

Business Statistics

Kevin C. Kaufhold

Working Paper 2012:2, latest revision January, 2012

Preamble and Sources

Statistical analysis plays a critical role in the consideration of all financial matters. One must have a firm background in Statistics to understand the complexities of Finance. The following is a brief review of some of the important mathematical concepts pertinent to investment activity. This document is essentially an outline of college level discourses including: *Statistical Techniques in Business and Economics*, 10th Edition, Robert D. Mason, Douglas A. Lind, William G. Marchal, Irwin/McGraw-Hill, 1999; *Statistics for Management and Economics*, Gerald Keller and Brain Warrack, Thomson, 6th ed. 2003; and *Quantitative Methods for Investment Analysis*, Richard A. Dufusco, Dennis W. McLeavey, Jerald E. Pinto, and David E. Runkle, 1st ed. 2001.

Table of Contents

Business Statistics	1
Preamble and Sources	1
Contents	1
Introduction.....	2
Descriptive Statistics.....	4
On Sampling	13
On Probability	18
Probability Distributions	25
Hypothesis Testing.....	33
Linear Regression Analysis	41
Times Series and Forecasting	54
Basic Statistical Concepts Regarding Investments	56

Introduction

Statistics provide tools for analyzing data and then drawing conclusions. Descriptive statistics is the study of how data can be summarized effectively to describe important aspect of large data sets. Inferential statistics makes forecasts, judgments from a smaller group that is actually observed. Probability theory is the foundation of inferential statistics.

A few introductory statistical items follow. A population is a set representing all observations of interest to a researcher. An observation is a unit of something observed or a unit of analysis. It includes the elements that the population is composed of (examples: people, states, lab mice, stocks). A census is the process of collecting information from all members of a population. A sample is a subset of observations taken from the population. A variable is a characteristic or descriptor of a member of an observation.

Data are the values of one or more variables recorded from a census or sample. Data can be organized in three forms: cross sectional data; time series data; and panel or pooled data. Cross section data has the unit of observation being an object. With time series data, the unit of observation is a time period. Panel data has pooled time series and cross section information.

Data is arranged as ratios, intervals, ordinal, or nominal information. Ratio and interval data are sometimes called quantitative data, while ordinal and nominal data are considered qualitative data. Ratio data contains numeric measurements of an amount or quantity. Data of this type must have a unique origin. This is an interval scale with zero as the point of origin. With this measurement, we can add and subtract as well as compute ratios (rates of return, etc). Interval data have numeric measurements with order and distance, but no unique origin. Interval data are real numbers. Calculations can be performed on interval data. Ordinal data, or ranked data, use measurements (numeric or non-numeric) that have order but no distance. Ordinal data are nominal data that are in a ranking or ordering. Calculations can be made on the ordering process. Nominal data, or categorical data, are measurements (numeric or non-numeric) of an attribute or quality that cannot be ordered. Each possible value that a qualitative variable can take on is called a level. Values of nominal data are categories. Nominal data can only have calculation performed based on the number of frequencies of occurrence.

Statistics are used to make inferences about parameters of a sample. Statistics are often used to estimate parameters. A parameter is a descriptive measure of a population characteristic (such as mean value, range of investment returns, variance). A sample statistic estimates an unknown population parameter (such as 2.52 for the mean value).

Summation Algebra and rules.

$$\sum X_i = x_1 + x_2 + \dots + x_n$$

Rules include:

$$\sum cx_1 = c \sum x_1, \text{ where } c \text{ is a constant.}$$

$$\sum (x_1 + y_1) = \sum x_1 + \sum y_1$$

$$\sum x_1 * \sum y_1 = \sum \sum x_1 y_1$$

Descriptive Statistics

Tools and procedures used to summarize data are referred to as descriptive statistics. Inferential Statistics are tools and procedures used to make inferences about a population based upon information in a sample.

A Confidence level is the proportion of times that an estimating process will be correct, while the significance level is the proportion of times that a conclusion will be wrong in the long run.

Frequency Distribution. A frequency distribution is a grouping of data points placed into categories sharing the number of observations in each mutually exclusive category. Raw data, on the other hand, is unorganized data. A distribution is organized into classes with upper and lower limits. The classes can be mutually exclusive or inclusive, with some overlap between the classes.

The class frequency is the number of units in a class. The class interval is the amount of units between the upper and lower limits of the class. The class midpoint is halfway between the upper and lower limits, and is also called the class mark. Typically, class intervals should be equal. Professional judgment can determine the number of classes. But there are a few rules that are normally followed – the lower limit of a class should be an even multiple of the class interval, and overlapping class limits should be avoided, as well as open ended classes. Additionally, there is a 2 to the k class rule in determining the appropriate number of classes.

Histograms. This is a graph in which classes are marked on the horizontal axis, and class frequencies are on the vertical axis. This generates a visual representation of a frequency distribution. A histogram is a visual, graphical display of a frequency distribution. With a Normal distribution, the histogram takes on a bell curve shape.

A frequency polygon is the same things as a histogram, except that interval midpoints are plotted with a line, instead of intervals being plots in sequential bars.

Relative Frequency Distribution. Class frequencies can be divided by the number of observations, generating a percentage for each class. A relative frequency is the number of times that a certain interval has been observed compared to the total frequency distribution. Thus, it is expressed as a percentage, and is: $\frac{\# \text{ observations}}{\text{total observations}}$.

A stem and leaf display can be used, whereby the leading digit is the stem and every trailing digit is located in the stem. For instance, 9/6434567 would display the values 96, 94, 93, 94, 95, 96, 97. Then, the leaf can be sorted by frequency of unit value – 9/3445667 would be 93, 94, 94, 95, 96, 96, 97. The stem and leaf display thus eliminates a problem with frequency distributions that do not show the exact identity of each value,

thereby making exact determination impossible of the distribution of values inside that class.

A relative frequency histogram is similar to a frequency histogram, except it shows the percentage of observations in each class. A cumulative relative frequency adds the relative frequencies of other, preceding intervals. Other frequently used diagrams include an Ogive (which is a graphical representation of cumulative relative frequencies); a Stem and Leaf Display (this is similar to a histogram, only turned on its side. The purpose of the display is to actually see the data in each class); A pie chart (This focuses on the proportion of occurrences); a bar chart (focusing on the frequency of occurrences); and a line Chart (which is a time series plot).

Measures of Central Tendency includes the Mean, Median, and the Mode. The measure of central tendency is the average or center of values. The Expected Value of the measure of central tendency is the average value of a probability distribution. If x and y represent values in one or more populations, and c is a constant:

$$E(x+c) = E(x) + c$$

$$E(cx) = cE(x)$$

$$E(x+y) = E(x) + E(y)$$

$$E(xy) \text{ may, or may not, be equal to } E(x)E(y)$$

The Population Mean (μ , pronounced mu) is the arithmetic mean. It includes all values included in the study. The arithmetic mean is the point in which the deviation of each value from the mean is zero. The mean is the balance point between all values of the study. The mean can be unduly affected by large or small values, though, and a mean cannot be determined from open ended data (such as \$100,000 or more). When one speaks of “an average”, the arithmetic mean is what is usually referred to. Where μ is the mean, X_i is the i th observation of the population, and N is the total number in the population, the population mean is expressed as:

$$\mu = \sum X_i / N$$

The population mean is a measure of central tendency for a population, of course, and is considered the “expected value” of the population, $E(x)$. It is known as the “first moment” of the population. μ need not be an observable value in the population. It is only meaningful for quantitative data, and is the balancing point of the histogram.

The weighted mean weights equal values by the number of values of equal weight. It is normally a short-cut to derive the arithmetic mean, and will generate the same value. Example: $((3*6.25) + (5*6.50)) / 8$ is an example of a weighted mean. It will be the same value as if one would add each of the values and then divide by the number of units. The weighted mean weights the different values of a sample. If 75% of a portfolio has 10% return, and the other 25% has a 5%, then the weighted mean would be $.75*.10 + .25*.05$. The equation is: $\bar{X} = \sum w_i X_i$, where $\sum w_i = 1.00$. In terms of portfolio returns, a long position will have a positive weight, and a short will have a negative

weight. Each X_i would be the stock and bond percentage of a portfolio, or the individual assets within a portfolio. When forward data is used, a portfolio expected return results from the weighted mean.

The geometric mean is used for the average rate of change over time. The arithmetic mean is more useful for a mean at any one point in time. The geometric mean is the square root of the products of each observation. $R_n = \text{square root of } ((1+r_1)(1+r_2) \dots (1+r_n)) - 1$.

The geometric mean is: $G = \text{nth } \sqrt{(X_1 * X_2 * X_3 * \dots * X_n)}$. Note that this is the nth root of the geometric expression, and not $n * \text{the square root of the geometric expression}$. In natural log terms, $G = (1/n) * \ln(X_1 * X_2 * X_3 \dots X_n)$, or shorter, $G = (\sum \ln X) / n$. To avoid a problem with taking the nth root of a negative return (which would generate an error term on the calculator), when returns are negative, X_i becomes $(1 + R_i)$. So 10% is 1.10 while -10 is .90.

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocal of a set of numbers. It is also described as the reciprocal of the mean of two extremes. $b = 1 / ((1/a) + (1/c)) / 2$. It is also $b = 2ac / (a+c)$, if $(b-a) / c = (c-b) / c$. Example, if 12 and 6 are the two extremes, 8 is the harmonic mean. The harmonic mean comes from the early Greeks, and is actually used in music applications.

The sample mean measures something less than the entire population. A sample mean is the selection of a sampling from the general population of the study, with an average then generated. This is used when selection of all units is impossible, such as where the sampling of the product would destroy the product itself (example: measuring the average breaking tensile strength of wires). A sample mean is instead generated. The sample mean is the mathematical average of the sample. Note the change to lower case letters and \bar{x} to distinguish a sample from a population.

$$\bar{x} = \sum x_i / n$$

\bar{x} is a sample statistic, but needs not be an observed value in the sample. \bar{x} is only meaningful for quantitative data.

The median (M_d or m) is the middle value in a population or sample. It satisfies the requirements that no more than half of the observations are above the median, and no more than half are below. It is the mid-point, so the median will be unaffected by extreme values. For even number of values, the median is the average of the two middle values. Throw out the highs and lows. The last value remaining will be the median. Note that the median is not influenced by extreme values, as it is the middle term of the sample. The median is the value of the $(n + 1) / 2$ observation. If the series is even, then the median is the mean of the $(n + 1) / 2$ and $(n + 2) / 2$ observations. An equal number of observations lie above and below the median. The median is useful when working with a skewed distribution, since both the mean and the mode are affected by a skew.

The median may be calculated for a population or a sample. M_d is not meaningful for nominal data.

The mode (M_o) is the most frequently occurring value in a population or sample. The mode could be far removed from the mean and median. With one maximum interval, the distribution is said to be unimodal. If two intervals are at a max, then it is bimodal. If there are three intervals with the same high value, it is tri-modal. There may also be no modality, for frequencies that have no real maximum interval.

The mode may be calculated for a population or a sample. M_o must be an observed value in the population or sample. M_o is meaningful for all types of data.

Advantages of each measure of central tendency. The mode is useful because it is the only measure that has meaning for nominal data. The median is especially useful for skewed data, such as home prices or income. The mean has many nice properties that will make it quite useful for probability and investment work.

For Grouped Data, the arithmetic mean of a frequency distribution is normally, the midpoint of each class in the distribution. Then, the sum of the frequency of each class (i.e. the number of units in the class) multiplied by the midpoint of the class / total number of frequencies or unit = the mean. So, the mean of the data grouped by the frequency distribution may be different than the raw data's mean (due to the assumption of the midpoint as being the mean).

The median of grouped data can be only be estimated. First, the class containing the mid-point is located; and then second, that particular class can be reviewed to ascertain the middle value. Thus, the median can be determined for open-ended frequency distributions.

The mode is the class mid-point of the class containing the largest number of frequencies (or values). The mode could be bi-modal with 2 distinct modes, or multi-modal, with numerous major groupings, each of which are equal.

The Skew of a population or sample indicates the relative placements of the mean, median and mode. A symmetrical histogram has two identical sides around a center. Symmetricals are also referred to as "normal" curves or bell shaped curves. Skewness is the degree to which the frequency distribution is non-symmetrical. With a positive skew (the skew is towards the left of a graph) the arithmetic mean is larger than the median or mode. This is because the mean is affected by a few large values. With a negative skew (the skew is towards the right of the graph) the mean is the lowest value of the median and mode. A few low numbers affect the mean. With non-symmetrical graphs, the mean is not a good average to use, and the median may thus be a more appropriate fit.

Nonsymmetrical distributions are skewed, with a non-Normal type of shape. Skewness describes the degree to which a distribution is not symmetric about its mean. A normal will have symmetry about its mean, with the median and mode being equal to the mean.

A positive skew (skew to the right; tail to the right; big distribution to the left) has many small losses and a few large gains. The shape has a long tail to the right, with a few positive gains on the right, and lots of small losses on the left. A negative skew (skewed to the left; tail to the left; big distribution to the right) has many small gains and a few large losses. The tail is to the left, with a few losses to the left and lots of gains to the right. Skewness is calculated similar to variance, by using each observation's deviation from the mean. Also note, that when the skewness equation is calculated, a distribution with a tail to the right and lump to the left will develop a positive number for the skew (and thus, positively skewed), while a tail to the left and lump to the right will have a negative number for the skew (and thus be negatively skewed).

