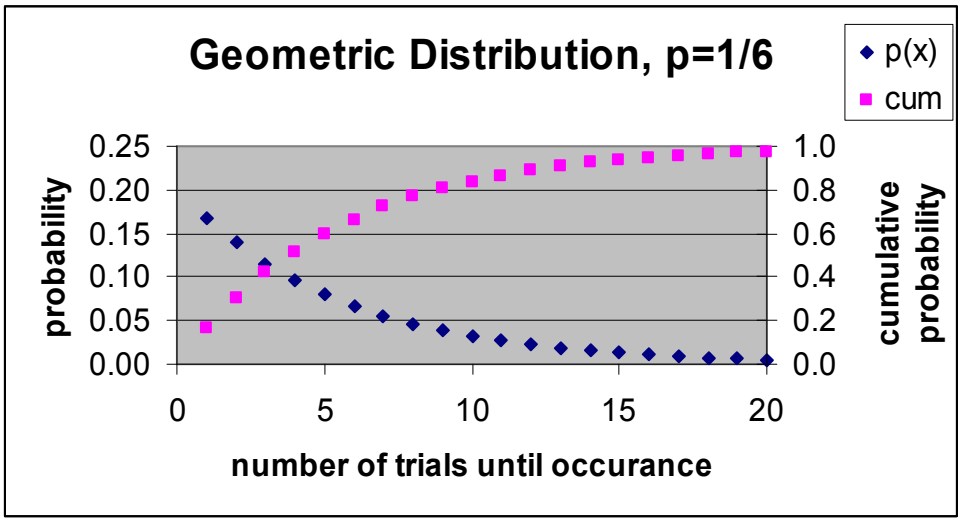
$$p_H(x,n \mid M,N) = \frac{\binom{N-M}{n-x}\binom{M}{x}}{\binom{N}{n}}$$

This is the probability of x occurrences of interest in a sample of size n, given that the population from which one draws the sample is of size N and that the population contains exactly M possible occupancies of interest.
Mean = n*M/N, the Variance = [(N-n)/(N-1)]n(M/N)(1-M/N). If I let M/N = p then we have Mean = n*p and variance = n*p*(1-p)*(N-n)/(N-1) and this looks a lot like the mean and variance of the binomial distribution but with a finite population correction for the variance.
[One can calculate the probability of winning a game of KINO, a gambling game that seems to be a favorite at most casinos, using the hypergeometric distribution. Try it, and then look at the payoffs the casino gives for winning numbers. You will never play KINO again!]
Geometric Distribution.
The geometric distribution is somewhat like the binomial distribution in that for any trial there can be only one of two possible outcomes. The distribution is used to determine the probability of first occurrence of an event given the probability of occurrence in any given trial = p. The pmf is given by,

$$p_g(x) = (1-p)^{x-1}\,p$$

The p(x) represents the probability that an event of interest will occur on the x trial i.e. there were x-1 trials in which the event did not occur that happened before we had a trail in which the event of interest did occur.

This is the classic calculation of Russian roulette in which only one bullet is loaded into one of 6 chambers and the probability of the bullet coming up is 1/6 so the probability that you get shot on the first pull of the trigger is $(1-1/6)^0(1/6) = 1/6$, on chamber is rolled to randomize the bullets location and a second try shows that probability of being shot on the second try is $(5/6)^1(1/6)=5/36=0.014$, for the third try $(5/6)^2(1/6)=25/216=0.116$, for 4^{th} try $(5/6)^3(1/6)=125/1296=0.096$, 5^{th} try $(5/6)^4(1/6)=625/7776=0.08$ and we could go on. Note that the highest probability occurred for the 3^{rd} trial. It is of interest however to look at the cumulative probabilities.
Noting for the geometric distribution one has Mean = 1/p and variance = $(1-p)/p^2$.
A plot of the pmf and the cumulative distribution for p=1/6 is shown below.

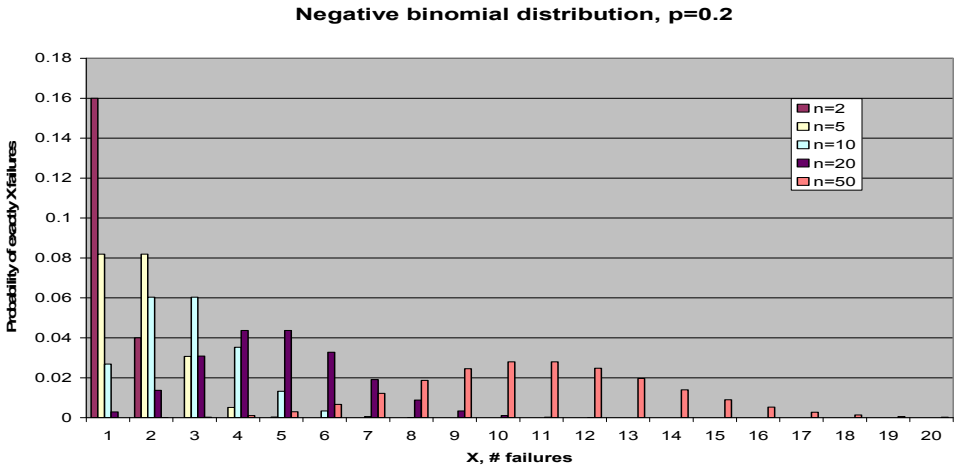


Mean = 1/p = 6, variance = $(1-p)/p^2=30$.

Negative Binomial Distribution.
Used to find the probability of the x^{th} occurrence on the n^{th} trial. The pmf is given by,

$$p_{NBIN}(x) = \frac{(n-1)!}{(n-x)!(x-1)!} p^x (1-p)^{n-x}$$

The mean = x/p and the variance = $x(1-p)/p^2$.



Uniform Distribution.

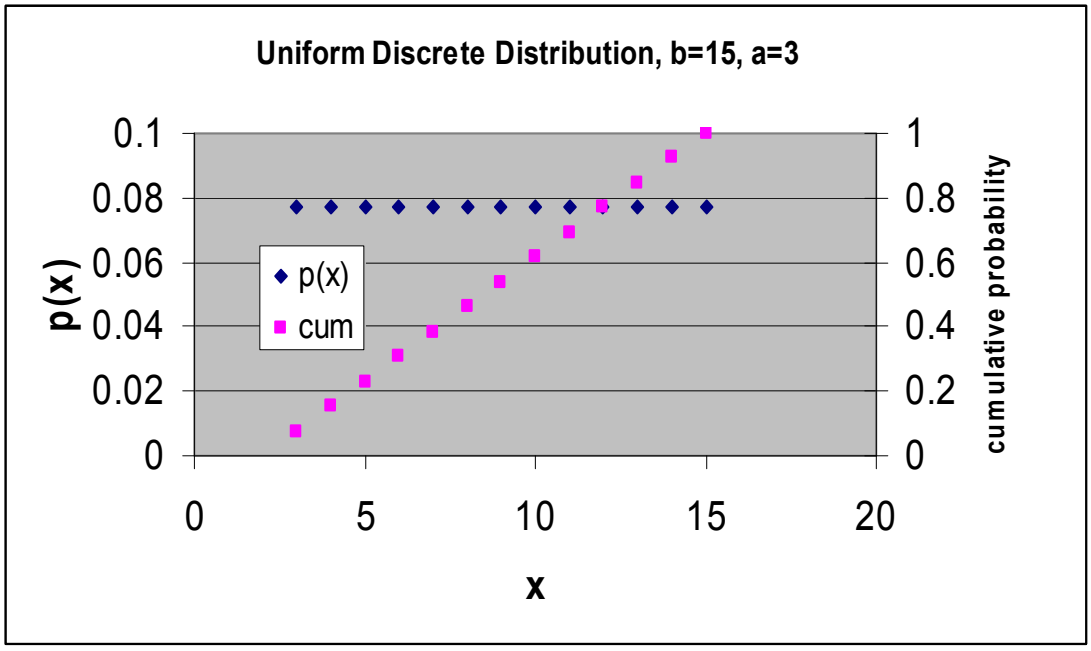
For a discrete uniform distribution ranging from x=a to x=b there are b-a+1 values and each probability must be equal so f(x)*(b-a+1) =1 therefore

$$p_U(x) = \frac{1}{(b-a+1)} = U(a,b), a \le x \le b$$

and the cumulative distribution is

$$P_U(X \le x) = \frac{(x-a+1)}{(b-a+1)}$$

The expected value of x = mean= (b+a)/2, the variance = (b-a)(b-a+2)/12



Fisher's F-Distribution

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

Let

$$y\equiv\frac{\nu_1 x}{\left(\nu_2+\nu_1 x\right)},\frac{dy}{dx}=\frac{\left(\nu_2+\nu_1 x\right)\nu_1-\nu_1 x\nu_1}{\left(\nu_2+\nu_1 x\right)^2}=\frac{\nu_1\nu_2}{\left(\nu_2+\nu_1 x\right)^2}$$

$$f(y\mid \nu_1,\nu_2)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}y^{\frac{\nu_1}{2}-1}(1-y)^{\frac{\nu_2}{2}-1},0\le y\le 1$$

$$P\{y\}=\int_0^y\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}z^{\frac{\nu_1}{2}-1}(1-z)^{\frac{\nu_2}{2}-1}dz=\frac{B_y\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}$$

Gamma Function (Complete):

$$\Gamma(x)\equiv\int_0^\infty e^{-t}t^{x-1}dt,\ \text{x=real}$$

$$\Gamma(n+1)=n!=n\Gamma(n),\ \ n=\text{integer}\ge 1$$

Beta Function:

Complete Beta Function

$$B(\alpha,\beta)\equiv\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},\ \text{where}\ \alpha,\beta\ge 1$$

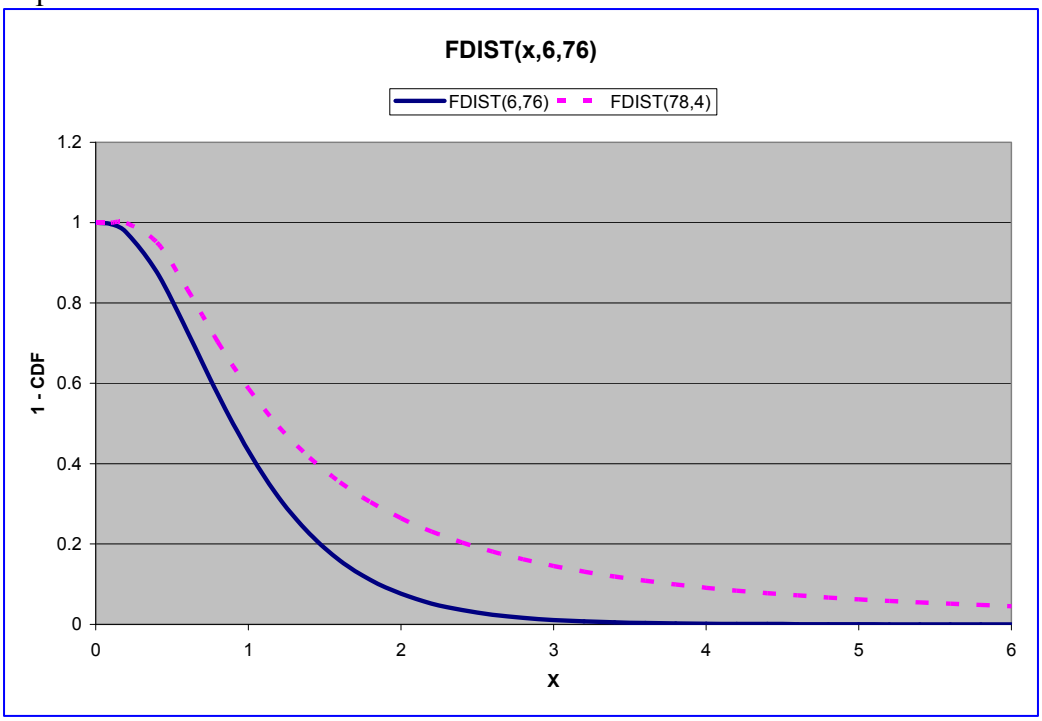
$$\frac{1}{B(n,m)}=m\binom{n+m-1}{n-1}=n\binom{n+m-1}{m-1},\text{where n,m integers}\ge 1$$

$$\binom{\mathrm{a}}{\mathrm{b}}\equiv\frac{a!}{(a-b)!b!},\text{where a,b are integers}\ \ge 0.$$