Normal distributions have a bell shape curve, with the mean and the median being the same value. For a negative skew (tail to the left), the mean will be less than the median, and that will be less than the mode. Positive skews will have a mode of less than the median, which is less than the mean. Investors will want a positive skew, because (in statistical terms) the mean return will be above the median. In practical terms, people will have a number of small losses that will not devastate them, but will be able to counter with a few large gains.

Population Proportion (p) is the percentage of observations in the population that have some characteristic. $0 \leq p \leq 1$. A second population would produce a second population statistic. Thus, the population Proportion p^{\wedge} is the percentage of observations in a sample that have some characteristic. $0 \leq p^{\wedge} \leq 1$. Notice the change in notation, p^{\wedge} , to distinguish between a sample and the population, p.

Kurtosis is a peak that is greater or less than a Normal distribution. The distribution is still bell shaped, but a distribution that is more peaked than a Normal is called leptokurtic, with the overall shape being more slender and the tails being higher and fatter than a Normal. Less of peak is called Platykurtic, with the distribution being broader than a Normal and the tails being more flattened than a Normal. Kurtosis provides information about the probability of extreme outcomes, with a leptokurtic distribution (higher peak) have more extreme values than a Normal (or mesokurtic).

Skewness and kurtosis are both used to assess departures from a Normal, with skew measuring the movement away from symmetry and kurtosis measuring the peakedness of the distribution. Skew can visually be seen as a horizontal shift off of a Normal symmetry, while kurtosis is the vertical shift away from a Normal.

Describing Data: Measure of Dispersion. Dispersion measures the variation of a study. Further, dispersion can be used in a comparative way, to ascertain the spread of data between two studies with the same mean. Variability is the amount of dispersion in the data. The variability or dispersion of the data can be measured as a range; a percentile; variance; or standard deviation.

The range is a measure of variability in the sample or population. The range of returns on interval data is simply: **max value – min value**. It is rather limited in nature, since it only

indicates the extreme values, and not the bulk of the observations. It may be calculated for a population or a sample. It is only meaningful for quantitative data. The range has some drawbacks as a measure of variability, but will capture the entire spread of variable in the sample or population.

The mean absolute deviation addresses this by measuring the dispersion of values within the range. $MAD = \sum |X_i - \bar{X}| / n$. The mean deviation (or MAD) is the arithmetic mean of the absolute value of deviation from the mean.

Percentiles and quartiles are very important. A plot box can be visually deployed, allowing for the location of values that can be placed into equally spaced and divided parts of the box. The median is at the 50th percentile, of course, but visually, the median may be off of the center of the box – this would be the case for a positive or negative skew. Typically, the low end of the range would depict the 1st quartile, while the upper end of the range would be the 3rd quartile. An outlier is a value generally outside of the range that is inconsistent with the raw data. Thus, a plot box can be used with out of range data without destroying the validity of the range, itself.

The p%-tile satisfies the condition that no more than p% of the observations are below the p%-tile, and no more than (1-p)% of the observations are above. An example of this is the 50th percentile, where 50% of the sample is above and below one half of the sample. The 90th percentile would be where 90% of the population is below that percentile and 10% of the sample would be above it.

The p%-tile may be calculated for a population or a sample. When the data is sorted into ascending order, the p%-tile is at position $p(N+1)$ for the population, (or $p(n+1)$ for the sample). If $p(N+1)$ is not an integer, interpolate. The p%-tile is not meaningful for nominal data. The 50%-tile is the same as the median. The 25%-tile, 50%-tile and 75%-tile are sometimes called the quartiles, with the lowest quartile labeled Q_1 , the second as Q_2 , and the third or upper quartile as Q_3 . The interquartile range is the range of values between the upper and lower quartiles, and is calculated as $Q_3 - Q_1$. Quintiles are divided into five sets of data, and deciles are divided into tenths of data.

While a median divides a frequency distribution in half, a quartile divides the frequency into fourths; a quintile divides the frequency into fifth; deciles divide by ten; and percentiles divide by hundredths. The median is the 50th percentile; a quartile is the 25th percentile; a quintile is the 20th percentile; and a decile is the 10th percentile.

Population Variance (σ^2 or $\text{Var}(x)$). The variance and standard deviation are the most widely used forms of measures of dispersion, and are the basic definitions of risk in the CAPM and MPT. Variance is the average of the squared deviations around the mean, while the standard deviation is simply the square root of the variance. Note: the variance gets rid of the negative sign due to its squaring effect, and thus is a way around using the absolute value with MAD, or having a negative sign with a dispersion.

Variance is referred to as σ^2 , or variance of x . The variance is a measure of the variability of the sample or of the population. The unit of measurement is x^2 , and not simply as x . As an example, if the variance of sample data on the number of jobs is calculated to be 33.2, then the variance is 33.2 jobs². The deviation of the data values off of mean is squared so as to eliminate the possibility of a negative number.

The equation is the sum of the squares of the difference between the observation and the sample mean divided by the sample size. Variance measures the deviation of the observation off of the sample mean for each observation. Squaring the sum of the differences has the effect of canceling out the minus signs. The equation is as follows:

$$\sigma^2 = \sum (X_i - \mu)^2 / N.$$

Variance is a population parameter, measures the population's variability. It is only meaningful for quantitative data. Variance is sometimes called the second, or centered, moment of the population (and mean is the first moment of the population). $\sigma^2 \geq 0$. The numerator in the variance equation is sometime called the sum of the squared deviations, or more simply, the sum of the squares. Variance will be larger for many observations far from the mean (due to a greater spread in the data). Statistical rules include:

$$\text{Var}(cx) = c^2 \text{Var}(x)$$

$$\text{Var}(x+c) = \text{Var}(x)$$

$$\text{Var}(x+y) \text{ may, or may not, equal } \text{Var}(x)+\text{Var}(y)$$

$$\text{Var}(xy) \text{ may, or may not, equal } \text{Var}(x)\text{Var}(y)$$

Sample Variance (s^2). This is the almost same thing as a population variance, only with the variance of the sample. A sample variance (below) is referred to as “s squared”, and is written as s^2 . The population variance has N units as the divisor, whereas the sample variance has $n-1$ unit as the divisor.

$$s^2 = \sum (X_i - \bar{X})^2 / n-1.$$

The sample variance is a sample statistic, and measures the sample's variability. It is only meaningful for quantitative data. Sample variance will be larger for a sample with many observations far from the mean. $s^2 \geq 0$. Alternative Definition for Variance is: $\text{Var}(x) = E(x^2) - [E(x)]^2$.

Semivariance is often called down side risk, since it measures only the variance of values below the arithmetic mean. The target semi variance, is the variance of returns falling below any arbitrary target or benchmark index level. The importance of these types of variance is that many investors feel that upside variance is not really risk, since there is no actual loss incurred. This is also important due to the fact that pricing returns are not bell shaped, but are typically distributed lognormally. By eliminating the upside risk, an investor then concentrates on only the magnitude of the downside pricing movements.

Standard Deviation. The standard deviation is used to eliminate the squared effect of the answer, so that the unit of measurement is not expressed in squared terms. Also note the difference in population and sample calculations. With samples, the divisor is $n-1$, but is N with populations. This is done for statistical reasons (to produce an unbiased estimator of the population variance). The term $n-1$ is also referred to as the degrees of freedom.

Because the variance involves the sum of the deviations of each observation from the arithmetic mean, the unit of variance is also squared, such as jobs². To get back to the original unit of measurement (i.e. jobs), the standard deviation is used. For a population or sample, the standard deviation is simply the square root of the variance. Once the sum of the deviations off of mean is squared (to remove the minus signs) and then taken to the square root, the unit of measurement will be same as the original observations. The standard deviation then represents the total range off of mean that the observations have been, in either direction from mean.

Standard deviation of a population is: $\sigma = \sqrt{(\sum (X_i - \mu)^2 / N)}$.

Standard deviation of a sample is: $s = \sqrt{(\sum (X_i - \bar{X})^2 / n-1)}$.

The standard deviation can be shortened to: $s = \sqrt{s^2}$

Use the symbol σ_x for the population standard deviation, and the symbol s_x for the sample standard deviation. The standard deviation is a measure of the population or sample's variability, and it is only meaningful for quantitative data. The standard deviation will be larger for a population or sample with many observations far from the mean. Both σ_x and $s_x \geq 0$.

Chebyshev's Theorem. This theorem states that $x\%$ of the population will lie within 1, 2, or 3 standard deviation units of the mean. This theorem has the effect of putting deviations off of mean in terms of "standard units". If: $k \geq 1$, then at least $[1 - (1/k^2)]$ of the observations in a population or sample will lie within k standard deviations of the mean. Chebyshev's inequality gives the minimum percentage of observations within k standard deviations of the mean. $\% = 1 - 1/k^2$ for all $k > 1$. If k is 2, then at least 75% of all observations will lie within 2 standard deviation units of the mean. The result is another way to measure dispersion around the mean.

At least 3/4 of all observations must lie within 2 standard deviations of the mean. At least 8/9 of all observations must lie within 3 standard deviations of the mean. Chebyshev's Theorem states that at least 75% of the spread must lie between the mean plus or minus 2 standard deviations; that 88.9% will lie between 3 standard deviations, and 96% will be between 3 deviations.

The Empirical Rule holds that: if the population or sample is approximately bell shaped (This is called a normal distribution), then 68% of the observations are between 1 standard deviation; 95% are between 2 deviations; and 99.7% will be between 3 standard deviations.

The Coefficient of Variation (CV) is a very important concept. It is a measure of a population or sample's relative variability. We can compare deviations between two or more sets of data by standardizing the absolute dispersion. This is referred to as relative dispersion, and is done through the coefficient of variation. The coefficient of variation is useful to determine the magnitude of variance in the data. It has no units of measurement, and is merely a number. The coefficient of variation expresses how much dispersion exists relative to the mean, and thus is a ratio of the sample's or population's standard deviation divided by the arithmetic mean.

$$\begin{aligned} CV_{\text{pop}} &= \sigma_x / \mu_x \\ CV_{\text{sample}} &= s_x / \bar{x} \end{aligned}$$

In order to compare 3 or more measures of dispersion, one needs to convert to a relative value. The coefficient of variation is that relative value. It measures the ratio of the standard deviation to the arithmetic mean, expressed as a percent. So, $CV = s / \bar{x} (100)$.

On Sampling

Methods of collecting data include direct observation; experimental data collection; and surveys with personal interviews, telephone interviews, or self-administered questionnaires. Estimates are sample proportions used as an estimate of the entire population. A target population is the population that we want to draw inferences about. A sampled population is the actual population that has been sampled. It may differ from the target population due to sampling error. A sampling distribution is the distribution of all possible values that the statistic can assume when computed from samples of the same size randomly drawn from the same population.

Time series data is a data set that runs across time collected at equally spaced time intervals. Cross sectional data involves data collection on various items at one point in time.

There are various methods of sampling. A simple random sample has the same chance of inclusion for all items. Random numbers are generated for each item, thus preventing a bias. Systematic Random sampling is organized in some fashion (by alphabet, 20th invoice, file drawer, etc) and then the same kth number is selected for sampling. This method should not be used if there are patterns in the population mix. Stratified random sampling divides the population into subgroups, called strata, and then a sample is selected in each strata. With this type of stratified sampling, there can be different methods: the proportional sample selects an item in the same proportion as the general population, while the non-proportional sample weights a non-proportional item. With the later method, random chance of non-selection of a strata may be reduced or eliminated. Then, cluster sampling subdivides the population into small geographical units, and then does a random sampling in each of the units. This may reduce the cost of sampling a widely scattered population.

Self-Selected Samples are almost always biased in some fashion, with the conclusions drawn from such samples almost always wrong. Self selection occurs when someone replies to a survey or questionnaire because of keen interest in the subject matter by the sampled population. A person “self selects” to participate in a survey, thus generate a sample population of only people who are interested in a particular issue. The target population will invariably be more indifferent to the issue at hand and will tend to have different outlooks and answers than a self selected sample population.

Simple Random Sampling is a sample selected in such a way that every possible sample with the same number of observations is equally likely to be chosen. An example of this is pulling a ticket stub from a rotating drum. Each piece of paper has an equal chance of being pulled. Stratified Random Sampling is obtained by separating the population into mutually exclusive sets, or strata, and then drawing simple random samples from each stratum. An example of this is separating the population into male versus female; or into various ages, occupations, or household incomes. Inferences about each strata can then be made or between strata. The strata must be mutually exclusive with each member of the

population assigned to only one strata. Then, a simple random sample can be done to generate a complete sample. A Cluster sample is a simple random sample of groups of clusters or elements. This is done when it would be costly to do a simple random sampling on the entire population. It is also useful when the sample is geographically located or clustered together.

Sample Size must be determined in all samples. The larger the size, the more accurate the estimations. With a small sample size, the range of the CI will be large, since there is less precision in the calculations. As the sample size increases, the range of the CI will decrease, and the estimation will become more precise. Also, if the sample size is sufficiently large, we do not even need the distribution to be normal, since the shape of the curve will assume a normal with higher degrees of freedom in a t distribution, due to the CLT.

The Central Limit Theorem suggests that with large samples, the shape of the distribution is close to a normal probability distribution. The shape will become more symmetrical as the sampling size increases. Many statisticians feel that a sample of 30 or more units is large enough to see a normal sampling distribution. There will therefore be a convergence to normality as the sampling size approaches 30 units.