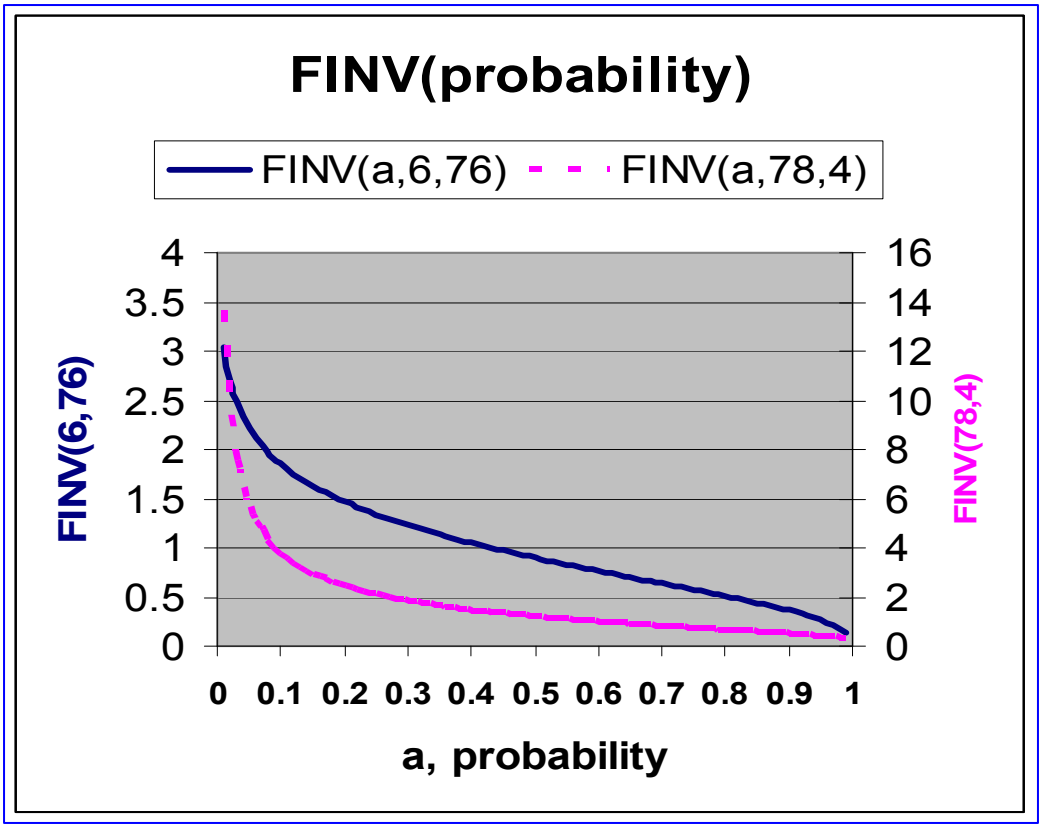
Incomplete Beta Function, $B_p(\alpha,\beta)$

$$B_p(\alpha,\beta)\equiv\int_0^p y^{\alpha-1}(1-y)^{\beta-1}dy=B(\alpha,\beta)-\int_p^1 y^{\alpha-1}(1-y)^{\beta-1}dy$$

A plot of the CDF for two F-distributions is shown below. Note that Excel actually



calculates 1- CDF for this distribution. The inverse of the (1-CDF) is shown below.



Appendix B: Goodness of Fit.

There are a number of methods for determining whether data fits a given distribution. None of the methods are perfect. The three methods we wish to discuss are 1) the chi-square test, 2) the Kolmogorov-Smirnov or K-S test, and 3) the Anderson-Darling test.

Chi Square Test.

As there is always variation in samples taken from populations, one can use the chi square test to test whether the observed results from the sample fit the expected results derived from an assumed distribution. The null hypothesis is that the sample data came from the assumed distribution. We examine the data to see if we can reject this hypothesis at a given level of significance (α).
The chi-square test is a hypothesis test, the level of significance is determined and the expected frequencies are calculated and compared to the observed frequencies. When using the chi-square test it is recommended that any single interval should have a frequency count of at least 5. The test statistic is given by,

$$\chi^2 = \frac{\sum_{i=1}^{k-\text{int}ervals} \left(O_i - E_i\right)^2}{E_i}$$

where O_i is the observed frequency in the i^{th} interval and E_i is the calculated expected value using the assumed population distribution. The degrees of freedom for this test is dof = k-m-1, where k= number of intervals or cells (need not be all of the same size), m = number of parameters that need to be estimated using the sample data. If the distribution is given with its parameters known then m=0. If the mean and standard deviation of the assumed distribution are calculated from the sample data then m=2.

If $\chi^2 \geq \chi^2_{\alpha,dof}$ then reject the null hypothesis i.e. the data does not come from the assumed distribution at level of significance α.
Example.

Anderson-Darling Test.

Ref: D'Agostino & Stephens, Goodness of Fit Techniques, Marcel Dekker Inc., 1986.

For an assumed lognormal or normal distribution calculate

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln(w_i) + \ln(1-w_{n-i+1})\right]$$

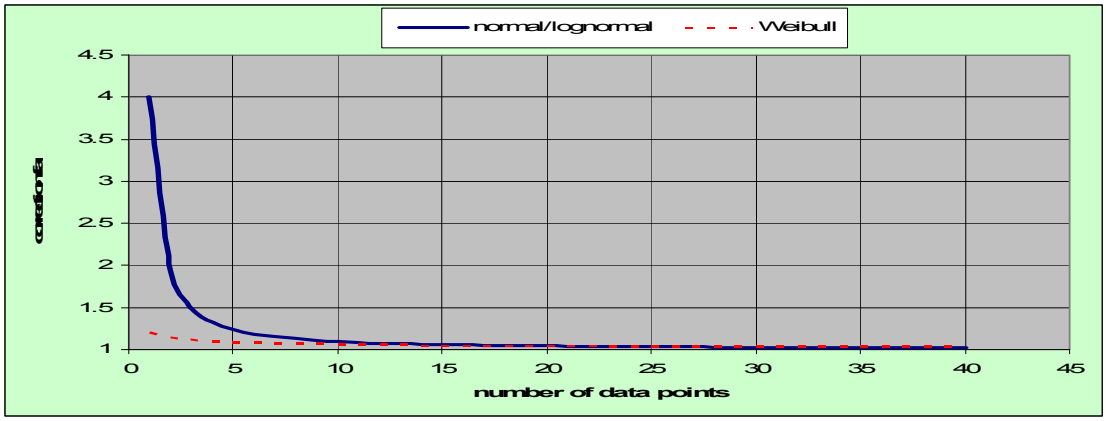
where

$$w_i = \Phi\left(\frac{x_i - \bar{x}}{s}\right)$$

and Φ is the cumulative normal distribution. The mean and standard deviation are computed from the test data.
If $A^2 > A_{crit}^2$ you reject the null hypothesis that the data fits the assumed lognormal or normal distribution. (If the sample size is small there is a correction to A^2)

$$A_m^2 = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$$

and the rejection criterion is $A_m^2 > A_{crit}^2$.



For a Weibull or Gumbel distribution the definition of A^2 is same as above but the definition of w_i is different and given by,

$$w_i = F(x_i) = 1 - \exp(-(x_i/\eta)^{\beta})$$

Again for small sample sizes on uses

$$A_m^2 = A^2\left(1 + \frac{0.2}{\sqrt{n}}\right)$$

and the rejection criterion is $A_m^2 > A_{crit}^2$. The scale and shape parameters are computed from the test data.

The critical values are tabulated in various sources but some are given below.
For normal or lognormal distributions with level of significance α.

α=	0.1	0.05	0.025	0.01
A^2_{crit}=	0.631	0.752	0.873	1.035

For Weibull and Gumbel distributions the critical values are

α=	0.1	0.05	0.025	0.01
A^2_{crit}=	0.637	0.757	0.877	1.038

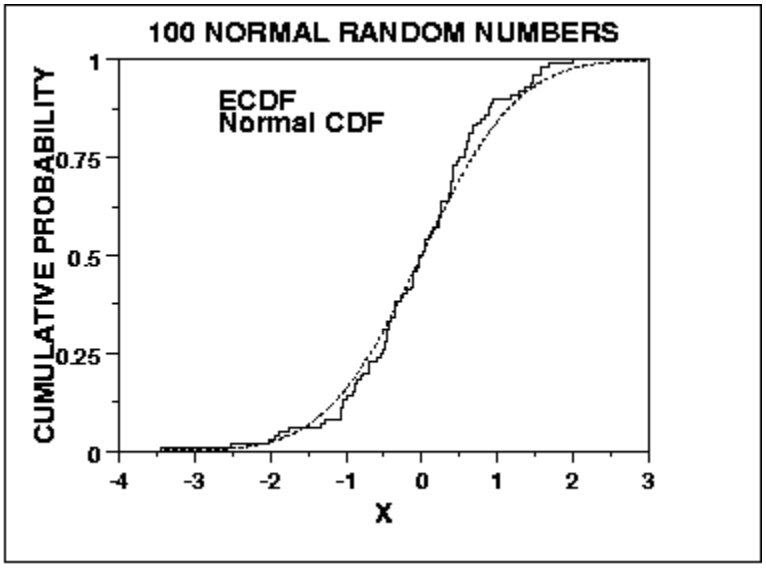
For an Exponential Distribution one uses a Weibull with

Kolmogorov-Smirnov Test.

The Kolmogorov-Smirnov test (Chakravart, Laha, and Roy, 1967) is used to decide if a sample comes from a population with a specific distribution.
The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given N ordered data points Y1, Y2, ..., YN, the ECDF is defined as

$$E_N = n(i)/N$$

where n(i) is the number of points less than Yi and the Yi are ordered from smallest to largest value. This is a step function that increases by 1/N at the value of each ordered data point.
The graph below is a plot of the empirical distribution function with a normal cumulative distribution function for 100 normal random numbers. The K-S test is based on the maximum distance between these two curves.



An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the K-S test has several important limitations:

1) It only applies to continuous distributions.
2) It tends to be more sensitive near the center of the distribution than at the tails.

3) Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation. Due to limitations 2 and 3 above, many analysts prefer to use the Anderson-Darling goodness-of-fit test.
However, the Anderson-Darling test is only available for a few specific distributions if the parameters of the assumed distribution are determined from the data itself. If the test distribution is known completely w/o making use of the sample data then both the K-S and the A-D tests are distribution independent.