The sampling distribution of same mean \bar{x} computed from samples of size n from a population (and given a population described by any probability distribution with mean μ and variance σ^2) will be approximately normal with mean μ and variance σ^2 / n when the sample size n is sufficiently large. In lay terms, if the sample size is large enough, a distribution will assume a normal shape. The question then is: when is the sample size large enough?

Sampling Error is the difference between the sample and the population that exists only because of the observations that happened to be selected for the sampling. This is the error that is expected to occur when we draw an inference about a population when only a smaller sample of the population has been tested. We are in effect assuming that the sample represents the same values as the entire population, and that assumption is subject to sampling error, wherein the sample does not exactly represent the values of the entire population. Given a fixed sample size, we can state the probability that the sample error is less than a certain amount (ie plus or minus $x\%$; or between $x\%$ and $xx\%$). As the size of the sample increases, the probability of sampling error decreases.

Sampling errors will creep into the results, from time to time. The mean of any sample may not be identical to the population mean. This difference between the sampling mean and the population mean is called sampling error. A sampling distribution of the sample mean involves the means of all possible samples of a given size and the probability of occurrence for each sample mean.

The standard error of sampling means is the difference between the sample mean (\bar{X}) and the mean population (μ). This difference is due to chance and the randomness of the sample itself. The equation is: $\text{Std error} = \bar{X} - \mu$.

The standard error of a sample is: $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, with $\sigma_{\bar{x}}$ being the sample error, and σ being the population standard deviation. When σ is not known, then: $s_{\bar{x}} = s / \sqrt{n}$, where s is sample standard deviation. Sample deviation of the sample is slightly different than for the population, with: $s = \sqrt{(\sum (x_i - \bar{x})^2) / (n-1)}$. Note that the divisor is now $n-1$, instead of n as with a population. As to the distribution population, $X \sim N(u, \sigma^2)$. As to the sample distribution, $\bar{x} \sim N(\bar{x}, s^2)$.

Non-sampling Errors are due to mistakes made in the:

- 1) acquisition of data; or
- 2) due to observations being selected in an improper manner, as is the case with a non-response error. A non-response may result in a biased self selected sample error.
- 3) selection bias occurs where a member of the target population cannot be (or is not) included in the sampled population to be observed.

Increasing the size of the sample does not decrease this type of error. Non-sampling errors are potentially more serious than sampling errors because the changing the size of the sample cannot change the rate of error in the survey.

Examples of Sample biases. Data mining bias can occur when extensive searching occurs through a data-base in search of patterns. The bias exists when several different models are tested on the data until one is proven to be successful, on a back-testing basis. To avoid the problem, the result (or strategy) should be tested on out-of-sample data to see if it holds up there.

Sample selection bias occurs when the data availability leads to certain assets being excluded from the sample. Tracking only companies still in existence, for instance, may lead to survival bias.

Look-ahead bias exists with data that was not available on the test date. For example, the FY book value may not be available for a quarter after the end of the year, when P to BV may be calculated.

Time period bias results for time specific data. This occurs when the time period is not sufficiently large to accurately reflect investment performance. A longer period may be better in such instances, but could also then lead to structural changes occurring across two or more time periods, such that the results may still not hold in the future after the change occurs.

A Confidence level is an important concept to grasp. The point estimate is a value that is used to estimate the population parameter. It is an estimate of the total population. It is very useful when the population is unknown, to begin with. A confidence level is a range of values constructed from the sample so that the point estimate (or data point) occurs

within a range at a specified probability. 95% of the sample means will occur within +/- 1.96 standard deviation units of the population mean (μ). 99% of the sample mean will be within +/- 2.96 standard deviation units of the population mean. The amount outside of the confidence level is basically the tail of the sample. By using Z values on a graph, the confidence levels can be calculated (from Appendix D). The standard deviation of the sampling distribution = std dev. of the pop. mean / square root of the sample size.

A point estimate is a single estimate of a population parameter. A sample mean is an example of a point estimate. Where we take several samples of a population parameter, however, and thereby develop a varying estimate of a population mean with several different sample means, we end with up a range of values of the parameter. A probability distribution can then be generated, and this results in a confidence interval of the population parameter.

Estimations (this is also covered later in this text). Estimators are unbiased, with an expected value equaling the parameter it is intended to estimate; they are also efficient, in that an unbiased estimator is efficient if no other estimator has a sampling distribution of smaller variance; and they are consistent, with the probability of accurate estimates increasing as sample size increases. This is also referred to as “consistent estimators”. [KCK note: these properties lead to the BLUE assumption of classical statistics that is the best linear unbiased estimator].

The z test. A $(1-\alpha)\%$ CI for a parameter has the characteristics of:

point estimate +/- reliability factor * the standard error.

The point estimate is of a parameter (i.e. a value of a sample statistic). The reliability factor is a number based on the assumed distribution of the point estimate and the degree of confidence $(1-\alpha)$ for the CI. The standard error is of the sample statistic providing the point estimate. For example, a 95% confidence interval indicates that in repeated sampling, 95% of the confidence intervals will contain the true population mean. We cannot know whether the particular sample under test contains the population mean, however. The confidence interval of $(1-\alpha)\%$ for the population mean of a normally distributed population with a known variance will be:

$$\bar{x} \pm z_{\alpha/2} \sigma / \sqrt{n}$$

The reliability factor for CI's based on a normal:

For 90% CI, use $z_{0.05} = 1.645$;

For 95% CI, use $z_{0.025} = 1.96$;

For 99% CI, use $z_{0.005} = 2.575$.

Notice the general form in the above equation of point estimate (i.e. \bar{x}) +/- reliability factor (i.e. z) * standard error (i.e. standard deviation / square root of n).

The t distribution can be used in situations where:

- 1) When the sample size is small, but the distribution is still normal.
- 2) When the sample size is large, and the distribution is unknown. The central limit theorem suggests that the sample size may still be approximately normal by using a t factor.

Note that with a t distribution, the tails are fatter than with a normal, and height of the curve is less than a normal. As the degrees of freedom increases (see below), the tails become less fat, and the shape becomes more aligned with a normal distribution. Thus, the importance of the t distribution is that a t can be used with a large size sample but an unknown distribution (or even where the size is small but the distribution is normal), as the higher degrees of freedom allows the t distribution to assume a normal shape, due to the CLT.

The equation for the t distribution is:

$$\bar{x} \pm t_{\alpha/2} s / \sqrt{n}$$

Note that the only differences with a t from a normal are that we are now using 1) the sample standard deviation in the standard error term; and 2) a t table describes the reliability factor instead of a z format.

On the degrees of freedom. The term **n-1** describes the degrees of freedom in estimating the population variance, as calculated in the standard error equation. In a random sample, we assume that the observations are selected independently. The sample mean is computed from a total of n independent observations, and only n-1 observations can be chosen independently of each other.

With a known population variance and a normal, the z tables can be used to calculate a CI. Use the t distribution for situations where the population variance is unknown but the distribution is normal, or for a large size and unknown population variance and distribution shape.

On Probability

The use of frequency distributions are historical in nature – the statistics all study known, past numbers. Probability, on the other hand, studies the likelihood of future numbers. By the use of inferential statistics, known risks can be identified, and possible outcomes estimated. Probability therefore measures the relative likelihood that an event will occur in the future. Three Approaches to Probability include the following.

1. The Classical Approach to probabilities helps determine the probabilities with games of chance. Examples include flipping a coin, rolling dice, taking a chance on the lottery. Objective Probability includes the Classical method, whereby the outcomes are equally likely (by assumption). The outcomes are defined as being mutually exclusive (no two outcomes can occur at the same time) and collectively exhaustive (at least one event must occur in an experiment). The empirical concept is also part of the classical method – the fraction of time the events have occurred in the past are measured, and then the future extrapolation of the past is done for future projections. Empirical probability is estimated from a known set of data. A priori probability draws on logical analysis rather than subjective feelings of odds or probabilities. Empirical and a priori probabilities are considered objective in nature.
2. The Relative Frequency Approach believes that probability is the long run frequency of an outcome or event occurring. Probability is the long run likelihood that an event will occur. A history of past outcomes is necessary to utilize this approach. This approach allows us to link the population of an experiment to the sample population through statistical inferences. This approach is also used for an infinite number of experiments or trials.
3. The Subjective Approach is used when no games of chance are involved and there is no history of past outcomes upon which probability can be estimated. Subjective probability draws on a reference to a personal or subjective thought without reference to data. Probability is the degree of belief that we hold in the occurrence of an event. An example is the probable future rate of return on an asset.

Definitions. A random variable is a quantity whose outcome is uncertain. An outcome is a particular value of a random variable. An event is any outcome or specified set of outcomes of a random variable. A mutually exclusive event can occur only one event at a time. An exhaustive event is an event that covers all possible outcomes.

A Random Experiment is an action or process by which an observation is obtained. The outcome cannot be predicted with certainty. An example includes flipping a coin and seeing if it lands heads or tails.

Exhaustive probable outcomes occur when all possible outcomes are included in the experiment.

Two general rules regarding probabilities are considered to be the requirements of probabilities. Given a sample space, the probabilities assigned to the outcomes must satisfy two requirements: First, the probability of an event is between 0 and 1, or $0 \leq P(E) \leq 1$. Second, the sum of the probabilities of any list of mutually exclusive and exhaustive events = 1.0, or $\sum P(E) = 1.00$.

A sample space (S) of a random experiment is a list of all possible outcomes of the experiment. The outcomes must be exhaustive and exclusive. The sample space and its outcomes is noted as: $S = \{O_1, O_2, \dots, O_k\}$. Probabilities are then assigned to the outcome. An example of a sample space is the flipping of two coins.

A Simple Event (E_i) is a single possible outcome of an experiment. The Probability of a Simple Event, $\Pr(E_i)$ is the relative frequency of the event. It may be written as $P(E_i)$, $\Pr(E_i)$ or $\text{Prob}(E_i)$.

Then, a compound event is considered to be a collection or set of one or more simple events occurring in a sample space. The probability of a compound event is the sum of the probabilities of the simple events that constitute the event. An example is tossing two coins. $A = \text{"Observe at least one H"}$. It is noted as: $A = \{HH, HT, TH\}$. Then, the probability of a compound event is the summation of each separate probability. It is described by the following equation:

$$\Pr(A) = \sum \Pr(E_i)$$

A valid assignment of probabilities must satisfy $0 \leq \Pr(E_i) \leq 1$. The total probability of all simple events of the sample space must equal 100%, or 1.00. The Probability of a Null Event is: $\Pr(\emptyset) = 0$.

The Union (\cup) of event A or B ($A \cup B$) is the event containing all simple events that are in A or B, or both. The union can be A and B, or it can be B and A. This reversal of additions is summarized as: $A \cup B = B \cup A$. The union of events A and B is the event that occurs when either A or B occurs, or both A and B occurs. The probability of the union of A or B is noted as $\Pr(A \cup B)$, and is used to calculate the probability of the union of two events. It is:

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

This result is sometimes called the addition rule. For the special case where A and B are mutually exclusive, then the probability of the union of A and B is as follows: $P(A \cup B) = P(A) + P(B)$.

One should be careful not to over-count the joint probabilities, as the two or more events may no longer be mutually exclusive. (See, the example with the probability of picking either a king or a heart – there is a king of hearts, so one should not count that card twice). The Venn diagram would now show two or more circles with some intersections of circles (to account for the overlap). The equation is $P(A \text{ or } B) = P(A) + P(B) -$

$P(A+B)$, and would be the sum of the two or more individual events probabilities minus the sum of both of the event's probability.

The addition rule for probabilities is also expressed as: $P(A \text{ or } B) = P(A) + P(B) - P(AB)$, and is the probability that either A or B occurs, or both A and B occurs. The equation can be visually displayed by a Venn diagram, with two circles, one for A and another for B. Since A and B can overlap at some points, just adding the probabilities that A or B will separately occur overestimates or double-counts the probability of either A or B occurring. The overlap or intersection between the A and B circles, which is the joint probability of $P(AB)$ must be subtracted from the summation of $P(A)$ and $P(B)$.

An intersection (\cap) of events A and B is the event that occurs when both A and B occur. It is denoted as "A and B". The probability of the intersection of A and B is called the joint probability of A and B. The intersection of events A and B is referred to as " $A \cap B$ " or mostly simply as "AB". It is the combined event containing all simple events that are in both A and B. A and B is equal to B and A, or in math notation: $A \cap B = B \cap A$. The joint probability of A and B is noted as: $\Pr(A \cap B)$. The Probability Rule for Intersections calculates the probability of two events both occurring (i.e. the probability of an intersection of A and B occurring).

$$P(A \cap B) = P(A|B) P(B).$$

This result is sometimes called the multiplication rule. A special case of the multiplication rule exists for the intersection of two independent events.

An intersection also is referred to as a joint probability, and is the likelihood that two events both occur, and is expressed as $P(AB)$, or the probability of A and B. The multiplication rule for joint probability is: $P(AB) = P(A | B) P(B)$, or the joint probability is the probability of A given that B has occurred * probability that B will occur. Also, note that $P(AB) = P(BA)$. Also, $P(A | B) = P(AB) / P(B)$, where $P(B) \neq 0$.