Definition The Kolmogorov-Smirnov test is defined by:
H_0: The data follow a specified distribution
H_1: The data do not follow the specified distribution

Test Statistic: The Kolmogorov-Smirnov test statistic is defined as

$$D = \xrightarrow[1 \le i \le N]{\sup} \left| F(Y_i) - \frac{i}{N} \right|$$

where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data).
The practical instantiation of the KS statistic was given by Smirnov. Compute $D^+ = \max(i/N - F(Y_i))$ and $D^- = \max(F(Y_i) - (i-1)/N)$ for i=1,2,…,N then calculate $D = \max(D^+, D^-)$. Use is made of tables of critical values for D that depend on the level of confidence (1-a) desired and sample size. The statistic, D, is distribution independent IF the distribution against which you are comparing the data is fully known prior to gathering the data, i.e. you are not using the test data to find the parameters of the distribution and then doing the KS test. If the latter is the case then the distribution of D does depend on the type of distribution you are assuming for F(Y).

Significance Level: α.
Critical Values: The hypothesis regarding the distributional form is rejected if the test statistic, D, is greater than the critical value obtained from a table. If $D > D_{max,\alpha}$ reject null hypothesis.
(There are several variations of these tables in the literature that use somewhat different scalings for the K-S test statistic and critical regions. These alternative formulations should be equivalent, but it is necessary to ensure that the test statistic is calculated in a way that is consistent with how the critical values were tabulated.)

The Kolmogorov-Smirnov test can be used to answer the following types of questions:
Are the data from a normal distribution?
Are the data from a log-normal distribution?
Are the data from a Weibull distribution?
Are the data from an exponential distribution?
Are the data from a logistic distribution?

Importance: Many statistical tests and procedures are based on specific distributional assumptions. The assumption of normality is particularly common in classical statistical tests. Much reliability modeling is based on the assumption that the data follow a Weibull distribution.

Bartlett's Test for Exponential Distribution.

This is a specific test just for the exponential distribution.

H0: Failures times are exponential
H1: Failure times are not exponential

Test statistic is called B and is computed as follows;

$$B = \frac{2r\left[\ln\left(\frac{1}{r}\sum_{i=1}^{r} t_i\right) - \frac{1}{r}\sum_{i=1}^{r}\ln\left(t_i\right)\right]}{1+\left((r+1)/6r\right)}$$

Where r is the number of failures and t_i is the failure time of the i^{th} unit. Under the null hypothesis B is distributed as a Chi-square variable with r-1 degrees of freedom. The acceptance region for B is given by

$$\chi^2_{1-\alpha/2,r-1} < B < \chi^2_{\alpha/2,r-1}$$

If B falls outside this interval then H0 must be rejected at the 1-α level of confidence.

Mann's Test for Weibull Distribution.

H0: failure times are Weibull
H1: failure times are not Weibull

The test statistic is given by;

$$M = \frac{k_1\left[\sum_{i=k_1+1}^{r-1}\left[\ln\left(t_{i+1}\right)-\ln(t_i)\right]/M_i\right]}{k_2\left[\sum_{i=1}^{k_1}\left[\ln\left(t_{i+1}\right)-\ln(t_i)\right]/M_i\right]}$$

$$k_1 = \left[\frac{r}{2}\right], k_2 = \left[\frac{r-1}{2}\right], M_i \equiv Z_{i+1} - Z_i$$

$$Z_i \equiv \ln\left[-\ln\left(1-\frac{i-0.5}{n+0.25}\right)\right]$$

The symbol [x] means take the integer portion of the number x (this is not a rounding operation). M_i is an approximation. The M statistic under the null hypothesis is approximately an F-distribution variable with numerator dof = $2k_2$, and denominator dof=$2k_1$. If $M > F_{crit}$ then reject H0. This is a test for the tw0-parameter Weibull distribution it does not rule out the three-parameter Weibull or any other distribution. Many times one tries the lognormal as an alternative. Note: Data must be rank ordered.

There are many non-parametric and robust techniques that are not based on strong distributional assumptions. By non-parametric, we mean a technique that is not based on a specific distributional assumption. By robust, we mean a statistical technique that performs well under a wide range of distributional assumptions. However, techniques based on specific distributional assumptions are in general more powerful than these non-parametric and robust techniques. By power, we mean the ability to detect a difference when that difference actually exists. Therefore, if the distributional assumptions can be confirmed, the parametric techniques are generally preferred.

If you are using a technique that makes a normality (or some other type of distributional) assumption, it is important to confirm that this assumption is in fact justified. If it is, the more powerful parametric techniques can be used. If the distributional assumption is not justified, using a non-parametric or robust technique may be required.

Appendix C. Sampling by Variables

Reference: "The Statistical Analysis of Experimental Data" by John Mandel, Dover, 1984, pp 230-232.

Suppose we are testing subsystems and the variable being measured is a voltage,V, that has a lower spec limit of 41 volts. We would like to have a voltage reading of at least 42 volts to give us some margin. If the voltage measurement is as low as 41 volts, we feel compelled to reject that unit with probability 1-β, (beta is called the consumer's risk). If the voltage reading is at least 42 volts, we wish to readily accept the unit with probability 1-α, (alpha is the producer's risk). These two conditions establish a sampling plan and determine what is called an operational characteristic (OC) curve.

Let N be the number of tests to be run. Assume the standard deviation of the measurement apparatus is known, σ_t. The sampling plan consists in giving a value for N and stipulating that a lot be accepted whenever the average $\bar{V}$ of the N tests is at least V_k, where V_k is to be determined. The number V_k must be between 41 and 42 volts. The problem is to determine N and V_k.

Suppose the true voltage is 42 volts. If the sample average, $\bar{V}$, is normally distributed with mean = 42 volts and standard deviation $\sigma_t/\sqrt{N}$. The probability of accepting the voltage is the probability of $\bar{V}$>k which is the probability that t=($\bar{V}$-42)/($\sigma_t/\sqrt{N}$) is greater than (V_k-42)/($\sigma_t/\sqrt{N}$). This probability must be 1-α. Thus we infer that

$$P\left\{\frac{\bar{V}-42}{\sigma_t/N} > Z_{0.05}\right\} = .95 \text{ or } \frac{V_k-42}{\sigma_t/\sqrt{N}} = Z_\alpha = -1.645 \text{ for } \alpha = 0.05.$$

Consider the alternative situation in which we presume the true voltage to be 41 volts. In this case we wish to have the probability of acceptance equal to β. This is the probability that t=($\bar{V}$-41)/($\sigma_t/\sqrt{N}$) > (V_k-41)/($\sigma_t/\sqrt{N}$) and must be ≤β, or

$$\frac{k-41}{\sigma_t/\sqrt{N}} = Z_{1-\beta} = +2.33 \text{ for } \beta=0.01$$

This system of two equations can be solved for N and k giving,

$$V_k = \frac{Z_{1-\beta}V_{accept} - Z_\alpha V_{reject}}{Z_{1-\beta} - Z_\alpha}$$

and

$$N = \left(\frac{\left(Z_{1-\beta} - Z_{\alpha} \right) \sigma_t}{V_{accept} - V_{reject}} \right)^2$$

Using the numerical example with $1-\alpha=0.95$ and $\beta=0.01$, and assume $\sigma=0.45$ volts. $V_k = [(2.33)(42)-(-1.645)(41)]/(2.33 - (-1.645))=[97.86+67.445]/3.975 =41.58$ volts, and $N=[(3.975)(0.45)/(42-41)]^2= 3.2$, This means that you need to test at least 3 more units and the average value of those 3 tests must be ≥ 41.58 volts. It is an interesting result that V_k can be determined without the need to know the standard deviation.

Appendix D: Simple Linear Regression Analysis

(for more detailed information refer to Introduction to Regression Techniques: a whitepaper by Allan Mense, also books by Montgomery et al. and Draper & Smith)

The Basic Idea:

Let's first explore simple linear regression (SLR). The basic SLR regression model uses y as the dependent variable and assumes it depends on only one variable, x. There are many possible forms that could be written. It could be a very complicated function of x such as: $y= \beta_0+\beta_1*x +\beta_2*x^2+\beta_3*\sin(x^5)+ \ldots+\varepsilon$. It helps to know the physics of the situation if such a complex model is to be used. In general much simpler models will be sufficient.

In this note when the coefficients of the model are designated using the Greek letter β, the model is referred to as the population regression model. It is called a linear model because it is linear in the β coefficients. It is nonlinear in terms of the regression (independent) variable x. The simplest version of SLR is. $y = \beta_0 + \beta_1 x + \varepsilon$, where ε is an error term that carries with it the "randomness" characteristic of the real data (you don't get exactly the same value of y for the same values of x). The mathematical job is to find the statistically significant β values used in the regression equation.

Regression analysis is much more complicated than simply "fitting a curve to data." Anybody can put data into excel or some other program and generate a curve fit, but how good is the fit? Are all the input variables important? When regression on more than one variable are there interactions between the input variables and how do these interactions affect the response(s)? How does one know if some terms are significant and others are not? Does the data support the postulated model for the behavior? These questions are not answered by simply "curve fitting."

References:

1. D.C. Montgomery & E.A. Peck, Introduction to Linear Regression Analysis, 2nd edition, Wiley-Interscience Publication, 1992. ,

2. D.N. Gujarati, Basic Econometrics, 3rd edition, McGraw-Hill, 1995.

3. F.A. Graybill, H.K. Iyer, Regression Analysis: Concepts and Applications. This entire textbook can be downloaded from Professor Iyer's web site at Colorado State University.

Simple Linear Regression (SLR).

A simplest linear regression (SLR) model implies we expect to fit response data to a straight line dependent on x. The model equation is written as follows:

1) $$y = \beta_0 + \beta_1 x + \varepsilon$$

or

$$(y - \bar{y}) = \beta_1 (x - \bar{x}) + \varepsilon$$

This is the regression equation of y on x (population regression function or PRF), y is the response variable, x is the regressor variable (or variate), and ε is a (presumed random) error introduced in some manner into the problem. The error could be due to randomness in the process of making a component, randomness in the measurement of a property such as shear stress residing in the component, and there could even be causal

information (i.e. we need to add another regressor variable to the problem). The variance of y is determined by the variance of ε, the only random variable in the problem. SLR assumes 1) E[ε]=0 and 2) $E[\varepsilon^2] = \sigma^2$ which is the same for every value of x (homoscedasticity). This last condition can be extended to cases in which the variance depends on x (heteroscedasticity) using a procedure called generalized or weighted least squares analysis. [*Note: We can never know the parameters β_0 and β_1 exactly and we can only estimate the error distribution, f(ε) or its moments.*] So where does one start?

First, let us assume that the regressor variable, x, is under the control of the analyst and can be set or determined with arbitrarily good precision. It is not a stochastic or random variable. This restriction will be relaxed later. y is of course a random variable since ε is a random variable. It goes without saying that if we do not have an error term that has some random variation then we do not have a statistics problem.

For any given value of x, y(x) is distributed about a mean value and the distribution is the same as the distribution of ε , i.e. var(y(x))=var((ε)). Since the expected value of ε is assumed to be zero, the expected value or mean of this distribution of y, given the regressor value x, is

2) $$\boxed{E[y \mid x] = \beta_0 + \beta_1 x}$$

It is this equation we wish to model and it should be noted that we are "calculating the mean value of y at a prescribed value of x. The variance of y given x is given by the formula

3) $$V[y \mid x] = V(\beta_0 + \beta_1 x + \varepsilon) = V(\varepsilon) \equiv \sigma^2$$

and again we note that equation 3) presumes the variance of the error term is NOT a function of x.

We cannot determine β_0 and β_1 exactly. We need estimators for these two parameters. Call the estimators b_0 and b_1. How do we find b_0 and b_1?

One of the most often used, and easily understood, methods is called the method of least squares.

In brief, what is done is to first construct the deviations of y (taken from the data) from the proposed expected mean value $E[y|x]=b_0+b_1*x$. These differences are called residuals ($e_i= y_i–b_0–b_1*x_i$). The deviation, e_i, is called a residual for the i^{th} value of y. The residuals will be used to estimate ε, the error term, and thus estimate σ^2, the variance. We square these residuals then sum them up and call this the error sum of squares (SSE).

$$SSE = \sum_{i=1}^{n} \left(y_i - (b0 + b1 * x_i) \right)^2$$

This sum is then minimized with respect to b_0 and b_1, the estimators for β_0 and β_1. The two equations, called the normal equations, that are produced have two unknowns (b_0, b_1) and are solved simultaneously. The results are shown below.

4) $$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b0} = 0 = 2(-1)\sum_{i=1}^{n} \left(y_i - b0 - b1 x_i \right) \Rightarrow b0 = \tfrac{1}{n}\sum_{i=1}^{n} y_i - b1 \tfrac{1}{n} \sum_{i=1}^{n} x_i$$

This is well detailed in the literature (Ref 1, pg 9) and the result is usually written

4a) $b_0 = \overline{y} - b_1 * \overline{x},$ where ($\overline{y}$,$\overline{x}$)
are the averages of the measured values.
The second equation is
5)

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b1} = 0 = 2(-1)\sum_{i=1}^{n} x_i \left(y_i - b0 - b1x_i\right) \Rightarrow b1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

5a)

$$b1 = {}^{S_{xy}}\!/_{S_{xx}}, S_{xy} = \sum_{j=1}^{n} \left(y_j - \overline{y}\right)\left(x_j - \overline{x}\right), S_{xx} = \sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2$$

We now have an equation to estimate the expected or average value of y in terms of the independent variable x. This equation gives the <u>expected value of y at a specified value of x</u> and is simply

6) $E[y \mid x] \equiv \hat{y} = b_0 + b_1 * x$

This is the fitted regression line or <u>sample regression function</u> (SRF). The SRF is the estimator for the PRF. It measures the location of the expected mean value of a number of measurements of y given x. This all seems easy! Just wait. We can make the easiest of methods very complicated.

A note to the wise. Do not use equation, 6), to predict vales for E[y|x] outside the limits of the x and y values you used to determine the coefficients b_0 and b_1. This is NOT an accurate extrapolation formula this is an interpolation formula. Sometimes we do violate this rule but be careful when you do.

Adequacy of Regression Model:
To discuss this topic there are a few other terms that need to be defined. Define a total corrected sum of squares SST = $\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = S_{yy}$ that is a measure of whether or not the response is anywhere close to being a constant i.e. simply equal to the mean value everywhere. It is called corrected because the sample mean, $\overline{y}$, is subtracted from the data value

This sum can be divided into two parts, a sum of squares regression SSR = $\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2$ that measures how much of the variance in the data can be accounted for using the SLR model, plus a sum of squares error that measures the randomness in the response about the fitted regression curve. i.e. SSE = $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$.

Now, if one defines a quantity R^2 = SSR/SST, called the coefficient of determination, we will find that R^2 (where $R^2 \leq 1$) measures the variation of y that is explained by the regressor variable x. For the above problem $R^2 = 0.90$ or better indicate a good fit for engineering problems. For econometrics and social science problems R^2 ~.4 to .6 is considered good! Go figure. Values of R^2 close to 1 imply that most of the variability of y is explained by x.

Variations in estimated parameters (coefficients):

The variance of the estimators (or coefficients) b_0 and b_1 are given by (ref 1, pg 14),

7) $$V(b_0) = \sigma^2\left(\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}\right), V(b_1) = \frac{\sigma^2}{S_{xx}}, Cov(b_0, b_1) = -\frac{\bar{x}\sigma^2}{S_{xx}}, Cov(\bar{y}, b_1) = 0$$

We can see several interesting features. First, we cannot determine the variability of the estimators, b_0 and b_1, without knowing the variance of y, i.e. σ^2, which we do not know! The variance can however be estimated from the data we have available. Secondly, the coefficients b_0 and b_1 are negatively correlated.

Finally, one finds that $\bar{y}$ is not correlated with b1, the slope of the regression line.

A General rule: You may have noticed that I call β_0 and β_1 parameters of the population regression equation. I call b_0 and b_1 the estimators of the parameters or coefficients of the sample regression equation. σ^2 is the variance of the random errors in the population regression equation $\hat{\sigma}^2$ is the estimator of that variance using the data that created the sample regression equation. The general rule is that one must find an estimator for every parameter of interest in the population.

Estimating the variance of y:

Estimating the population parameter σ^2 is done by evaluating the residuals.

The j^{th} residual is defined as $e_j \equiv y_j - \hat{y}_j$ and again noting that the sum of the squares of all the residuals is given the name "Error Sum of Squares" or

8) SSE =. $$\sum_{j=1}^{n} e_j^2 \equiv \sum_{j=1}^{n} \left(y_j - \hat{y}_j\right)^2$$

In this simple regression one finds that an unbiased estimator of the population variance is given by SSE/(n-2) = MSE, called the mean squared error, this is not trivial to prove but it is true. There is a 2 in the denominator because we are estimating 2 parameters in this example (b_0 and b_1). Again it can be shown (ref 1, pg 16) that the best estimator of the variance of y, $E[\hat{\sigma}^2] = \sigma^2$, is given by

9) $$\hat{\sigma}^2 = \frac{SSE}{(n-2)} = \frac{1}{n-2}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2$$.

Up to this point in the analysis, nothing has been said about how the errors are distributed statistically. It turns out that least squares estimators (b_0,b_1) are the minimum variance linear estimators. No other estimators that are linear in its coefficients will produce a regression equation that has less variance of $\hat{y}$ than using the b_0,b_1 values determined above. Since $\hat{\sigma}^2$ is an estimator of the variance we can prescribe a confidence interval for

the population variance, σ^2, only if we can say something about how the errors are distributed, i.e. if we presume that errors are normally distributed, then we can use the chi-squared distribution and bound σ^2 by the confidence interval shown below.

Confidence Interval for variance.

Assume a two sided 95% confidence level. If one takes α=0.05 and n=20 data points this interval would become,

$$\frac{(n-2)MSE}{\chi^2_{0.025/2,20-2}} \le \sigma^2 \le \frac{(n-2)MSE}{\chi^2_{0.975/2,20-2}} \quad 10)$$

(Note: For the above confidence interval to be correct the random errors should be normally distributed with variance =σ^2 = constant for all values of x.).

Digression: Note that we can calculate a number of point estimators (b_0, b_1, $\hat{\sigma}^2$) but they do us little good unless we know something about how the random variation is distributed. When we calculated estimators that make use of the sum of many small errors then we can invoke the central limit theorem (and the theorem of large numbers) to let us use the normal distribution as a good guess for the distribution of the estimator. This normal distribution assumption allowed us to use χ^2 distribution to find a confidence interval for the population variance. This assumption is even useful for small samples (e.g. n=20).

The general form for confidence intervals for some parameter, let me call the parameter θ, is given by: $\hat{\theta} - (table-value) \times SE(\hat{\theta}) \le \theta \le \hat{\theta} + (table-value) \times SE(\hat{\theta})$ and the factor called (table-value) is taken from the student t distribution or the normal distribution tables and $SE(\hat{\theta})$ is called the standard error of sample estimator $\hat{\theta}$.

Confidence interval for β_1:

This is also a confidence interval problem. In the analysis we just performed we wished to know β_1 and to do this we calculated the estimator b_1. Reference 1 again gives the following formula for the 95% confidence interval for β_1 using the n=20, α=0.05 example.

$$\left(b1 - t_{.05/2,20-2} \frac{\hat{\sigma}}{\sqrt{20}} < \beta_1 < b1 + t_{.05/2,20-2} \frac{\hat{\sigma}}{\sqrt{20}} \right) = \quad 11)$$

Again, one assumes the errors are normally distributed and the sample size is small (n=20). What does this mean?

For this specific case, and stating it loosely, one says the above interval is constructed in such a manner that we are 95% confident that the actual β_1 value is between the constructed lower limit and the constructed upper limit. If we want a tighter interval about β_1 then we need to have a larger sample size, n or be willing to live with a lower level of confidence. If n=180 instead of 20 we would find the actual interval would be smaller by about a factor of 3 on each side of β_1.

Confidence interval for β_0:

The confidence interval for β_0 for this specific example (n=20, ref 1, pg 25) is given by the following for confidence level = 95%:

12) $$\left(b0 - t_{.05/2,20-2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}, b0 + t_{.05/2,20-2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}\right)$$

Confidence interval for the mean value of y at a given x value, E[y|X=x$_0$].

The formula is $\hat{y}(x_0) = E(y \mid x_0) = b0 + b1 * x_0$ and the confidence interval for E[y|x$_0$] is given by the following formula, again using n=20 data points and setting the confidence level to 95%:

13)

$$\left(\hat{y}(x_o) - t_{.05/2,20-2}\sqrt{MSE\left(\frac{1}{20}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}, \hat{y}(x_o) + t_{.05/2,20-2}\sqrt{MSE\left(\frac{1}{20}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right)$$

Prediction interval for y itself at a given value of x, y(xi).

Finally, if we wish to know the prediction interval for a <u>new set of m observations</u> there is the formula

14) $$\left(\hat{y}(x_i) - \Delta\sqrt{MSE\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right]}, \hat{y}(x_i) + \Delta\sqrt{MSE\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right]}\right)$$

where $\Delta = t_{\alpha/2,n-2}$ for $m = 1$. Note the interval for the prediction about y(x$_i$) is greater than the confidence interval for the mean value. [Why?]

There are also methods for the simultaneous bounding of b_0 and b_1 instead of the "rectangular bounds" imposed by the equations 11) and 12). For m>1 there are other methods for determining the correct value of Δ. (See Ref 1, pg 32 for Bonferroni method).

All this is can get complex fast but it is all well known and documented in the references given at the beginning of this note.

The above information establishes a base from which we can explore other regression techniques.

Lack of fit (LOF):

There is a formal statistical test for "Lack of Fit" of a regression equation. This procedure assumes the residuals are normally distributed, independent of one another and that the variance is a constant. Under these conditions and assuming only the linearity of the fit is in doubt one proceeds as follows. We <u>need data with at least one replicated observation for one or more levels of x</u>. We want two or more values of y for the same values of say x$_i$.(See appendix for discussion of replication vs. repetition). Begin by partitioning the sum of squares error into two parts.

SSE=SS$_{PE}$ + SS$_{LOF}$.

The parts are found by using the formulation $\left(y_{ij} - \hat{y}_i\right) = \left(y_{ij} - \bar{y}_i\right) + \left(\bar{y}_i - \hat{y}_i\right)$ and squaring both sides followed by summing over the m-levels of x and the n$_i$ values measured at each level. i.e.