The joint probability of any number of independent events is the multiplication of the probabilities of each independent event occurring, or the joint probability of all events occurring. $P(AB) = P(A) P(B)$; for more than two independent events: $P(AB...N) = P(A) P(B) \dots P(N)$.

The Complement (\bar{A} , A' , or A^c) of event A is the collection of all simple events in the sample space that are not in A. It is expressed as: $(A')' = A$. A complement can be stated as "everything that is not in A" or even shorter: "not A". The probability rule for complements is: $\Pr(\bar{A}) = 1 - \Pr(A)$.

Marginal probabilities are computed by adding across the rows and down the columns of a probability matrix, and are so named because students will often calculate them in the margins of the matrix table. Then, adding all the rows will equal 100%, or adding all the columns equal 100%.

Conditional Probability, $\Pr(A|B)$, is the probability that A occurs, given that B occurs. The probability of an event not conditioned on another event is an example of unconditional probability, $P(A)$, whereas a conditional probability involves the likelihood of an event occurring (A), given the fact that another event has already occurred (B). In math terms: $P(A | B)$.

$$\Pr(A|B) = \Pr(A \text{ and } B) / \Pr(B), \text{ if } \Pr(B) \neq 0$$

Conditional probability asks the question: if B occurs, then what is the probability that A also occurs? The “|” means “given”. So the probability of A occurring, given that B occurs, is noted as: $\Pr(A|B)$. The reverse of the above theorem is also true, where the probability of A, given that B occurs is:

$$\Pr(B|A) = \Pr(B \text{ and } A) / \Pr(A), \text{ if } \Pr(A) \neq 0$$

Conditional Probability of an event occurring, given that the other event has also occurred is part of the Rule of multiplication, and the general rule is that the probabilities of the two events are deduced by simple multiplication of the two individual probabilities. A Tree Diagram can be drawn to visually portray all of the secondary stages of conditional event occurrences. So, if one event occurs, then what is the probability of one of more events occurring from there. The probabilities of all possible outcomes of the tree is therefore 1.00 or 100%, with each individual outcome on any one branch of the tree being a small component of the overall tree.

Unconditional Probability can be expressed by using the total probability rule. For S and not S (noted as S^c), not S is the complement of S. So, $S + S^c = 1$. The abbreviated form of the total probability rule with only two scenarios, of S and not S, is: $P(A) = P(A | S) P(S) + P(A | S^c) P(S^c)$. The general form of the total probability rule is: $P(A) = P(A | S_1) P(S_1) + P(A | S_2) P(S_2) + \dots + P(A | S_n) P(S_n)$. In lay terms, the probability of an event occurring is the weighted average of the probabilities of the event, given the various scenarios. The weights are the probabilities of each scenario, and the events must be mutually exhaustive and exclusive.

The probability weighted average concept can be seen in investment work with conditional probabilities. Conditional expected valuation in investments uses relevant information to make forecasts. The conditional expected value, $E(X | S)$, is the sum of the probability-weighted value of each X, conditional upon a scenario occurring, and is: $E(X | S) = \sum P(x_i | S) * x_i$.

An expected value for investments (or otherwise) can be calculated by using total probability rule, above. Note that a typical tree diagram is now possible, with the probability of each event being sketched out as branches of a tree. The equation is $E(X) = E(X | S) P(S) + E(X | S^c) P(S^c)$. The general form is: $E(X) = E(X | S_1) P(S_1) + E(X | S_2) P(S_2) + \dots + E(X | S_n) P(S_n)$. It then is abbreviated as: $E(X) = \sum E(X_i | S_i) P(S_i)$.

Events A and B are independent if the occurrence of one does not affect the probability that the other occurs. This allows us to prove whether A and B are affected by each other somehow, or whether they are independent of each other. Two events A and B are independent of each other if:

$$\Pr(A|B) = \Pr(A), \text{ or}$$

$$\Pr(B|A) = \Pr(B)$$

Two events are independent of each other if the probability of one event is not affected by the occurrence of another event. So, long as the conditional probability of an event occurring, given that another event also occurs, is the same probability as the event occurring by itself, then the two events are independent of each other. Further, if the marginal probability of the events equals the conditional probability of the events, then the two events are independent of each other.

Two events are independent if the occurrence of one event does not affect the occurrence of another event. In math terms, two events are independent IFF $P(A | B) = P(A)$, or $P(B | A) = P(B)$. When two events are not independent, they are said to be dependent upon each other.

The Special Rule of Multiplication requires that both events be independent, i.e. that the occurrence of one event does not affect the probability of the other event. The equation is $P(A+B) = P(A) * P(B)$. There is no need to subtract the joint probabilities of both outcomes, as is the case with joint probability, due to the combination of outcomes not being affected by each individual outcome.

A description of $\Pr(A|B)$: This measures the probability that an odd number will be rolled from a single die given that same die also rolls a number under 4.

A Description of $\Pr(B|A)$: This measures the probability that a number under 4 will be rolled from a single die given that the same die also rolls an odd number.

The two events are dependent upon each other, since the conditional probability of A is not the same as the probability of A occurring apart of any other occurrence. A and B are related to each other, since the condition of B occurring will limit the number rolled on the die to a number less than the entire set of A.

Events that are not independent may, or may not, be mutually exclusive.

Mutually exclusive outcomes occur when no two outcomes can occur at the same time. Events A and B are mutually exclusive if $A \cap B = \emptyset$. If one event occurs, the other cannot occur. If A occurs, then B cannot occur. The Rule of Addition is important. Events that are mutually exclusive are used in the special rule of addition: The probability

of one event occurring equals the sum of all of the probabilities on the event. This is stated as $P = P(A) + P(B)$. The complement rule to the rule of addition is the probability of the event not happening. It is stated as $P = 1 - P(\text{not } A)$. It is sometimes easier to find out the probability of an event not occurring than the chances of that event's occurrence.

Given a probability $P(E)$, the odds for the event E happening are: $E = P(E) / (1 - P(E))$. The odds against E happening are: $E = (1 - P(E)) / P(E)$. For example, if the probability of an event happening is 33%, then the odds for the event are: .33 / .66, or 1 : 2, described as 1 to 2. The odds against the event are: .66 / .33, or 2 : 1, described as 2 to 1.

Mutual exclusion can be seen in the world of investments. Where probabilities of two or more assets (or other items) are inconsistent with each other, profit opportunities result. This is often the case with arbitrage, where buying and selling of the same item (or of two or more items) will ultimately bring the pricing of the item back to equilibrium and the inability to profit any further from just the trading activity.

Probability trees. Joint probabilities can now be more easily done by multiplying the probabilities of the linked branches.

Bayes Theorem is based upon a clergyman in the 18th century that asked the question "Does God Really Exist?" The prior probability in the initial probability is based upon the present level of available information, while the posterior probability is the revised chances based upon additional information. Bayes' formula is the inverse of the total probability rule, with the occurrence of an event inferring the probability of the scenarios generating it. It is basically the updating of a priori probabilities, given the occurrence of new information. The updated probability of an event given new info = (probability of new info given the event / unconditional Probability of the new info) * prior probability of event. A tree diagram can be used to sketch out all of the possibilities, and such a tree diagram is thus a good depiction of Bayes Theorem. The Theorem holds that:

$$\Pr(B|A) = \Pr(A|B) P(B) / P(A)$$

The multiplication rule of counting allows us to count the numbers of ways that specific numbers of tasks can be performed. The number of ways k things can be done = $n_1 * n_2 * n_3 \dots n_k$, where n is the number of ways each thing (1 through k) can be done. Basically, all possible ways each thing can be done are multiplied together.

A Factorial expression is noted as $n!$, where $n! = n * (n-1) * (n-2) * (n-3) * \dots * 1$. Basically, this is a $3*2*1$ type of multiplication. Factorials involve the assigning of every member of a group of size n to n tasks or slots.

A Permutation is an ordered arrangement of r distinct objects chosen from n objects. A permutation is an ordered listing, counting the number of ways that we can choose r objects from a total of n objects, when the order in which the r objects is listed does matter. The equation is: ${}_n P_r = n! / (n - r)!$. The number of possible permutations is:

$${}_n P_r = n! / (n-r)!$$

A Combination is a unordered selection of r objects from a set of n distinct objects. The number of possible combinations is:

$${}_n C_r = n! / r! (n - r)!$$

This value is sometimes referred to as the binomial coefficient, or binomial and multinomial factorial combinations. The multinomial equation with $n_1 + n_2 + \dots + n_k = n$, is given as: $n! / (n_1! * n_2! * \dots * n_k!)$. The special case of combinations is: ${}_n C_r = n! / ((n-r)! * r!)$, where n chooses r, or n combination r, and where the ordering does not matter.

The appropriate probability counting method. Using the above equations, when the listing matters, the permutation equation is: ${}_n P_r = n! / (n - r)!$. When the order does not matter, use the combination equation: ${}_n C_r = n! / ((n-r)! * r!)$.

The number of possible outcomes must be finite in order to use the above equations. Factorials are used where every member of a group of n size is assigned a task. Multinomial equations are used for counting the number of ways to apply many labels to a group. Combination formulas are used for choosing r objects from a total of n, where the ordering does not matter. Permutations are used where the ordering does matter.

Probability Distributions

Introduction. Probability distributions play a vital role in the calculation and estimation of event occurrences. A probability distribution specifies the probabilities of the possible outcomes of a random variable. Various kinds of distributions exist, both discrete and continuous, and can be Normally distributed, skewed or shaped fatter or thinner than a normal (i.e. kurtotic considerations).

Discrete Random Distributions. This is the entire range of values that can occur based on an experiment. It describes something that may happen in the future, more particularly, the outcome of an experiment and the probability associated with each outcome.

Discrete random variables are results from a random experiment. It stems from Bernoulli's random variable, with the result of an outcome being either p or not p (and, thus binomial in nature). $p(1) = p$; $p(0) = 1-p$. The binomial random variable is defined as the number of success in n Bernoulli trials. Where n is the number of trials, and p is the probability of success for all trials (and is a constant probability), then X is a binomial probability distribution with parameters n and p . $X \sim B(n, p)$. Note the factorial expression of $p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{((n-x)! * x!) } p^x (1-p)^{n-x}$. The probability can be sketched out as a binomial tree. A discrete random variable can only assume clearly separated values. A discrete uniform random variable is a finite, specific value for a variable, and a discrete uniform distribution is the probability distribution generated on the random variable (say, the integers 1 through 8 are possible outcomes).

These discrete variables can then be sorted into a probability to become a discrete probability distribution, $P(x)$. The two probability rules, above, still apply to the distribution: $0 \leq p(x) \leq 1$; and $\sum p(x) = 1$. Statistical averages can be generated from the distribution. A probability function specifies the probability that the random variable takes on a specific value.

The discrete distribution is also finite in nature with specified probability weighted values for each possible outcome. $f(x) = \sum p(x_i) * x_i$, with the probability function of $p(x) = P(X = x)$ and the cumulative distribution function of $f(x) = P(X \leq x)$.

The mean, or μ , is the long run average of a random variable. The mean is also referred to as the expected value, or $E(x)$, and is the probability-weighted average of the possible value of a random variable of: $E(X) = n * p$, and a variance $\sigma^2 = n * p * (1-p)$. The equation for the mean is: $E(x) = \mu_x = \sum x * p(x)$.

The variance measures the spread of the distribution. It is expressed as: $\text{Var}(x) = \sigma_x^2 = \sum (x - \mu_x)^2 p(x)$. Basically, the mean is subtracted from the actual value, and the difference is squared; then, that number is multiplied by the probability of the event

occurring, and all of the products are summed. The standard deviation is simply found by taking the square root of the variance.

With binominal random variable distributions, there are only two possible outcomes, and they are mutually exclusive. True or False are examples of binominal probability distribution. The random variable is the result of the success in the total number of trials. Each trial is independent of each other for every trial – there is no rhythmic pattern, and the probability of each observation never changes. Flipping a coin 100 times is another example of a binominal probability distribution experiment. Tables and graphs can be developed with a two dimensional matrix, with the plot being of the probability of success versus the number of correct answers. As the probability of success approaches 50%, the distribution would become more symmetrical. Further, as the number of trials increases, the distribution also becomes more symmetrical. $\Pr(T) = p$; $\Pr(F) = 1 - p$. The distribution takes the form of:

$$p(x) = {}_n C_x p^x (1 - p)^{n-x};$$
$$u_x = n p.$$
$$\text{Var}(x) = n p (1 - p).$$

A cumulative probability distribution is a probability for an exact number of events, with the distribution being sketched out on a cumulative basis.

The sampling can be done without replacement, so that the member of the population cannot be sampled more than once; or with replacement, so that the member can be sampled more than once.

When a binomial is sampled without replacement from a finite population, and with predetermined outcomes, a Hyper-geometric distribution results. This allows for the probability of each event to change. It does so by having no replacements in the trials – the probability of success increases with each pick of the observation pool. The variable is generated by experiment with N members of a population, with n selections; there are k members of the population with a characteristic; and x sample members with eth characteristic. The equations are rather cumbersome, but a close approximation is $p(x) = n / N <= .05$. The mean can be specifically determined by $u_x = n (k / N) = n k / N$.