15) SSE= $$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}_i\right)^2=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+\sum_{i=1}^{m}n_i\left(\bar{y}_i-\hat{y}_i\right)^2$$

This decomposition produces the obvious definitions for a "pure error" term

16) $$SS_{PE}\equiv\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2$$

and a "lack of fit" term,

17) $$SS_{LOF}\equiv\sum_{i=1}^{m}n_i\left(\bar{y}_i-\hat{y}_i\right)^2$$

The justification of this partitioning is that the double sum in the SS_{PE} term represents the corrected sum of squares of the replicated observations at each level of x and then pooling those errors over the m levels of x. *If the assumption of constant variance is true then this double sum is a model independent measure of pure error since only the variability of y at each x level is used to compute SS_{PE}.* Since there are n_i-1 degrees of freedom for pure error at each of the m levels then $\sum_{i=1}^{m}(n_i-1)=n-m$ degrees of freedom total. Similarly for the SS_{LOF} term we see that *it is a weighted sum of squared deviations between the response $\bar{y}_i$ at each level of x and the fitted values $\hat{y}_i$. If the fitted values are close to each level mean value then one has a good fit.* To the extent that SS_{LOF} is large, one has a lack of fit. There are m-2 degrees of freedom associated with SS_{LOF} since one needs 2 degrees to obtain the b_0 and b_1 coefficients in the model. The test statistic for lack of fit is $$F_0\equiv\frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)}=\frac{MS_{LOF}}{MS_{PE}}$$

and since the expected value of $MS_{PE}=\sigma^2$ and the expected value of MS_{LOF} can be shown to be (ref 1, pg 87)

$$E[MS_{LOF}]=\sigma^2+\frac{\sum_{i=1}^{m}n_i\left(E[y_i]-b_0-b_1x_i\right)^2}{m-2}$$

If the regression function produces a good linear fit to the data $E[MS_{LOF}]=\sigma^2$ this implies $F_0 \sim 1.0$. Under this situation F_0 is distributed as an F distribution, more specifically as $F_{m-2,n-m}$. Therefore if $F_0 >> F_{\alpha,m-2,n-m}$ one has reason to reject the hypothesis that a linear fit is a good fit with some confidence level 1-α.. If rejected, then the fit must not be-linear either in the variables or in the functional form of the fitting equation. We will address higher order fitting later in this note and will discuss nonlinear regression in a later note.

Analysis of Variance:

To begin this analysis one takes the difference form of the regression equation as given by

18) $$(\hat{y}-\bar{y})=b_1(x-\bar{x})$$

Viewing the variability of y by expanding the left hand side of the equation, one obtains the basic ANOVA (analysis of variance) equation. Squaring the left hand side and summing over all n data values for y and using the regression fit for E[y|x] and noting that the cross product terms sum to zero one obtains;

19) $$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

SST (or S_{yy}) = SSR + SSE

As noted earlier, SST is called the total (corrected) sum of squares; SSR is called the regression sum of squares and measures the amount of the variability that can be accounted for by the regression model, and finally the term SSE is recognized as the error sum of squares or residual sum of squares.

The SSR term can be rewritten as b_1*S_{xy} using equation 15) and has one degree of freedom associated with it since we are only evaluating b_1 in this expression for SSR. SST has n-1 degrees of freedom since we need only one degree of freedom to evaluate b0 or equivalently $\bar{y}$. The SSE term has n-2 degrees of freedom since this term has b_0 and b_1 to evaluate. To test the significance of the regression model one again makes use of the F-test where the F statistic is given by

20) $$F_0=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE}$$

Noting that the expected values of these terms are given by

21) $$E[MSR]=\sigma^2+\beta_1^2 S_{xx}$$

and

22) $$E[MSE]=\sigma^2$$

The F statistic represents in the limit $F_0 \approx 1+\beta_1^2 S_{xx}/\sigma^2$

Thus F>>1 implies the regression analysis is significant i.e. β_1 is not zero, so there is a significant linear dependency that is represented by the regression. How big must F_0 become to be considered significant?

Up to this point, we have calculated variances and standard deviations from the statistical data. One cannot however interpret these measures of variability without a model of how the randomness of the variables is distributed. For example, what is the probability that β_1 might be between $b_1 \pm 2\hat{\sigma}_{b1}$? The answer is you don't know unless you know how b_1 is distributed? If it is distributed normally, then ~95.4% of the range of β_1 lies within 2 standard deviations of b_1. If it is distributed in some other way then this is not true. How one interprets the exactness of the "fit" is dependent on the statistical distribution of the error term in the population regression equation and the assumptions about how the independent variables are or are not correlated.

Design of Experiments (DOE) vs. Multiple Linear Regression

The question often arises about what is the difference between using the statistical Design of Experiment method versus simply taking data and doing a MLR. The answer is best quoted from Mark J. Anderson, the designer of Design Expert 7, a commercial DOE software program used widely in industry.

Question: 'How is DOE different than multiple linear regression and similar techniques?'"

Answer: " … The big difference with DOE [versus MLR] is that it [DOE] controls the factors that generate changes in the responses rather than just relating (regressing) factors to responses. Ideally, one tries to keep the changes in the factors independent of each other. The easiest way to accomplish this is to run all combinations in a full-factorial design. DOE and regression use the same math for the analysis, and use the same diagnostics, but regression methods without the benefit of design and forethought will have less power than a comparable controlled, designed experiment."

You get more information or the same information with much more accuracy and lower cost using DOE.

MATRIX Math

Let A,B,C be matrices that have the appropriate number of rows and columns to be commensurate with the operations required and assume they are not singular if their inverse is required.

$A+B = B+A$

$(A+B)+C = A+(B+C)$

$(AB)C = A(BC)$

$C(A+B) = CA + CB$

$l(A+B)=lA+lB$, l = scalar

$(A')' = A$, The prime symbol ($'$) means transpose

$(A+B)'= A'+ B'$

$(AB)'=B'A'$

$(ABC)'=C'B'A'$

$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ The superscript ($^{-1}$) means inverse

$(A^{-1})^{-1} = A$

$(A')^{-1} = (A^{-1})'$

$$\underline{\underline{I}} \equiv \begin{bmatrix} 1 & 0 & \mathrm{L} & 0 \\ 0 & 1 & \mathrm{L} & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & \mathrm{L} & 1 \end{bmatrix}, \quad \underline{\underline{J}} \equiv \begin{bmatrix} 1 & 1 & \mathrm{L} & 1 \\ 1 & 1 & \mathrm{L} & 1 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 1 & 1 & \mathrm{L} & 1 \end{bmatrix}, \quad \underline{1} \equiv \begin{bmatrix} 1 \\ 1 \\ \mathrm{M} \\ 1 \end{bmatrix}$$

$$\underline{1}'\,\underline{1} = [n] = n, \quad \underline{1}\,\underline{1}' = \underline{\underline{J}}$$

Let A = matrix of constants, Y = random vector

Define $\underline{W} = \underline{\underline{A}}\,\underline{Y}$, the expected value of the constant is a constant thus E[A]=A

$$E[\underline{W}] = \underline{\underline{A}}E[\underline{Y}], \quad Var[\underline{W}] = \sigma_W^2 = \underline{\underline{A}}Var(\underline{Y})\underline{\underline{A}}'$$

Appendix E: Approximation for Expected Value and Variance of nonlinear function.

Suppose we have a function y=g(x1,x2, … , xn) and we have a set of nominal values of the x variables about which we wish to evaluate y. In close proximity to these nominal values we can perform a Taylor series expansion of g(x). See the last section of this appendix for a short derivation of a Taylor series for those who have forgotten. The derivation and results follow:
Define a vector of nominal values, the operating point so to speak.

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \Lambda\ , \bar{x}_n)$$

$$g(x) = g(\bar{x}) + (x-\bar{x})\cdot\nabla g\,|_{x=\bar{x}} + \tfrac{1}{2}(x-\bar{x})\cdot\nabla\nabla g\,|_{x=\bar{x}}\cdot(x-\bar{x}) + h.o.t.$$

Interest is in finding the expected value of g(x) since x is a set of random variables. Noting that $E[x] \equiv \bar{x}$ the first task is to evaluate E[g(x)].

$$E[g(x)] = E[g(\bar{x})] + E\left[\sum_{i=1}^{n}\frac{\partial g(x=\bar{x})}{\partial x_i}(x_i-\bar{x}_i)\right] + E\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 g(x=\bar{x})}{\partial x_i \partial x_j}(x_i-\bar{x}_i)(x_j-\bar{x}_j)\right] + h.o.t.$$

Since g and its partial derivatives are all evaluated at the nominal operating point values the expectation operator has no effect on those numbers.

$$E[g(x)] = g(\bar{x}) + \left[\sum_{i=1}^{n}\frac{\partial g(x=\bar{x})}{\partial x_i}E(x_i-\bar{x}_i)\right] + \left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 g(x=\bar{x})}{\partial x_i \partial x_j}E[(x_i-\bar{x}_i)(x_j-\bar{x}_j)]\right] + h.o.t.$$

Noting that $E[(x_i-\bar{x}_i)] = E[x_i] - E[\bar{x}_i] = \bar{x}_i - \bar{x}_i = 0$ and $E[(x_i-\bar{x}_i)(x_j-\bar{x}_j)] = \mathrm{cov}(x_i, x_j)$ this equation can be rewritten as follows,

$$E[g(x)] = g(\bar{x}) + \left[\frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2 g(x=\bar{x})}{\partial x_i^2}\mathrm{var}(x_i) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq 1}\frac{\partial^2 g(x=\bar{x})}{\partial x_i \partial x_j}\mathrm{cov}(x_i, x_j)\right] + h.o.t.$$

and in the special case *when there is no covariance* this simplifies to

$$E[g(x)] \cong g(\bar{x}) + \left[\frac{1}{2} \sum_{i=1}^{n} \frac{\partial^2 g(x=\bar{x})}{\partial x_i^2} \sigma_i^2 \right] = g(\bar{x}) + \left[\frac{1}{2} \sum_{i=1}^{n} T_i \sigma_i^2 \right]$$

If g(x) is not a linear function of one or more of the x_j values, the second derivative, T_j is nonzero Making the term in the square brackets nonzero. Remember this is nonlinearity in the region surrounding the nominal operating point. Note that the importance of the nonlinearity is weighted by the variance of the random variable itself. An often heard phrase is that the mean value of a function is not equal to the function evaluated at the mean value. Here is the basis for that statement.

The second job is to evaluate the variance of g(x) about the nominal operating point.

$$Var(g(x)) \equiv E\left[\left(g(x) - g(\bar{x}) \right)^2 \right] = E\left[\sum_{i=1}^{n} \frac{\partial g(x=\bar{x})}{\partial x_i} (x_i - \bar{x}_i) \sum_{j=1}^{n} \frac{\partial g(x=\bar{x})}{\partial x_j} (x_j - \bar{x}_j) \right] + h.o.t.$$

$$Var(g(x)) \equiv E\left[\left(g(x) - g(\bar{x}) \right)^2 \right] = \left[\sum_{i=1}^{n} \frac{\partial g(x=\bar{x})}{\partial x_i} \sum_{j=1}^{n} \frac{\partial g(x=\bar{x})}{\partial x_j} E\left[(x_j - \bar{x}_j)(x_i - \bar{x}_i) \right] \right] + h.o.t.$$

$$Var(g(x)) \equiv E\left[\left(g(x) - g(\bar{x}) \right)^2 \right] = \left[\sum_{i=1}^{n} \frac{\partial g(x=\bar{x})}{\partial x_i} \sum_{j=1}^{n} \frac{\partial g(x=\bar{x})}{\partial x_j} \operatorname{cov}(x_i, x_j) \right] + h.o.t.$$

Defining $S_j \equiv \dfrac{\partial g(x=\bar{x})}{\partial x_j}$, which is called the sensitivity coefficient, the above formula for the variance of g(x) in the neighborhood of the operating point as;

$$Var(g(x)) \equiv E\left[\left(g(x) - g(\bar{x}) \right)^2 \right] \cong \left[\sum_{i=1}^{n} (S_i)^2 \operatorname{var}(x_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} S_i S_j \operatorname{cov}(x_i, x_j) \right]$$

Again *if the variables are not correlated* this expression simplifies to,

$$Var(g(x)) \cong \sum_{i=1}^{n} (S_i)^2 \operatorname{var}(x_i) = \sum_{i=1}^{n} (S_i)^2 \sigma_i^2$$

The equations for the mean and variance are used extensively in the statistical design methods approach to handling variability.