A Poisson distribution occurs with very small success probabilities and very large population samples. This is also called the law of improbable events. The probability of an event is extremely small. The poisson variable is produced by experiment from the following: Events occur at an average rate of λ (this is the arrival rate). The occurrences are independent. x is the number of occurrence (arrivals) during the time interval tested. $u_x = \lambda$; $\text{Var}(x) = \lambda$; $p(x) = (\lambda^x e^{-\lambda}) / x!$ When n is large, binomial computations become cumbersome, and a poisson can be used to approximate a binomial distribution. Then, we can use $\lambda = u = n p$.

Continuous Random Distributions. The probability of a discrete single event, $p(x)$, now has no meaning, as the variables under test are continuous in nature. Probability now involves a continuous function. Thus, $\Pr(x) \approx 0$.

A probability density function now describes the continuous random variable. $f(x) \geq 0$, for all values of x , and the total area under the function must equal one: $\int_{-\infty}^{+\infty} f(x) dx = 1$. The probability that x is between two numbers can be found by calculating the area under the curve between the two numbers. $\Pr(a \leq x \leq b) = \int_a^b f(x) dx$. Note that because $\Pr(x) \approx 0$ and the function is continuous, there is no meaningful difference between $\Pr(a \leq x \leq b)$ and $\Pr(a < x < b)$.

The Expected value of a continuous function is: $E(x) = \int_{-\infty}^{+\infty} x f(x) dx$, and the variance is: $\text{Var}(x) = \int_{-\infty}^{+\infty} (x - u_x)^2 f(x) dx$.

A cumulative distribution function defines the probability that a random variable is less than or equal to a particular value (and thus the term “cumulative” is used). $f(x) = P(X \leq x)$.

For uniform distributions taking on any values between a and b , x is equally likely to fall in any range of a particular width. A uniform distribution is the simplest continuous uniform distribution, leading to other continuous types of distributions (such as normals and lognormals). The possible outcomes are not countable, as they are continuous. A uniform distribution describes a straight line on a graph with $f(x)$ on the y axis, as it always describes the probability of equally likely events. With limits of a and b (for the minimum and maximum values), $f(x) = 1 / (b-a)$, for $a \leq x \leq b$; otherwise, 0 . The general proposition is that: $x \sim U(a, b)$. Now:

$$\begin{aligned} f(x) &= 1 / (b-a); \\ u_x &= (a + b) / 2; \\ \text{Var}(x) &= (b - a)^2 / 12; \\ \Pr(x_0 \leq x \leq x_1) &= (x_1 - x_0) / (b-a). \end{aligned}$$

For poisson distributions, x is the waiting time until the next event occurs, and x takes on an exponential distribution, and the equations for the function, mean, variance, and probabilities can be found in standard statistics books.

Normal Probability Distributions. The central limit theorem states that a large number of independent random variables are approximately normally distributed. A normal is a continuous, symmetrical and bell shaped curve (unlike the skewed and leptokurtic types of graphs previously identified). A Normal is a continuous random distributions with an infinite number of values within a specified range. It is continuous probability curve, with the tails never touching the X axis. $X \sim N(u, \sigma^2)$, where X is normally distributed with mean u and variance σ^2 . N has a skewness of zero, and a defined kurtosis of 3. The excess kurtosis is zero. The mean, mode, and median are all the same values, centered on the peak of the curve, due to the bell shape of the Normal. This is a symmetrical

distribution, and has a smooth run-off from the central value. In theory, the curve extends to infinity, and never touches the edge of the graph.

Standard Normal Probability Distributions are those distribution patterns that are limited to standard deviations between 0 and 1.00. x can take on any real number. x assumes a normal shape, with mean of u and variance of σ_x^2 . This is stated as:

$$x \sim N(u_x, \sigma_x^2).$$

Increasing the mean will not affect the shape of the Normal, but only shift the entire distribution to the right. Increasing variance will not location of the distribution, but will increase the thickness of the tails (makes the distribution fatter).

Adding variables to the Normal. Constants can be added to a normal, and just moves the mean to the right.

$$x + c \sim N(u_x + c, \sigma_x^2).$$

The constant can also multiply the distribution, and has the effect of moving both the mean and increasing the overall size of distribution, while still keeping the Normal shape.

$$cx \sim N(c u_x, c^2 \sigma_x^2).$$

A univariate distribution describes the probabilities of a single random variable. A multivariate distribution specifies the probabilities for a group of related random variables. With multivariate distributions, correlations must be specified between the values of the random variables. For example, a multivariate distribution of returns will have the mean rate of return, the variance of the return, and all pair-wise return correlations between random variables. This is opposed to univariate distributions, where no correlations can be specified due to there being only a single random variable.

Two independent normal variables can be added together to generate:

$$x + y \sim N(u_x + u_y, \sigma_x^2 + \sigma_y^2).$$

A Z value is the distance between a selected value, x , and the mean of the distribution, u , divided by the standard deviation. So, $z \sim (x - u_x) / \sigma_x$. z can be any real number. $z \sim N(0, 1)$.

The total area of a normal “Standard” curve is 1.00. The area under a normal curve within ± 1 standard deviation is .6826. The area within 2 deviations is .9547, and the area within ± 3 deviation units is .9974 – essentially the entire curve. The center of the curve is, of course, at .500. Z values are basically areas of the normal curve, so by solving for Z from the above formula, and then adding or subtracting from the center of a curve (at .50), one can calculate the area of the curve covered by the Z value, along with the associated standard deviation of the distribution.

Since the center of the distribution, c , is at 0.5 standard unit, the critical value of z , will be $\Pr(z > z_c) = c$.

Approximating a binomial. If n is large, we can use a Normal to approximate a binomial with $u_x = n p$ and $\text{Var}(x) = n p (1 - p)$. Since the binomial is discrete while a normal is continuous, we use the normal if $p(a \leq x \leq b)$.

Confidence Intervals. Probability statements about a random variable often use confidence intervals built around point estimates. For example, within a 95% confidence level, x will be within 1.96 standard deviation units. This is: $P(X) = \bar{x} \pm 1.96 s$. The practical part of this is that we expect the random variable to fall within the confidence interval 95% of the time. We have no way of knowing however whether this sample is within that 95%. Note that at the 95% CI, the outlier range lies 2.5% on either side of the tails.

The Standard Normal. A standard normal distribution has a $u = 0$ and $\sigma = 1$, or a mean of zero and one standard unit of deviation. As deviation increases (say, for example, to 2 standard units, or $\sigma = 2$), the distribution fattens out and the peak is lowered, since the deviation about the mean is much greater.

To standardize a random variable: $Z = (\bar{x} - u) / (\sigma / \sqrt{n})$, where Z is the notation for a standard normal random variable, and X is a Normal random variable. We can then answer all probability statements about X in standard normal terms and probability tables for Z , and a cumulative density function of a Normal is denoted $N(x)$ or $N(z)$.

Relationship to Lognormal. A random variable, Y , follows a lognormal distribution (i.e. skewed to the right, with a long right tail and peak to the left, with a “positive” skew) if its natural logarithm, $\ln Y$, is normally distributed. This relationship between normal and lognormal is important in investment work because most stock returns are lognormal distributed, with lots of small losses and a few huge gains.

Sampling Distributions. Sample statistics are considered to be random variables because they are calculated based upon a sample that is chosen at random. A sampling distribution is a probability distribution of a sample statistic. For a simple random sampling, if n is large enough, the CLT would suggest that the sample of a population would approach that of a normal.

$$\bar{X} \sim N(u_x + u_y \sigma_x^2 + \sigma_y^2).$$

If the population is distributed close to Normal, then the approximation of the sample to a Normal would work better, of course. If $X \sim N$, then $\bar{x} \sim N$. The approximation also improves as n increases. $E(\bar{x}) = u_x$, or the expected value of the sample mean will be the population mean, as the sample approaches a normal. As n increases, $\text{Var}(\bar{x})$ decreases.

However, if the sampling is done from a finite population without replacement, a correction factor has to be applied. The correction can be ignored when the sample size is very small in relation to the population size, usually at: $n / N \leq .01$. True variance of the sample, with the correction factor is:

$$\text{Var}(\bar{x}) = \sigma_x^2 / n * (N - n) / (N - 1)$$

Standard error is the standard deviation of the estimator used for a population parameter. An estimator with a small standard deviation is preferred, as we will be closer to the true value of the parameter. A relatively small standard error is referred to as an estimator with precision. Many times, the exact standard deviation of the estimator cannot be calculated and must be estimated instead.

Because of the CLT, the probability of the estimator will assume a normal in large sample sizes with:

$$\hat{p} \sim N(p, p(1 - p) / n).$$

The approximation works better as $n \rightarrow \infty$ and as $p \rightarrow .5$. The $E(\hat{p}) - p$. As n increases, Variance of the estimator decreases.

Sometimes, two samples from the same population will be compared (or a two separate populations will be compared). In such cases:

$$\bar{x}_1 - \bar{x}_2 \sim N(u_1 - u_2, \sigma_1^2 / n_1 + \sigma_2^2 / n_2).$$

Approximation works better when the population is close to Normal, and as n_1 and $n_2 \rightarrow \infty$. $E(\bar{x}_1 - \bar{x}_2) = u_1 - u_2$. As n increases the standard error of $\bar{x}_1 - \bar{x}_2$ decreases.

Estimation. A point estimator is a rule or formula that gives a numeric estimate of a population parameter. For instance, \bar{x} is a point estimator for u .

An interval estimator is gives a range of numbers in which the population parameter is likely to fall within.

An unbiased estimator. For a population parameter of θ and a point estimator is θ^w , if the $E(\theta^w) = \theta$, then θ^w is said to be an unbiased estimator.

A consistent estimator. The point estimator, θ^w , of the population parameter θ is said to be consistent if, for any ϵ :

$$\lim_{n \rightarrow \infty} \Pr [|\theta^w - \theta| < \epsilon] = 1.$$

Confidence coefficient of an interval estimator is the probability that an interval estimator will contain the true value of the parameter being estimated. The coefficient must be between 0 and 1. This is also known as the confidence level.

Estimation error and levels of confidence. This is the difference between the value of the point estimator and the true value of the population estimated. We would like for the estimation error to be small, but may never know in practice the size of the error. If the point estimator θ^w is distributed normally and is unbiased, there is a 95% probability that the estimation error is $< 1.96 \sigma \theta^w$. This is referred to as 95% error bound.

The 95% level of confidence when using \bar{x} to estimate μ_x is $1.96 \sqrt{\sigma_x^2 / n}$. Increasing n will normally reduce the size of the estimation error.

The confidence interval can now be introduced. Assuming that θ^w is an unbiased point estimator for the population parameter θ , and σ is distributed normally. The interval estimate for θ with a confidence coefficient of $1 - \alpha$ will generate the following confidence interval: $\theta^w \pm z_{\alpha/2} \sigma \theta^w$.

The $1 - \alpha$ confidence interval (CI) for determining the mean is:

$$\begin{aligned} \mu_x &= \bar{x} \pm z_{\alpha/2} \sqrt{(\sigma_x^2 / n)}; \text{ or} \\ \mu_x &= \bar{x} \pm z_{\alpha/2} (\sigma_x / \sqrt{n}); \end{aligned}$$

Increasing α will normally make the CI smaller or narrower. Increasing the sample size will make the interval narrower. The endpoints of the interval are called the upper and lower confidence limits.

Choosing a sample size. Since sampling is expensive and time consuming to do, often the sample size is just big enough to generate the widest confidence interval that researchers are willing to tolerate. Smaller α 's, narrower CI's, and larger σ_x^2 will require larger samples. When variance is not known, then the variance can be estimated using the empirical rule.

$$n \geq [z_{\alpha/2} \sigma_x / B]^2$$

where, B is the CI radius, namely: $z_{\alpha/2} (\sigma_x / \sqrt{n})$.

The Chi-square test. For multinomial problems, the experiment has the following characteristics:

- 1) n identical trials
- 2) k possible outcomes (often called cells)
- 3) $\text{Prob}(\text{outcome } i) = p_i$
 $0 \leq p_i \leq 1; \sum_{i=1 \text{ to } k} p_i = 1.$
- 4) Trials are independent
- 5) Define n_i = number in sample that "land" in cell i

$$E(n_i) = p_i n.$$

Hypothesis testing (described in the following section) can be conducted regarding p_1, p_2, \dots, p_k . $H_0: p_1 = \theta_{0,1}, p_2 = \theta_{0,2}, \dots, p_k = \theta_{0,k}$; H_A : at least one hypothesized value is incorrect. The test statistic is: $\chi^2 = \sum (n_i - n \theta_{0,i})^2 / n \theta_{0,i}$. The test statistic should be close to zero under null hypothesis. The rejection region is: reject if $\chi^2 > \chi^2_{(k-1)}$. The test statistic is only approximately distributed χ^2 . When the expected cell counts are under 5, the approximation doesn't work well.

A contingency table can be developed showing a two-dimensional table matrix with the values taken on by two qualitative variables in a sample of size n . If the classifications are independent, $E(n_{ij}) = np_i p_j$. If p_i and p_j are unknown, we can approximate them using r_i/n and c_j/n .

Hypothesis Testing

Developing the Testing Procedure. It is often not feasible to study all items in a population, so it is common to develop a sample and then determine whether the empirical evidence supports the statement or hypothesis. Inferential statistics have two broad fields of study: estimation and hypothesis testing. As seen in prior sections of this outline, estimation is done through confidence levels and probability distributions. A hypothesis is a statement or proposition made about one or more populations. Hypothesis testing ascertains whether evidence supports a hypothesis. It involves the positing of a value for one or more population parameters, and then examining how well the data conforms to the hypothesis. The hypothesis must be formed before examining the data. A sample should only be used for one hypothesis test.