A more rigorous analysis of the variance using the skewness and kurtosis in the Taylor series produces the following expression for the variance.

$$Var(g(x)) = \sum_{i} \left\{ S_i^2 \sigma_i^2 + S_i T_i \sigma_i^3 Sk_i + \left(\frac{T_i \sigma_i^2}{2} \right)^2 (Ku_i - 1) \right\}$$

$\underline{Sk_i \text{ is the coefficient of skewness}}$ and $\underline{Ku_i \text{ is the coefficient of kurtosis}}$ for the i^{th} variable, x_i. T_i is the second derivative of g(x) with respect to x_i. Therefore if you know apriori something about the asymmetry and "peakedness" of the distribution of one or more of the x variables, that information can be inserted into the variance formula through these higher moments.

A derivation of the Var(g(x)) formula follows: It does not require evaluating any higher derivative terms in the Taylor expansion.

$$Var[g(\underset{\rightarrow}{x})] = E[(g(x) - E[g(x)])^2]$$

$$\left(g(\underset{\rightarrow}{x}) - E[g(x)]\right)^2 = \left(\sum_{i=1}^{n} S_i(x_i - \bar{x}_i) + \frac{1}{2}\sum_{i}^{n} T_i\left[(x_i - \bar{x}_i)^2 - \sigma_i^2\right]\right)^2$$

$$E[\left(g(\underset{\rightarrow}{x}) - E[g(x)]\right)^2] \approx \sum_{i=1}^{n} S_i^2 E[(x_i - \bar{x}_i)^2] + 2\sum_{i}^{n} S_i \frac{T_i}{2}\left[E(x_i - \bar{x}_i)^3\right] + \sum_{i=1}^{n}\left(\frac{T_i}{2}\right)^2\left[E(x_i - \bar{x}_i)^4 - \sigma_i^4\right]$$

$$\boxed{Var[g(x)] = \sigma_g^2 \approx \sum_{i=1}^{n} S_i^2\sigma_i^2 + \sum_{i}^{n} S_i T_i \sigma_i^3 Sk_i + \sum_{i=1}^{n}\left(\frac{T_i\sigma_i^2}{2}\right)^2\left[Ku_i - 1\right]}$$

$where\ Sk_i = Sk[x_i] \equiv E[(x_i - \bar{x}_i)^3]/\sigma_i^3, Ku_i = Ku[x_i] \equiv E[(x_i - \bar{x}_i)^4]/\sigma_i^4$

Taylor Series.

Appendix F: Basic Probability Theory.

There are a few basic rules of probability that come in handy when doing statistics.
If A and B represent two events.
The probability of observing both events, called the intersection of events A and B, is given by

P(A and B) = P(A ∩ B)= P(A|B)*P(B).

P(A ∩ B)= P(A)*P(B), if A is independent of B.

The notation P(A|B) is called a conditional probability and is read as "the probability of event A given that event B has already occurred."
This is called the "rule of ANDs." Obviously if the probability of event A does not depend on the event B one has P(A|B)=P(A) and therefore P(A ∩ B)=P(A)*P(B). This implies that events A and B are independent of one another.
Then the probability of observing either event A or event B (or both) is called the union of events A and B, written A∪B. This leads to the "rule of ORs" This rule is simple stated as

P(A or B) = P(A ∪ B) = P(A) + P(B) – 2P(A ∩ B).

P(A ∪ B) = P(A) + P(B), if A and B are mutually exclusive.

If the events A and B are mutually exclusive then P(A∩B) = 0 by definition and this implies P(A ∪ B) = P(A) + P(B).
If there are a number of events each of which must occur for one to have a success and if the probability of the first event is p_1, the second event p_2, …, the n^{th} event p_n. Then the probability of them all occurring is simply the product of each of the events occurring i.e. $p_{Total} = p_1 * p_2 * \ldots * p_n$ or using pi notation $p_{Total} = \prod_{i=1}^{n} p_i$. If all the p_i values are equal to p, then the total probability is simply p^n. If we perform n experiments and the probability of success in any one experiments is p and if p does not change from experiment to experiment, then the probability of x successes and n-x failures in n experiments is certainly proportional to $p^x(1-p)^{n-x}$. Since we do not know in what order the x successes might occur in the n experiments, and given that p is the same for every experiment, we assume a success could occur anywhere in the n experiments. There are n!/[(n-x)!x!]

ways that we could have x successes distributed across n experiments. (one possibility would be to have the first x experiments be successes and the remaining n-x experiments produce failures, another possibility would be to have the first n-x experiments be failures and the last x experiments are successes) One needs the number of combinations of x successes in n experiments and that is given by the factorial expression shown above.) Thus, the probability of x successes in n trials is takes the probability expression, $p^x(1-p)^{n-x}$ and multiplies it by the number of combinations, n!/[(n-x)!x!]. This is the binomial probability function

$$\mathbf{b(x|n,p) = \{n!/[(n-x)!x!]\}\, p^x(1-p)^{n-x}.}$$

It can be shown that the mean of the binomial probability function is

$$\mu \equiv \sum_{x=0}^{n} x * b(x \mid n, p) = np$$

and the variance is found to be

$$\sigma^2 \equiv \sum_{x=0}^{n} (x-\mu)^2 * b(x \mid n, p) = np(1-p).$$

The binomial probability function covers any random situation in which there are only two possible outcomes of an experiment or trial.
If n is large and p small but the product np is of reasonable size (say np >5 and n(1-p)>5), one can approximate b(x|n,p) with a Poisson distribution

POI(k|μ). The Poisson distribution is used extensively in counting experiments.

$$f_{POI}(k) = \frac{\mu^k}{k!} e^{-\mu}.$$

There is another approximation that is useful when n is large. Taking

$\mu = np$ and $\sigma^2 = np(1-p)$,

we can construct a normal probability function that approximates the binomial distribution for n large.

$$f(x) = \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{1}{2}\left(\frac{x-np}{\sqrt{np(1-p)}}\right)^2} \approx b(x \mid n, p)$$

Permutations & Combinations

Suppose we want to find the number of ways to arrange the three letters in the word CAT in different two-letter groups where CA is different from AC and there are no repeated letters.

Because order matters, we're finding the number of *permutations* of size 2 that can be taken from a set of size 3. This is often written 3_P_2. We can list them as:

CA CT AC AT TC TA

Now let's suppose we have 10 letters and want to make groupings of 4 letters. It's harder to list all those permutations. To find the number of four-letter permutations that we can make from 10 letters *without repeated letters* (10_P_4), we'd like to have a formula because there are 5040 such permutations and we don't want to write them all out!

For four-letter permutations, there are 10 possibilities for the first letter, 9 for the second, 8 for the third, and 7 for the last letter. We can find the total number of different four-letter permutations by multiplying 10 x 9 x 8 x 7 = 5040. This is part of a factorial (see note).

To arrive at 10 x 9 x 8 x 7, we need to divide 10 factorial (10 because there are ten objects) by (10-4) factorial (subtracting from the total number of objects from which we're choosing the number of objects in each permutation). You can see below that we can divide the numerator by 6 x 5 x 4 x 3 x 2 x 1:

$$ {}_{10}P_4 \equiv \frac{10!}{(10-4)!} = \frac{10!}{(6)!} = \frac{10*9*8*7*6*5*4*3*2*1}{6*5*4*3*2*1} = 10*9*8*7 = 5040 $$

From this we can see that the more general formula for finding the number of permutations of size k taken from n objects is:

$$ {}_nP_k \equiv \frac{n!}{(n-k)!} $$

For our CAT example, we have:

$$ {}_3P_2 \equiv \frac{3!}{(3-2)!} = 3*2 = 6 $$

We can use any one of the three letters in CAT as the first member of a permutation. There are three choices for the first letter: C, A, or T. After we've chosen one of these, only two choices remain for the second letter. To find the number of permutations we multiply: 3 x 2 = 6.

Note: What's a factorial? A factorial is written using an exclamation point - for example, 10 factorial is written 10! - and means multiply 10 times 9 times 8 times 7... all the way down to 1.

Derivation of the exact confidence limits for proportion data.

Assume you perform n tests and of those tests r of the tests result in failures. The proportion of failures becomes <p>=r/n. This is however a sample of n tests. What if we perform another set of n tests would we get exactly r failures again? Usually the answer is no. The question becomes what is the proportion of bad units (i.e. those that would fail a test) in the population of all the units of interest from which we draw these samples. Again as in many statistics problems we wish to infer a population parameter (p) from a sample estimate of the parameter <p>.
As stated the population proportion is labeled p and we would like to find some interval $[p_L, p_U]$ such that when calculated from sample data the (1-α)100% of the intervals so-calculated would contain the true population proportion, p. The probability α is called the significance level and 1−α is called the confidence level. Here is the probability proposition we wish to make $\Pr\{p_L(\alpha,r,n) <= p <= p_U(\alpha,r,n)\}=1-\alpha$.

The following is a derivation of the interval bounds $[p_L, p_U]$.

Introduction:

When one measures the outcomes of a series of binary events and calculates say a proportion of failures <p> = r/n where r = number of failure in n trials, there is always a question of what is the confidence interval around this proportion.

The classical method relies on calculating the cumulative binomial distribution.

These bounds, labeled $[p_L,p_U]$, depend on the confidence level (1-α) desired for each bound, the number of failures and of course the number of trials.

In most textbooks the procedure is assume the sample size is large enough to call upon the central limit theorem and use a normal approximation that in its least accurate form looks like

$$[p_L, p_U]= \left[< p > -Z_{1-\alpha^*}\sqrt{\frac{< p > (1- < p >)}{n}}, < p > +Z_{1-\alpha}\sqrt{\frac{< p > (1- < p >)}{n}} \right]$$

Let α be the significance level for the upper bound and α* the significance level for the lower bound. The confidence level is therefore $1 - \alpha - \alpha^*$. A more general statement would take the form of the probability statement below.

$$\Pr\left\{ p_L(\alpha^*,r,n) < p < p_U(\alpha,r,n) \right\} = 1-\alpha^*-\alpha$$

Given a hypothetical proportion in a *population* we wish to know what is the chance of that population value, p, being within the above interval. The sample proportion <p> = r/n and we use this information in finding p_L and p_U.

To find the upper (1-α)100% confidence bound, $p_U(\alpha)$, we need to pose the following question: If the population proportion equals p_U, *what is the probability that a sample of size, n, has r or fewer failures.* Call this probability α? The equation that needs to be solved is setting the CDF of the binomial distribution equal to α and finding p_U. That is given $\sum_{i=0}^{r}\binom{n}{i}p_U^i(1-p_U)^{n-i}=\alpha$, solve iteratively to find the p_U value. Example: r=2, n=20, <p>=2/20=0.1, Find the 95% upper confidence bound (α=0.05). Using SOLVER in Excel gives p_U = .2826.

This is a way to get the exact (nonparametric)answer.

To find the lowest (1-α*)100% confidence level value we pose the opposite question: *what value of proportion, p_L, will produce r failures or more in n trials with a probability* α^*. Again the equation to solve is given by $\sum_{i=r}^{n}\binom{n}{i}p_L^i(1-p_L)^{n-i}=\alpha^*$ (or stated in another way $\sum_{i=0}^{r-1}\binom{n}{i}p_L^i(1-p_L)^{n-i}=1-\alpha^*$. Again it can be solved iteratively for p_L. The lower 95% confidence bound is found for the above example to be p_L=0.0181.