Components or steps of a Hypothesis Test. Note that different texts will vary in the exact steps or sequences to follow, but generally cover the following:

1. State the Null Hypothesis and Alternative Hypothesis
2. Select the level of significance.
3. Identify the test statistic and its probability distribution.
4. Use the Decision Rule / Rejection Region
5. Decision to reject or not reject (statistical decision); and the economic decision.

For a short discussion on the steps, first, state the hypothesis. A failure to reject the null does not automatically prove the proposition. We have just failed to disprove the null. The alternative hypothesis is accepted if the null is proved false. The null hypothesis, H_0 is the hypothesis to be tested. H_a is the alternative hypothesis that is accepted when the null is rejected. For example, $H_0: \theta = \theta_0$; $H_a: \theta \neq \theta_0$. H_0 and H_a should be stated in such a way that all possible events are provided for by the test. If H_0 is equal to, then H_a should be \neq . If H_0 is \geq , then H_a is $<$.

Second, select a level of significance. This is the probability level for rejecting the null proposition. It is the risk of rejection when the null is really true. .05 for consumer tests, .01 for quality tests, and .10 for political polling are all common. This would be 5%, 1% and 10% respectively.

Third, compute the test statistic. The test statistic is a value used to determine whether to reject the null. The difference between the population and the mean may be statistically significant by finding the standard deviation from the mean.

Fourth, use the Decision Rule. This is the condition under which the null is rejected or not rejected, and will depend upon the level of significance. The critical value (the confidence level) is the critical value.

Fifth, Make a decision to reject or not reject. Instead of accepting, some would say that the proposition is not rejected. This is due to the possibility of a type I or type II error

still existing. Instead of proving a statement or hypothesis, it would be better to say that the proposition cannot be disproved.

Step 1: The null and alternative hypothesis. The null hypothesis (H_0) is the hypothesized value for the population parameter(s). It is typically expressed as $H_0: \theta = \theta_0$. The null receives the benefit of the doubt, and is assumed to be true throughout the test until the final decision is made to reject or not reject. We cannot prove whether the null is true or false. We can only say whether data is consistent with the null hypothesis. Thus, we do not accept the null, we fail to reject the null.

The Alternative hypothesis (H_A or H_1) is the hypothesis that the alternative is true if the Null is False. There are three possibilities:

$$H_A: \theta > \theta_0$$

$$H_A: \theta < \theta_0$$

$$H_A: \theta \neq \theta_0$$

The first two possibilities are called “One-Tail Tests”. The third is called a “Two-Tailed Test.” The Alternative Hypothesis is usually the interesting result.

Type I Error and II Errors. A type I error rejects a true null hypothesis, whereas a type II error fails to reject a false null hypothesis. These are known as Alpha and Beta probabilities. In practice, we will not know if we made an error, and only calculate the probability of a correct conclusion, within a certain level of confidence. Four possible outcomes occur with H_0 testing. A type I error occurs where we reject a true null hypothesis. In a type II error, we do not reject a false null. No errors are made with two other possible decisions: where we reject a false null, and where we do not reject a true H_0 .

Usually, where an attempt is made to reduce the probability of making a type I error, the likelihood of a type II error increases. The chief way to reduce the probability of making either a Type I or a Type II error is to increase the sample size, n . This has the effect of reducing the standard error term.

Step 2 – State the level of significance (one or two tails). A one-tail test is either to the left or right of the normal distribution curve. A two-tail test involves both sides of the probability distribution. Rejection occurs if the finding is beyond 1.96 (at the 95% confidence level), and beyond 2.58 for 99% confidence level. Note: the critical values are different for a one tail versus a two tail. The previously noted significance level (α) can be used in Hypothesis testing, with $\Pr(\text{Reject Null} \mid \text{Null is True}) = \alpha$; and $0 \leq \alpha \leq 1$. All else equal, we'd prefer for α to be small.

The two-sided hypothesis test (or two-tailed test) would be to reject H_0 in favor of H_a if evidence exists that the population is either larger or smaller than θ_0 , and would be stated as: $H_0: \theta = \theta_0$; $H_a: \theta \neq \theta_0$. A one-sided test (or one-tail test) is to reject the null if evidence suggest that the population parameter is $>$ (or $<$) H_0 . A one-side hypothesis test

would be stated as $H_0: \theta \leq \theta_0$; $H_a: \theta > \theta_0$. With a two-tail test of $z_{0.025}$ (i.e. a level of significance of 0.05), there will be a 95% confidence interval that the true population will be within that area of the distribution that is not rejected.

The level of significance tells us how much evidence is needed to reject the null. Alpha, α , represents that level of significance. A 5% probability of rejecting the null, H_0 , would be shown as $\alpha = 0.05$. The level of significance is really the probability of making a type I error (rejecting a true H_0).

The Power of the Hypothesis Test. Some texts discuss the “power” of a Hypothesis Test, as opposed to the significance level. $\Pr(\text{Don't Reject the Null} \mid \text{Null is False}) = \beta$, and $1 - \beta$ to be the power of the hypothesis test. $0 \leq \beta \leq 1$, thus the power must be between 0 and 1 (inclusive). All else equal, we'd prefer for β to be small, and for the power to be large. Decreasing the significance level will also decrease the power. Typically, we will select a significance level, keeping in mind that our selection also affects the power.

Step 3 – the Test Statistic / Probability distribution. This is a sample statistic that has one distribution under the Null Hypothesis, and another distribution under the Alternative Hypothesis. Most discussions of Hypothesis testing start with the Z Test Statistic, which we have previously encountered in discussions of probability theory. The Z statistic is used for Normal distributions (or close to Normal), or where the sample size is so large that the distribution will approach that of a normal through the CLT.

If we wish to test the Null Hypothesis, $H_0: \theta = \theta_0$, versus any of the three alternative hypotheses, and we have some estimator, $\theta^\wedge \sim N(\theta, \sigma_\theta^2)$, we will typically use the test statistic:

$$Z = (\theta^\wedge - \theta_0) / (\sigma_\theta / \sqrt{n}); \text{ or:} \\ Z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$$

Note that if the Null Hypothesis is true, this test statistic is distributed Standard Normal. If the Null Hypothesis is false, this statistic is still distributed Normal, but not Standard Normal.

The denominator of the Z test, σ_θ , is the standard error of the sample statistic, which we previously encountered in discussions of probability distributions. The standard error for the population is $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, and the standard error of the sample is $s_{\bar{x}} = s / \sqrt{n}$. Since we are testing the null hypothesis at the point of equality, the value of the parameter θ_0 or μ_x will be zero, in many instances.

The confidence interval displays that part of the probability distribution being evaluated. This establishes the ultimate range of values that might contain μ , within the designated confidence level (and again, never being certain that the CI does in fact contain μ).

$$CI = \theta^\wedge \pm Z_{\alpha/2} \sigma_x / \sqrt{n}$$

Step 4 – the Rejection Region. This is a range of real numbers where the test statistic is unlikely to fall if the Null Hypothesis is true. On a bell shaped curve, the rejection region is that of α , and it will be on the far end of one or both tails of the curve.

If the test statistic takes on a value in the rejection region we reject the Null Hypothesis. The Rejection Region may depend upon the Alternative Hypothesis. The rejection region for Z Test Statistic is as follows. If the Null Hypothesis is true, the test statistic, Z, is distributed Standard Normal, and thus should be near zero. Thus:

$H_A: \theta > \theta_0$: Reject if $Z > z_\alpha$

$H_A: \theta < \theta_0$: Reject if $Z < -z_\alpha$

$H_A: \theta \neq \theta_0$: Reject if $Z > z_{\alpha/2}$ or $Z < -z_{\alpha/2}$

Stating the probability of correctly rejecting the null Hypothesis is the power of the test, or rejecting H_0 when it is false. The decision rule involves the level of significance used to reject H_0 . At the .10 level of significance, there is some evidence that the null is false. If we reject the null at .05, there is strong evidence that the null is false. At 0.01, there is very strong evidence of rejecting a false null. We can then calculate the critical or rejection point for the test statistic. For a two sided test at the .05 level (or $z / 2$), H_0 is rejected if $z_{0.025} > 1.96$ or < -1.96 . For a one-sided test at 0.05, reject H_0 if $z_{0.05} > 1.96$, for example.

Step 5 – Reject or Not Reject. For a given test, there is always some α large enough to cause a rejection, and some α small enough not to cause a rejection. On the terminology, H_0 is not “accepted”, it may be “failed to be rejected”. The area of not rejecting is the area under the bell shaped curve with the tails (or one tail) cut off at $z > 1.96$ or < -1.96 .

If the test statistic $>$ the rejection point, H_0 is rejected in favor of H_a . This is the statistical decision. The economic decision involves broader issues, such as whether a strategy that is statistically significance in a hypothesis test works once all economic or business factors are considered (such as including trading expenses and tax impacts in th investment arena).

P Values. This is the probability of observing a sample value as extreme or more extreme than the value observed. The P value is another at looking for a Type I or II error, and it is basically the likelihood that the H_0 is not true. If P value = .10, there is some evidence that the H_0 is not true. If P = .05, there is strong evidence. If P = .01, the evidence is very strong. If P = .001, the evidence that the H_0 is not true is extremely strong. Most software packages report a p-value for tests.

The p value is the smallest level of significance at which H_0 can be rejected. With a p value approach, we do not need to establish a rejection point and the corresponding level of significance. Instead, the p value indicates that the small level of significance (as in 0.01, for example), at which the H_0 is rejected. It does not create an artificial level of significance at which we test H_0 . The p value approach may therefore be much better than the traditional H_0 testing, as it indicates the level of confidence that is needed at

which the null hypothesis will be rejected, instead of creating an arbitrary boundary line of rejection or not rejection.

Discussions of One Population Inferences. In many instances, the sample is not distributed normally. With inferences involving one population, testing procedures other than the z statistic are often utilized.

The Student's t Distribution. The t distribution is a continuous distribution that is symmetrical and bell shaped with a mean of 0, median of t, and $E(t) = 0$. A t distribution is more spread out than the normal. It has a standard deviation of > 1 (versus 1.00 for a normal), and has a greater probability of outcomes distant from the mean (due to it being more spread out than a normal. The t statistic has an additional parameter called a "degree of freedom" or df (≥ 1). We write that a variable is distributed t(df), or t_{df} . For testes with two or more degrees of freedom:

$$\text{Var}(t(df)) = (df / df - 2)$$

The z statistic is typically used when the population is close to a Normal, or the size of the sample is large (so that the sample approaches that of a Normal due to the CLT). When $\text{Var}(\bar{x})$ is unknown, we substitute the standard deviation of the sample for the standard deviation of the population. So, in the denominator of the test statistic, as we move from a z to a t test, we go:

$$\begin{aligned} \text{From } Z = Z &= (\hat{\theta} - \theta_0) / (\sigma_{\theta} / \sqrt{n}), \\ \text{and to } t_{df} &= (\hat{\theta} - \theta_0) / (s_x / \sqrt{n}). \end{aligned}$$

Substituting S_x for σ_x thus causes the distribution to switch from z to t(n-1). Note, however, when n is large, df will be large, and the distinction is unimportant. As df increases, the t becomes a z, and approaches the general shape of a Normal. When $df > 30$, the t and z are very close. So, whether we use a z or a t statistic depends upon: the distribution being close to normal; the size of the sample; whether we know σ_x . The t must be used for non-normal distributions unless n is large enough that a close to normal distribution can be assumed; or where the standard deviation of the population is unknown (almost always). With an unknown variance or for a small sample size, the t statistic is the theoretically correct test to use. From above, the t statistic is: $t_{n-1} = (\bar{x} - \mu_0) / (s / \sqrt{n})$, for a t statistic of n-1 degree of freedom. Where the sample is sufficiently large that the CLT approximates a normal, the t test is still appropriate, but the z test can then be used because the difference between a t and a z will be quite small.

The rejection region for the t statistic is:

$$\begin{aligned} H_A: \mu_x > \theta_0 &: \text{Reject if } t > t_{\alpha}(n-1) \\ H_A: \mu_x < \theta_0 &: \text{Reject if } t < -t_{\alpha}(n-1) \\ H_A: \mu_x \neq \theta_0 &: \text{Reject if } t > t_{\alpha/2}(n-1) \text{ or } t < -t_{\alpha/2}(n-1) \end{aligned}$$

The Confidence Interval for a t statistic becomes:

$$\bar{x} \pm t_{\alpha/2, n-1} s_x / \sqrt{n}$$

Where the population variance is actually known (very unlikely), the z test is theoretically correct. From the prior discussions, a z statistic is:

$$z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$$

A sample size can be chosen to form a confidence interval. This is similar to the prior discussion on selecting a sample size for use in with a normal distribution. Smaller α 's require larger samples, as do narrower confidence intervals. Generally, as more precision is called for either in terms of the degree of confidence required or the size of the confidence interval, the size of the sample must increase.

On the Chi-Square. In tests of a single normally distributed population, we can use the chi-square test statistic. This test statistic is asymptomatic, and is composed of a family of distributions, just as with the t distribution. A different distribution exists for each possible degree of freedom, $n-1$. The chi-square is bounded by zero (it does not take on negative values, and the hump on the distribution terminates at zero). The null hypothesis is: $H_0: \sigma^2 = \sigma_0^2$; $H_a: \sigma^2 \neq \sigma_0^2$, for a two tail test.

Describing Populations of Nominal Data. Where a sample is composed of categorical values, no means or variances calculations can be done, since an arbitrary value is assigned to each piece of data. For example, if we assign certain numbers (1 through 5) for certain brand products (5 different cereals, for instance), this would be a sample of nominal data.

All we can do with this type of information is to describe the population through a proportion, p . Then we can calculate the probabilities based on binomial experiment (success or failure outcomes). The formula for estimating the population proportion is as follows:

$$\hat{p} = x / n$$

Where x is the number of successes in the sample and n is the sample size. The sampling distribution of \hat{p} is approximately normal with the mean p and the standard deviation $\sqrt{p(1-p) / n}$. The sampling distribution, and the test statistic, is:

$$z = (\hat{p} - p) / \sqrt{p(1-p) / n}$$

The confidence interval estimator is:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p}) / n}$$

These equations are valid so long as $n * \hat{p}$ and $n(1 - \hat{p})$ are greater than 5.

The sample size depends upon the confidence interval and the size of the interval that will be produced. When the confidence interval is specified, we can then determine the value of $z_{\alpha/2}$, above. The width is determined by the value of the quantity following the plus or minus sign. The sample size to be used is as follows:

$$n = \left(\frac{z_{\alpha/2} \sqrt{p^{\wedge}(1-p^{\wedge})}}{W} \right)^2$$

Where W is the width of the proportion (it also has been identified in other sections of this outline as B , or the radius of the confidence interval). The variable p^{\wedge} is not known, due to the sample not yet being taken, so we have to estimate. If we have no knowledge of the approximate value of p^{\wedge} , then we use .5 for p^{\wedge} , since the maximum value of p^{\wedge} will be 0.5. This then generates a conservative value for n . If we have some idea of the value of p^{\wedge} , then we can use that value to determine n .

When data is interval, the parameters of interest are the population mean μ and the population variance. The student t distribution is used to test and estimate the population mean when the standard deviation is unknown. The chi-square distribution is used to make inferences about the population variance. When the data is nominal, the parameter to be tested and estimated is the population proportion, p . The sample proportion follows a n approximate normal distribution, which then produces the test statistic and the interval estimator. The sample size can also be determined to estimate the proportion.

Two Population Inferences. Often, it is desirable to compare two different populations. For simple random samples that are independent, the CLT suggests that $\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/n)$. Approximation works better when populations are distributed close to Normal, and as $n_1, n_2 \rightarrow \infty$. The expected value of the difference in populations is: $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$. As n increases, the standard error decreases.

When testing: $H_0: \mu_1 - \mu_2 = \theta_0$, if n 's are large enough, populations are normally distributed, or σ_x 's are known, the z statistic can be used. If the sample size is small, the populations are not Normal, or the σ_x 's are unknown (which is usually the case), the t statistic should be employed.

Testing unknown with assumed equal deviations. The next question is whether $\sigma_1 = \sigma_2$. Typically, this is unknown, and should be tested, with a hypothesis test of $H_0: \sigma_1 = \sigma_2$ using a different sample. If you assume $\sigma_1 = \sigma_2$, and they are not equal, your methodology will be flawed. But if $\sigma_1 = \sigma_2$, and you don't require them to be, the methodology will be valid, but the power will be reduced unnecessarily. Two populations with independent random samples will generate a t test based on a pooled sample. The test statistic becomes:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\theta_0 - \theta_1)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

The number of degrees of freedom is $n_1 + n_2 - 2$. Stating the null is in the form of: $H_0: \mu_1 - \mu_2 = 0$; $H_a: \mu_1 - \mu_2 \neq 0$, for a two sided test. Rejection Regions:

$$\begin{aligned}
H_A: \mu_1 - \mu_2 > \theta_0 &: \text{Reject if } t > t_{\alpha, df} \\
H_A: \mu_1 - \mu_2 < \theta_0 &: \text{Reject if } t < -t_{\alpha, df} \\
H_A: \mu_1 - \mu_2 \neq \theta_0 &: \text{Reject if } t > t_{\alpha/2, df} \text{ or } t < -t_{\alpha/2, df}
\end{aligned}$$

The CI becomes: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{(S_1^2/n_1 + S_2^2/n_2)}$.

For unknown variance assumed to be unequal. The same general equation is used, but the becomes more involved to calculate.

Interpreting tests of $H_0: \mu_1 - \mu_2 = 0$ / paired difference testing. We often test $H_0: \mu_1 - \mu_2 = 0$ to see if there is a difference between two populations. The test can identify whether there are differences between the populations.

For a paired-difference test, a new random variable is defined as $d=A-B$. Now we can test $H_0: \mu_d = \theta_0$. The pairs must be matched prior to the test. d will be distributed normally if the two individual populations (A & B) are distributed normally. A matched pair is sometimes called a block.

For paired observations, we calculate the standard error of the difference in the two samples, based on the t statistic. In most cases, we test to see if the results are statistically different from zero for the mean. The t test is now based on:

$$t = (\bar{d} - \mu_{d0}) / s_{d \text{ bar}}$$

where, n is the number of paired observations, \bar{d} is the sample mean of the differences, and $s_{d \text{ bar}}$ is the standard error of \bar{d} .

On the F Distribution. This is another important test that is frequently encountered in regression analysis and hypothesis testing involving a comparison of two population or scenarios. Assume independent standard normal variables $\chi^2(n)$ and $\chi^2(m)$, then:

$$(\chi^2(n) / n) / (\chi^2(m) / m) \sim F^{n, m}$$

Tests between variances of two populations use the F distribution. This is a family of distributions that are asymptotical and bounded below by zero, just as with the chi-square. Each F has two degrees of freedom. $F = s_1^2 / s_2^2$, with $df_1 = n_1 - 1$, and $df_2 = n_2 - 1$. The F test, like the chi square, is not robust in the violations of its assumptions.

On non-parametric tests. Parametric tests have parameters whose validity is dependent upon assumptions. The above processes are examples of parametric tests. A nonparametric test does not concern itself with the validity of parameters. Such tests make minimal assumptions about the population. Nonparametric tests are done when distributed assumptions of parametric tests are not appropriate (as in non-normal distributions); when the data consists merely of ranks; or when a parameter is not at issue. Often, a nonparametric result is stated alongside a parametric test conclusion.

Linear Regression Analysis

Introduction. Regression Analysis is the relationship of two or more variables that allows estimation of one dependent variable (y) based on the value of other variables ($x_0, x_1, \text{ etc.}$). Correlation analysis is a grouping of techniques measuring the strength of association between two variables. A scatter diagram is a chart that visually portrays the observations between two variables. The plot is shown on a typical $x - y$ coordinate graph. The dependent variable is the variable being estimated, while the independent variable is a known variable providing the basis for the estimation.

Simple Regressions. Assuming only one independent variable ($k=1$) and one dependent variable, y , the ordinary least squares method will generate a best fit on a scatter plot by minimizing the sum of the vertical distance between the central Y line axis and the probable and predicted Y axis. A linear equation of $y^{\wedge} = \beta_0 + \beta_1 x_1$ generates a regression line, where y^{\wedge} is the predicted value of Y , β_0 is the Y intercept, β_1 is the slope of the line, and x_1 is the value of the independent variable. The standard error of estimate measures the scatter or dispersion of the observed values around the regression line, and is described as e_i (or more correctly, ε_i). This is similar in concept to the standard deviation. ± 1 standard error of estimate within 68% confidence; ± 2 standard error within 95.5%; and ± 3 units are within 99.5%.

The classical assumptions of linear regression include:

- 1) A linear relation exists between the dependent and independent variables.
- 2) The independent variable, Y_i , is not random.
- 3) The expected value of $\varepsilon_i = 0$; or $E(\varepsilon_i) = 0$.
- 4) The variance of the error term is the same for all observations; or $E(\varepsilon_i^2) = \sigma_{\varepsilon_i}^2$; $i = 1 \dots n$.
- 5) The error term is uncorrelated across observations; Thus, $E(\varepsilon_i, \varepsilon_j) = 0$, for all $i \neq j$.
- 6) The error term is normally distributed; or, $\varepsilon_i = N$.

Assumptions 2 and 3 are needed in order that correct estimates are obtained for β_0 and β_1 . Assumptions 4, 5, and 6 are needed to determine the distribution of β_0 and β_1 . Assumption 4 is the homoskedasticity assumption, while assumption 6 allows us to test the regression model for validity.

Population Covariance (σ_{xy}). At the heart of linear regression analysis is the extent to which independent variable moves with or explains the co-movements of the dependent variable, y . Once the co-movements of the variables are determined, a “best fit” linear regression line can be estimated and sketched on an x, y coordinated scatter plot.

The population co-variance is a measure of the linear relationship between two variables of a population.

$$\begin{aligned}\text{Cov}(x,y) &= E[(x-\mu_x)(y-\mu_y)] \\ &= \sum (x - u_x) (y - u_y) / N\end{aligned}$$

Cov(x,y) is a parameter and is only meaningful for quantitative data. Covariance measures the co-movements of two or more variables, to ascertain to what extent the variables move together. Cov (x, y) > 0 implies a positive co-movement; Cov (x, y) < 0 implies that the variables move in opposite directions; Cov (x, y) = 0 implies no linear relationship between the two variables.

The sample covariance (σ_{xy}^{\wedge} , or s_{xy}) is a measure of the linear relationship between two variables of a sample. Sample statistics involves the same equations as population statistics, only with n-1 used in the denominators (instead of N), and the term "s" used as the sample standard deviation (instead of σ).

$$\begin{aligned}\text{Cov}^{\wedge}(x,y) &= E[(x - \bar{x})(y - \bar{y})] \\ &= \sum (x - \bar{x}) (y - \bar{y}) / (n-1)\end{aligned}$$

Cov[^] is similar to Cov (x,y) in that it is a sample statistic, is only meaningful for quantitative data, and the same implications exist for co-movements.

Note that a variance – co-variance matrix describes the relationship between the variables of an equation. The covariance of a variable with itself is simply the variance, and is usually noted on the diagonal of the matrix, while cov (a, b) is the same as cov (b, a).

The Basic Linear Model. There are several reasons to develop a model that studies f (x), including the prediction of y; measuring the effect that x has on y; and for hypothesis testing.

A Deterministic Model is where a variable, y_i , is a function of a group of other variables, $x_{1i}, x_{2i}, x_{3i}, \dots, x_{ki}$, such that: $y_i = f(x_{1i}, x_{2i}, x_{3i}, \dots, x_{ki})$. y is known as the dependent, or endogenous variable. The x variables are known as the independent, explanatory or exogenous variables. They are assumed not to be random variables (non-stochastic).

A Probabilistic Model is one with a variable, y_i , being a function of a group of other variables, $x_{1i}, x_{2i}, x_{3i}, \dots, x_{ki}$ and an error term, ε_i , such that: $y_i = f(x_{1i}, x_{2i}, x_{3i}, \dots, x_{ki}) + \varepsilon_i$. The model is assumed to be linear. The probabilistic model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The residual or error term, ε_i , captures coding errors, minor misspecifications of f(x)'s functional form, and unobservable x's. It is the amount of the dependent variable, y , that cannot be explained by the independent variable, x_i . Assumptions made regarding ε_i include: $E(\varepsilon_i) = 0$; $\text{Var}(\varepsilon_i) = \sigma_{\varepsilon}^2 < \infty$ (the ε_i 's are homoskedastic). ε_i 's are independent of each other and of the x values. We often say that error terms satisfying these assumptions are "well-behaved". Thus: $E(\varepsilon_i \varepsilon_j) = \sigma_{\varepsilon}^2$ when $i = j$. $E(\varepsilon_i \varepsilon_j) = 0$ when $i \neq j$. Residual, e^{\wedge} , is the difference between the true value of y and the estimated value.

Since there is only one x variable (with k=1), there is no need to refer to it as x_1 (although it is still common to do so). β_0 is the Y intercept on an x, y coordinate graph, while β_1 is the slope of the estimated regression line. In terms of calculus, e_i is the first derivative of y with respect to x. Thus we will interpret β_1 as: The change in y, caused by a one-unit increase in x, holding everything else constant. $\beta_1 > 0$ implies that increasing x increases y; $\beta_1 < 0$ implies that increasing x decreases y; $\beta_1 = 0$ implies that increasing x does not affect y; β_1 will have the same sign as $\text{Cov}(x,y)$. β_0 and β_1 are unknown population parameters. We will use sample information to generate estimates β_0^{\wedge} and β_1^{\wedge} .

The Estimated Model is a probabilistic model with estimates substituted for each parameter and ε_i . y^{\wedge} is called the is called the estimated, fitted, or predicted dependent value. The estimated model is also called the prediction equation. The regression line is often referred to as the “least squares” or “best fit” line since it minimizes the sum of the squared differences between the variables and the estimated line.

$$y_i^{\wedge} = \beta_0^{\wedge} + \beta_1^{\wedge} x_i$$

Note even though we interpret the model as capturing the effect of x on y, we cannot prove a causal relationship. The actual relationship may be x on y (i.e. causation); y on x (reverse causation); x affecting z which then affects y (z is a lurking variable that may not even be defined in the equation or known to the researcher); etc. The decision regarding which variable to put on the left hand side and which to put on the right hand side must be based upon theory and knowledge of the subject matter. Thus, it is not appropriate to infer that “x causes y”. It is better to state that “x explains y”.