To summarize the equations to be solved;

$$\Pr\{p\le r/n\,|p_U\}=\sum_{i=0}^{r}\binom{n}{i}p_U^i(1-p_U)^{n-i}=\alpha$$

$$\Pr\{p\ge r/n\,|p_L\}=\sum_{i=r}^{n}\binom{n}{i}p_L^i(1-p_L)^{n-i}=\alpha^*, \text{ or}$$

$$=1-\sum_{i=0}^{r-1}\binom{n}{i}p_L^i(1-p_L)^{n-i}=\alpha^*$$

The purpose of this note is to calculate the nonparametric expressions for p_L and p_U, i.e. show

$$p_L=\frac{r}{r+(n-r+1)F^{-1}_{1-\alpha^*,2(n-r+1),2r}},\ p_U=\frac{(r+1)F^{-1}_{1-\alpha,2(r+1),2(n-r)}}{(n-r)+(r+1)F^{-1}_{1-\alpha,2(r+1),2(n-r)}}$$

or

$$p_L=1-B_{1-\alpha^*}(n-r+1,r),\ p_U=1-B_{\alpha}(n-r,r+1)$$

where the F-distribution (actually the inverse of the CDF of the F-distribution) and B is the inverse beta distribution. These distributions are used instead of summing the terms in the binomial distribution. This makes calculations ever so much easier as excel has the FINV and BETAINV functions.

A short numerical example helps here.

Example 1: Calculate a 80% confidence ($\alpha=\alpha^*=.1$) interval for n=40 tests and r=2 (# failures). Note <p>=r/n=2/40 =.05 (The Excel function FINV calculates F from highest to lowest values so its argument in the FINV function uses α not $1-\alpha$ for the probability argument). Plugging in numbers to the formulas shown above and noting that $F^{-1}_{.1,78,4}=3.783$, and $F^{-1}_{.1,6,76}=1.853$, one obtains for this 80% confidence interval $\frac{2}{2+(40-2+1)*3.783} \le p \le \frac{(2+1)1.853}{(40-2)+(2+1)1.853}$ or Pr{0.013<p<0.128}=.8.

In statistics courses one learns to use the normal approximation to the binomial distribution. This is not the correct way to solve for the confidence intervals for binary data. It is taught because it is easy to teach not because it is accurate. To demonstrate this consider the above numerical problem. Using the normal approximation one obtains p_L=<p>-1.282*(.1*0.9/40)$^{1/2}$ = 0.006 and p_U=0.094. This gives a confidence interval 0.006<p<0.094 which is smaller than the exact interval by (0.094-0.006)/(0.128-0.013)=0.77. That is, the approximation gives a tighter interval than the exact calculation. What a mistake to make! Underestimating the interval by almost 25%. The correct answer in this counting problem is given by the nonparametric calculation of the interval. The sample size n=40 is considered by most to be "large enough" to use the normal approximation (Appendix C). It is not!

Derivation of the expression p_U and p_L:

Consider the integral

$$\int_p^1 y^{r+1-1}(1-y)^{n-r-1}dy \equiv I_{1-p}(r+1,n-r)$$

This is proportional to the "upper tail" of the cumulative beta distribution (see appendix A) with parameters r+1 and n-r.

The pdf of the beta distribution is of course $f_\beta(y|r+1,n-r) = y^{r+1-1}(1-y)^{n-r-1}/B(r+1,n-r)$.

Integrate I_{1-p} by parts. After the first integration,

$$I_{1-p}(r+1,n-r) = \frac{p^r(1-p)^{n-r}}{n-r} + \frac{r}{n-r}\int_p^1 y^{r-1}(1-y)^{n-r}dy\,.$$

Performing integration by parts again gives

$$\begin{aligned} I_{1-p}(r+1,n-r) &= \frac{p^r(1-p)^{n-r}}{n-r} + \frac{r}{n-r}I_{1-p}(r,n-r+1) \\ &= \frac{p^r(1-p)^{n-r}}{n-r} + \frac{p^{r-1}(1-p)^{n-r+1}x}{(n-r)(n-r+1)} + \frac{r(r-1)}{(n-r)(n-r+1)}\int_p^1 y^{r-2}(1-y)^{n-r+1}dy \end{aligned}$$

Multiplying the above expression by $n\binom{n-1}{r}$ gives a more familiar expression

$$n\binom{n-1}{r}I_{1-p}(r+1,n-r)=\frac{n!}{(n-r-1)!r!}\left\{\frac{p^r(1-p)^{n-r}}{n-r}+\frac{p^{r-1}(1-p)^{n-r+1}r}{(n-r)(n-r+1)}+\frac{x(x-1)}{(n-x)(n-x+1)}\int_p^1 y^{r-2}(1-y)^{n-r+1}dy\right\}$$

$$=\frac{n!}{(n-r)!r!}p^r(1-p)^{n-r}+\frac{n!}{(n-r+1)!(r-1)!}p^{r-1}(1-p)^{n-r+1}+\mathrm{L}$$

These terms begin to look familiar.

Performing this integration r-2 more times produces the result

$$n\binom{n-1}{r}I_{1-p}(r+1,n-r)=n\binom{n-1}{r}\int_p^1 y^{r+1-1}(1-y)^{n-r-1}dy\equiv\sum_{\nu=0}^{r}\binom{n}{\nu}p^{\nu}(1-p)^{n-\nu}\equiv P\{r\}$$

P{r} is the cumulative mass probability for the binomial distribution for r failures in n trials given we know p. It is also called the "lower tail" of the binomial distribution and it will be shown below to equal the "upper tail" of the beta distribution. ***This is a key insight.*** Rewriting the above expression by differencing two integrals with limits (0,1) and (0,p) respectively one obtains,

$$P\{r\}=\frac{1}{B(r+1,n-r)}\left[\int_0^1 y^{r+1-1}(1-y)^{n-r-1}dy-\int_0^p y^{r+1-1}(1-y)^{n-r-1}dy\right]$$

and defining B and B_P one obtains,

$$B(r+1,n-r)=\int_0^1 y^{r+1-1}(1-y)^{n-r-1}dy=\frac{\Gamma(r+1)\Gamma(n-r)}{\Gamma(n+1)}=\frac{r!(n-r-1)!}{n!}$$

$$B_P(r+1,n-r)=\int_0^P y^{r+1-1}(1-y)^{n-r-1}dy$$

the cumulative probability of the binomial distribution from 0 to r is also given by

$$P\{r\}=\left[B(r+1,n-r)-B_P(r+1,n-r)\right]\frac{n!}{(n-r-1)!r!}=$$

$$\left[B(r+1,n-r)-B_P(r+1,n-r)\right]\frac{\Gamma(n+1)}{\Gamma(n-r)\Gamma(r+1)}=\frac{B(r+1,n-r)-B_P(r+1,n-r)}{B(r+1,n-r)}=P\{r\}$$

$$P(r)=1-\beta_P(r+1,n-r),$$

where $\beta_P(a,b)$ is the incomplete beta distribution with parameters (a,b).

[*Note how the factorials are just the complete Beta function. Remember this form for P{r}.*].

If you have software or tables that compute the incomplete beta distribution, $\beta_P(a,b)$ then use those values to compute the needed P(r). Many times however one has only the tables for the F-distribution and not the tables for the beta distribution.

Now explore the form of the F-distribution. The pdf is generally written as shown below.

$$f_F(z\,|\,\nu_1,\nu_2)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\nu_1^{\frac{\nu_1}{2}}\nu_2^{\frac{\nu_2}{2}}\frac{z^{\frac{\nu_1}{2}-1}}{\left(\nu_2+\nu_1 z\right)^{\frac{\nu_1+\nu_2}{2}}},0\le z\le\infty$$

Using the variable $y=\nu_1 z/(\nu_2+\nu_1 z)$ transforms the F-distribution into the form (see Appendix A)

$$f_F(y\,|\,\nu_1,\nu_2)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}y^{\frac{\nu_1}{2}-1}(1-y)^{\frac{\nu_2}{2}-1},0\le y\le 1$$

The cumulative F-distribution is given by integrating over y from 0 to y_p

$$F_F(y_p\,|\,\nu_1,\nu_2)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\int_0^{y_p}z^{\frac{\nu_1}{2}-1}(1-z)^{\frac{\nu_2}{2}-1}dz=\frac{B_{y_p}\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}$$

Which looks a lot like the CDF of the Beta distribution if one lets $\nu_1 / 2 = (r+1)$ and $\nu_2 / 2 = (n-r)$. Thus the function $B_P(r+1,n-r)$ is related to the CDF of the F-distribution by the relation

$F((n-r)/(r+1)*p/(1-p)|2(r+1),2(n-r)) = B_p(r+1,n-r)/\ B(r+1,n-r) = \beta_p(r+1,n-r)$

Note the argument of the F distribution is given by
$((n-r)/(r+1))*(p/(1-p))$

Now

$$\frac{B(r+1,n-r)-B_P(r+1,n-r)}{B(r+1,n-r)}\equiv P\{r\}=1-\ B_p(r+1,n-r)/\ B(r+1,n-r)=1-\beta_P(r+1,n-r)=1-\alpha$$

From this one immediately sees that

$$P\{r\}=1-\Pr\left\{F(2(r+1),2(n-r))<\left(\frac{n-r}{r+1}\right)\frac{p}{1-p}\right\}=\alpha$$

$$\Rightarrow\Pr\left\{F(2(r+1),2(n-r))<\left(\frac{n-r}{r+1}\right)\frac{p_U}{1-p_U}\right\}=1-\alpha$$

Since $\Pr\{F<F_{1-\alpha}\}=1-\alpha$, this implies $\left(\frac{n-r}{r+1}\right)\frac{p_U}{1-p_U}=F_{1-\alpha}^{-1}(2(r+1),2(n-r))$

More practically one uses the FINV function in Excel. Solving for p gives

$$\frac{n-r}{r+1}\frac{p_U}{1-p_U} = F^{-1}_{1-\alpha,2(r+1),2(n-r)}$$

$$p_U = \frac{(r+1)F^{-1}_{1-\alpha,2(r+1),2(n-r)}}{(n-r)+(r+1)F^{-1}_{1-\alpha,2(r+1),2(n-r)}}$$ QED

Just to make the situation more complicated on finds in practice that finding the inverse using excel requires noting that $F^{-1}_{1-\alpha,2(r+1),2(n-r)}$=FINV(α,2(r+1),2(n-r)). Excel calculates the CDF from largest to smallest values of α. See Appendix B

Now it is a little more complex calculating p_L. $\sum_{i=r}^{n}\binom{n}{i}p_L^i(1-p_L)^{n-i} = \alpha^*$ and to use P{r} we need to rewrite this expression as
$1-\sum_{i=0}^{r-1}\binom{n}{i}p_L^i(1-p_L)^{n-i}$ = 1-P{r-1} = α*. Using

$$P\{r-1\} = 1-P\left\{Y < \frac{n-r+1}{r}\frac{p}{1-p}\right\}$$

Setting P{r-1} = 1-α*, p=p_L, and noting the identity for the CDF of the F distribution which is F(α*,2r,2(n-r+1)) = 1/F(1-α*,2(n-r+1),2r). Solve for p_L using r →r-1 one finds,

$$\frac{n-r+1}{r}\frac{p_L}{1-p_L} = F^{-1}_{\alpha^*}(2(r),2(n-r+1)) = 1/F^{-1}_{1-\alpha^*}(2(n-r+1),2r)$$

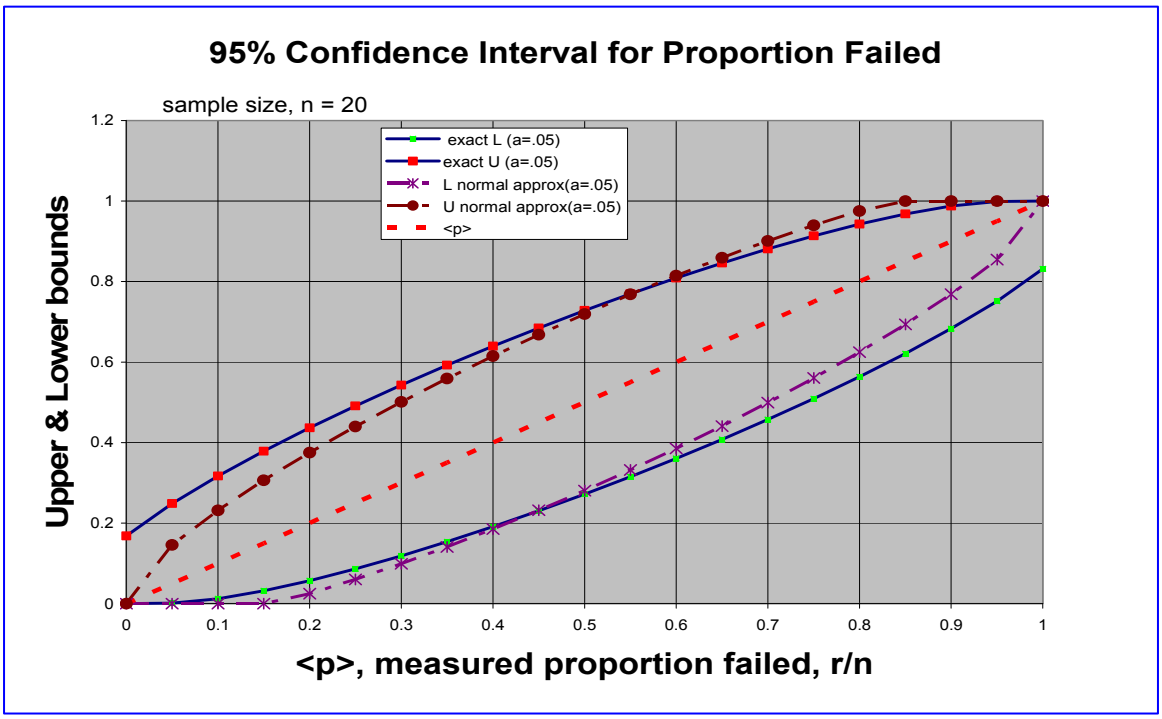
$$p_L = \frac{(r)F^{-1}_{\alpha^*}(2(r),2(n-r+1))}{(n-r+1)+(r)F^{-1}_{\alpha^*}(2(r),2(n-r+1))} = \frac{r}{r+(n-r+1)F^{-1}_{1-\alpha^*}(2(n-r+1),2r)}$$ QED

Thus I have shown that the lower and upper bounds for the nonparametric distribution for binary data can be found without recourse to summing over the binomial distribution. It requires instead the use of the FINV function in excel and also somewhat unfortunately learning the idiosyncrasies of the FINV function in excel, e.g.
$F_{1-\alpha^*,2(n-r+1),2r}$ =FINV(α*,2(n-r+1),2r).

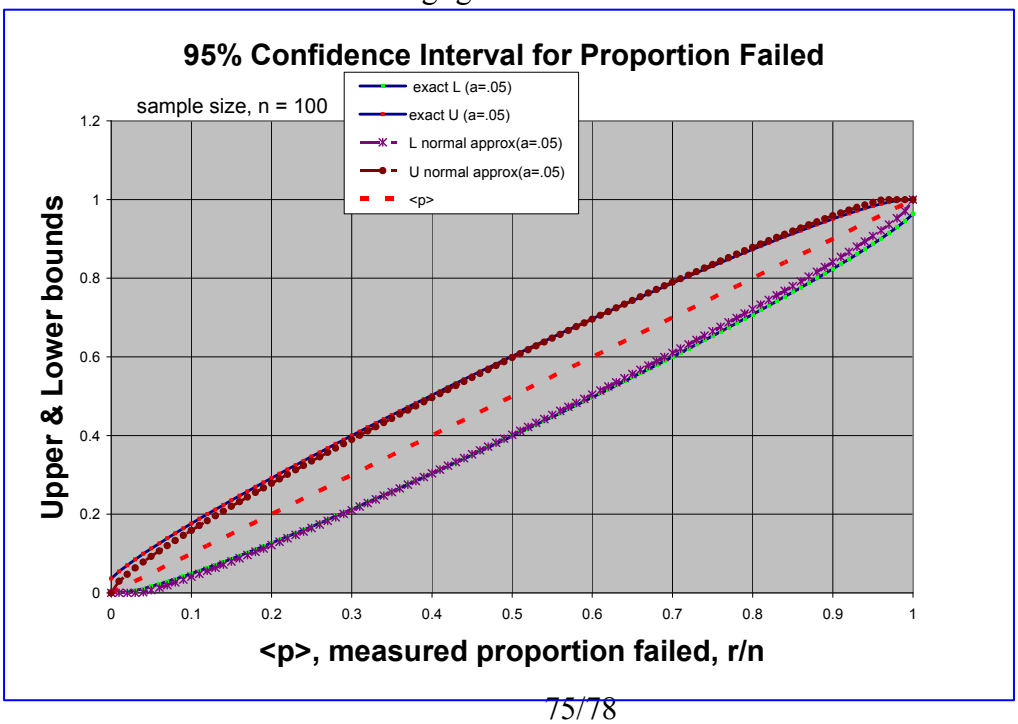
Summary
The bounds to the (1−α−α∗)100% confidence interval are given by the expressions shown below.

$$p_L = \frac{r}{r+(n-r+1)F_{1-\alpha^*,2(n-r+1),2r}},\ p_U = \frac{(r+1)F_{1-\alpha,2(r+1),2(n-r)}}{(n-r)+(r+1)F_{1-\alpha,2(r+1),2(n-r)}}$$

Appendix C



Shown above is plot of exact and normal approximation for the 95% confidence interval for a sample size of 20 with the abscissa = <p>=r/n=# measured failures in n trials. Note the normal approximation, when it is valid, always gives smaller intervals than the exact nonparametric calculation. The chart below shows the same calculation for a sample size n=100. The two calculations are much closer, except for small <p>~0 or large <p>~1 where the differences are non negligible.



Appendix G. Noncentral Distributions

Noncentral F Distribution

Definition of the Noncentral F Distribution.
Similar to the noncentral χ^2 distribution, the toolbox calculates noncentral F distribution probabilities as a weighted sum of incomplete beta functions using Poisson probabilities as the weights.

$$F(x \mid \nu_{num}, \nu_{den}, \lambda) = \sum_{j=1}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^j}{j!} e^{-\frac{\lambda}{2}} I\left(\frac{\nu_{num} x}{\nu_{den} + \nu_{num} x} \middle| \frac{\nu_{num}}{2} + j, \frac{\nu_{den}}{2}\right), \quad \text{noncentral F-distribution}$$

where $I(x \mid a,b) = \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\,dt,$ is the incomplete beta function

I(x|a,b) is the incomplete beta function with parameters a and b, and λ is the noncentrality parameter.

Background of the Noncentral F Distribution
As with the χ^2 distribution, the F distribution is a special case of the noncentral F distribution. The F distribution is the result of taking the ratio of χ^2 random variables each divided by its degrees of freedom. If the numerator of the ratio is a noncentral chi-square random variable divided by its degrees of freedom, the resulting distribution is the noncentral F distribution.

One of the applications of the noncentral F distribution is to calculate the power of a hypothesis test relative to a particular alternative.

Example and Plot of the Noncentral F Distribution
The following commands generate a plot of the noncentral F pdf with noncentrality parameter=10, ν_{num}=5, ν_{den}=20..

```
x = (0.01:0.1:10.01)';
p1 = ncfpdf(x,5,20,10);
p  = fpdf(x,5,20);
plot(x,p,'-',x,p1,'-')
```

F-Distribution (or Snedecor's F distribution) with degrees of freedom ν_{num} and ν_{den}.

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

$$\mu = \frac{\nu_2}{\nu_2 - 2}, \nu_2 > 2$$

$$\sigma^2 = \frac{2\nu_2^2\left(\nu_1 + \nu_2 - 2\right)}{\nu_1\left(\nu_2 - 4\right)\left(\nu_2 - 2\right)^2}, \nu_2 > 4$$

Noncentral t Distribution

Definition of the Noncentral t Distribution

The most general representation of the noncentral t distribution is quite complicated. Johnson and Kotz [27] give a formula for the probability that a noncentral t variate falls in the range [–t, t].

$$\Pr\{-t, x, +t | \nu, \lambda\} = \sum_{j=0}^{\infty} \frac{\left(\frac{\lambda^2}{2}\right)^j}{j!} e^{\frac{\lambda^2}{2}} I\left(\frac{x^2}{\nu + x^2} \middle| \frac{1}{2} + j, \frac{\nu}{2}\right)$$

I(x|a,b) is the incomplete beta function with parameters a and b, λ is the noncentrality parameter, and ν is the number of degrees of freedom.

Background of the Noncentral t Distribution

The noncentral t distribution is a generalization of Student's t distribution. Student's t distribution with n – 1 degrees of freedom models the t statistic, $t \equiv \frac{\overline{x} - \mu}{s/\sqrt{n}}$, where $\overline{x}$ is the sample mean and s is the sample standard deviation of a random sample of size n from a normal population with mean μ. If the population mean is actually μ_a, then the t-statistic has a noncentral t distribution with noncentrality parameter

$$\lambda = \frac{\mu_a - \mu}{\sigma/\sqrt{n}}$$

The noncentrality parameter is the normalized difference between μ_a and μ. The noncentral t distribution gives the probability that a t test will correctly reject a false null hypothesis of mean μ when the population mean is actually μ_a; that is, it gives the power of the t test. The power increases as the difference $\mu_a - \mu$ increases, and also as the sample size n increases.

Example and Plot of the Noncentral t Distribution

The following commands generate a plot of the noncentral t pdf with non centrality parameter=1 and df=10.

```
x = (-5:0.1:5)';
p1 = nctcdf(x,10,1);
p = tcdf(x,10);
plot(x,p,'-',x,p1,'-')
```

The student t-distribution with n degrees of freedom.

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, -\infty < t < +\infty$$

$$\mu = 0$$

$$\sigma^2 = \frac{\nu}{\nu - 2}, \nu > 2$$

$$\alpha_3 = \sqrt{\beta_1} = 0, \nu > 3$$

$$\alpha_4 = \beta_2 = 3 + \frac{6}{\nu - 4}, \nu > 4$$

References:

1) Statistical Theory with Engineering Applications by A. Hald, Chap 21, Wiley (1952),
2) Statistical Theory and Methodology for Science & Engineering by K. Brownlee, Chap 3, Sec. 3.5, Wiley (1965)
3) Reliability Engineering Handbook by B. Dodson & D. Nolan, Chap. 3, Quality Publishing (1999),
4) Statistical Methods for Reliability Data by Meeker & Escobar, Chap 3, Wiley (1998)
5) Nonparametric Statistical Methods by M. Hollander & D. Wolfe, Chap 1, pg 32, Wiley (1999)