The sum of squares for error (SSE) is the minimized sum of squared deviations. The SSE measures the uncertainty of the explanatory power between the dependent and independent variables. It is similar to standard deviation for a single variable, except that it measures the standard deviation of the error term in the regression. The residual, e_i , is the difference between the actual value and the predicted values of the dependent variables. The SSE gives some indication of how certain we can be about a particular prediction of y using the regression equation.

$$SSE = (n-1) [s_y^2 - (\text{cov}(x, y)^2 / s_x^2)]$$

Since the var (x) is the same thing as the cov (xx) = S_{xx} , the above equation can also be notated as: $SSE = S_{yy} - (S_{xy}^2 / S_{xx})$. And, various texts also note the following equation: $SSE = [\sum e_i^2 / (n-2)]^{1/2}$, with $e_i = y_i - B_0 - B_1 x_i$.

The standard error of estimate measures the suitability of using a linear model, and essentially is an estimate of the unknown population variance, σ_x^2 .

$$s_e^2 = SSE / (n - 2)$$

Ordinary Least Squares (OLS) is a method for estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. The estimates make SSE as small as possible. These estimates are called the “Least Squares Estimators”. The least squares coefficients are:

$$\begin{aligned}\hat{\beta}_1 &= \text{cov}(x, y) / s_x^2 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \text{Cov}^{\wedge}(x, y) &= S_{xy} = \sum (x - \bar{x})(y - \bar{y}) / (n - 1) \\ \text{Var}(x) &= s_x^2 = \sum (x - \bar{x})^2 / (n - 1) \\ \bar{x} &= \sum x_i / n \\ \bar{y} &= \sum y_i / n\end{aligned}$$

Mean Square Error (MSE) of Regression (S_e^2). Recall that $\text{Var}(\varepsilon_i) = \sigma_\varepsilon^2$. Since, typically, σ_ε^2 will be unknown, we will estimate it with:

$$S_e^2 = \text{SSE} / (n - 2)$$

It can be shown that S_e^2 is an unbiased estimator for σ_ε^2 . S_e is also known as the standard error of the regression. This formula only applies when $k=1$.

On the estimator, $\hat{\beta}_1$, the expected value of $\hat{\beta}_1$ is β_1 , or $E(\hat{\beta}_1) = \beta_1$. This makes $\hat{\beta}_1$ a unbiased estimator of β_1 . $\text{Var}(\hat{\beta}_1) = s_e^2 / S_{xx}$, or Var of the estimate is equal to the variance of the residual error term / variance of x_i . Thus, lower values of the residual's variance will decrease the standard error of the estimator. For $\hat{\beta}_1$ to be truly unbiased, the difference between the OLS weighting and the weights of the other estimators, $d_i = 0$, where the weights, $c_i = (x_i - \bar{x}) / S_{xx} + d_i$; and then, $\sum d_i = 0$; $\sum d_i x_i = 0$;

The Gauss-Markov Theorem establishes that the OLS method produces the best linear unbiased estimator, abbreviated as BLUE, such that $\hat{\beta}_1$ and $\hat{\beta}_0$ are the BLUE estimators of β_1 and β_0 .

A confidence interval on the estimator can be established to show the value of the true parameter value within a certain level of confidence. We must define the given level of confidence, and the standard error of estimate (SSE) must be known. For a hypothesized value of a parameter, β_1 , an estimated parameter value, $\hat{\beta}_1$, a t value of $t_{\alpha/2, n-2}$ and a two tail test, the CI will be:

$$\text{CI} = \hat{\beta}_1 \pm t_{\alpha/2, n-2} \text{SSE}_{\hat{\beta}_1}$$

Note that the CI will thus be dependent upon the number of degrees of freedom for the t distribution under the null hypothesis, as well as the level of confidence. The degrees of freedom are the number of observations minus the number of parameters estimated. In a regression with only one independent variable, there will be two parameters estimated: the Y intercept and the coefficient of the independent variable. Thus, for a two-tail test of β_1 in a single regression equation, t will be expressed as: $t_{\alpha/2, n-2}$.

Hypothesis tests on B_1^{\wedge} can be done (for single linear regressions when $k = 1$), and the CLT can be invoked where the residual, e , is close to Normal or where there is a large sample size. The null is $H_0: \beta_1 = \theta_0$. The alternatives are:

$$H_A: \beta_1 > \theta_0$$

$$H_A: \beta_1 < \theta_0$$

$$H_A: \beta_1 \neq \theta_0$$

The Test Statistic, assuming σ_e^2 is unknown, is: $t = (B_1^{\wedge} - \theta_0) / (S_e / \sqrt{S_{xx}})$. The confidence interval will be: $B_1^{\wedge} \pm t_{\alpha/2, n-2} \sqrt{S_e^2 / S_{xx}}$.

Rejection Regions are:

If $H_A: \beta_1 > \theta_0$, Reject if $t > t_{\alpha}(n-2)$

If $H_A: \beta_1 < \theta_0$, Reject if $t < -t_{\alpha}(n-2)$

If $H_A: \beta_1 \neq \theta_0$, Reject if $t > t_{\alpha/2}(n-2)$ or $t < -t_{\alpha/2}(n-2)$

If the conclusion is to reject the null hypothesis, we say that x is significant in explaining y . Note that if a large sample size is used to invoke the CLT, it is logically inconsistent to treat df as small and use the t distribution, as it is not meaningfully different from the z .

Another interesting example of testing regression coefficients follows. The hypothesis test is that a stock's beta is that of the market, 1.00; that the estimated stock beta is 1.5; that there are 62 observations; that the standard error of the estimated parameter is 0.200; and that the given significance level is .05 and a 95% CI. The test is as follows:

- 1) $H_0: B_1 = 1.0$; $H_a: B_1 \neq 1.0$;
- 2) $\alpha = 0.05$; $CI = B_1^{\wedge} \pm t_{\alpha/2} \text{SSE}_{B_1^{\wedge}}$;
- 3) $t_{.025, 60} = 2.00$; Reject if CI is $\neq B_1$;
- 4) $CI = 1.50 \pm 2.00 (0.200)$,
 $= 1.50 \pm .400$,
 $= 1.10 \text{ to } 1.90$.
- 5) Because the entire CI is above B_1 of 1.00, we can be 95% certain that the stock's estimated beta of 1.5 is different than the beta of the overall market, 1.00, and the null H_0 is rejected.

Note that the above method is slightly different than most Hypothesis tests, in that it uses a CI to conclude whether the null is rejected. The more typical method would be to calculate:

- 3) $t_{.025, 60} = 2.00$; Reject if $t^{\wedge} > t_{.025, 60}$, or $< -t_{.025, 60}$.
- 4) $t^{\wedge} = (b_1^{\wedge} - b_1) / \text{SSE}_{b_1^{\wedge}}$;
 $t^{\wedge} = 2.50$; $t = 2.00$; $t^{\wedge} > t$;
- 5) H_0 must be rejected in favor of H_a , being that the estimated stock beta of 1.5 is different than the average market beta, to a 95% level of confidence.

Measuring the Strength of the Relationship Between x and y. There are two ways of discussing the strength of the relationship between x and y.

- 1) A change in x could cause a change in y; and
- 2) Knowing x gives us great certainty in predicting y.

Looking at the strength of x causing y can be measured to some extent by B_1^{\wedge} . Note, however, that the magnitude of B_1^{\wedge} , by itself, does not tell us the strength of the relationship between x and y. In fact, by changing the units of measurement, we can always make B_1^{\wedge} larger or smaller. We would like a measure of the second type of strength, and this is measured by the correlation coefficient and R^2 .

The coefficient of correlation, rho, ρ , describes the strength of a relationship between two variables. The significance of the correlation coefficient is that the covariance of two variables (x_1 and y^{\wedge}) is transformed into a simple range of values (from -1 to $+1$). -1.00 is perfectly negative correlation, while $+1.00$ is a perfectly positive correlation. For no correlations, $\rho = 0$. Positive correlations show a direct relationship between the increase of the variables. Negative correlations indicate that there is an inverse relationship between the two variables: when one increases, the other decreases. Thus, the coefficient of correlation measures the linear relationship between the two variables.

$$\rho_{xy} = \sigma_{xy} / \sigma_x \sigma_y$$

ρ is sometimes referred to as the Pearson Correlation. ρ is a population parameter. It measures the second type of strength of the linear relationship between x and y in the population. ρ has the same sign as σ_{xy} and β_1 . Unlike covariance, ρ is independent of the units of measurement. $-1 \leq \rho \leq 1$.

The sample correlation coefficient (r), measures the correlation of sample variables.

$$r = S_{xy} / \sqrt{S_{xx} S_{yy}}$$

$E(r) = \rho$; r will have the same sign as $Cov^{\wedge}(x, y)$, S_{xy} , and B_1^{\wedge} .

Correlation coefficients also answer the question of spurious correlations by inquiring whether ρ actually comes from a population of paired observations with a zero correlation. The t test is used to establish the degrees of freedom and then deciding whether the hypothesis is rejected or not. Then the p values, previously mentioned, are used to determine the likelihood of finding a value of the test statistic as large or larger when the Hypothesis is true.

To test the proposition, $H_0: \rho = 0$; $H_A: \rho \neq 0$. The test statistic is: $t = r \sqrt{(n-2) / (1 - r^2)}$. Reject if $|t| > t_{\alpha/2, n-2}$. It can be shown that this test is algebraically equivalent to the test of $H_0: \beta_1 = 0$, versus $H_A: \beta_1 \neq 0$ (when $k=1$). A more fully developed hypothesis on whether the population correlation coefficient equals zero can be done. Following the procedure for Hypothesis testing:

1. $H_0: \rho = 0; H_a: \rho \neq 0;$
2. $t = r \sqrt{(n-2)} / \sqrt{(1-r^2)}.$
3. $\alpha = 0.05; CI = r \pm t_{\alpha/2, n-2} s / \sqrt{n}.$
4. Reject if $t_{\alpha/2, n-2} > 2.776$ or $t < -2.776$ (for a large n sample size).
5. Collect data; perform calculations on t.
6. Make the statistical decision of rejection or no rejection of H_0 ; State the investment or economic decision.

The coefficient of determination is the proportion of the total variation in the dependent variable, y, that can be explained by the variation in the independent variable, x. It is noted in the equations as r^2 , and when only one independent variable is used, it is simply the square of the coefficient of correlation, r, expressed as a percent. So, r^2 could be described as having 57.6% of the variation in the dependent variable explained by the independent variable. Thus:

$$r^2 = \rho(y_i, x_i)^2.$$

Spurious correlations occur when two variables with a strong correlation will exist, but for completely unrelated areas (like in the Super Bowl winner and subsequent performance of the stock market). The variables look to be related, but in fact are not related. So, r^2 then can be said to only show an apparent relationship between the two variables, and not a change directly caused by the other variable.

r^2 is a sample statistic. $0 \leq r^2 \leq 1$. Note that some information is lost as compared to using r. The SSE will be $\leq S_{yy}$. OLS always has the option of ignoring x, setting $B_1^{\wedge} = 0$, and letting $SSE = S_{yy}$. Since this is always an option for OLS, and OLS's goal is to make SSE as small as possible, we know that OLS will never choose B^{\wedge} that make $SSE > S_{yy}$.

ANOVA. This is a statistical procedure for analyzing the total variability of a set of data. The procedure usually generates several statistical results for a regression, including the estimated values for each coefficient, t values on each estimated coefficient, the SSE, the R^2 of the entire equation, the multiple R value, and the F test statistic. The F test examines how well the regression equation explains the variation in the dependent variable. With only one independent variable, however, the F statistic will be merely the square of the t value of B_1 , so the F test's real value lies with multiple regressions. The above calculations are typically wrapped together in a spreadsheet or software package.

Note that different nomenclature exists in various texts regarding the terms SSE and SSR. The SSE (the portion of sample variation in y that cannot be explained by the regression) is included in ANOVA, as is the Sum of Square Regression, or SSR. Then total sum of the squares, $TSS = SSE + SSR$. Some authors call Sum of Square Error the Sum of Squared Residuals, while other authors call Sum of Square Regression the Sum of Squares Explained.

Also note that r^2 can be interpreted as SSR / TSS . Thus, r^2 measures the proportion of total sample variation in y that can be explained by the x 's in our regression. If x is of no use in explaining y , $r^2 = 0$. If x perfectly explains y , $r^2 = 1$.

Confidence and Prediction Intervals. Confidence intervals report the mean value of y for any given x , while prediction intervals report the range of values of y for any given x . So, confidence intervals estimate the degree of confidence of y to x . Prediction intervals estimate the degree of likelihood of a y for a particular x . This prediction interval will be wider than the confidence interval.

Prediction of the dependent variable. A prediction of the dependent variable, \hat{y} , can be made, given a designated value for the independent variable. Basically, the ANOVA statistics can be used, with a calculation of the predicted value of the dependent variable using the regression equation, and then a confidence interval can be obtained for the predicted value. The process is as follows:

1. Take a designated value for the independent variable, x_1 , and then use B_0^{\wedge} , B_1^{\wedge} to calculate the predicted value of the dependent variable, \hat{y} , from the estimated regression equation, $\hat{y} = B_0^{\wedge} + B_1^{\wedge} x_1$.
2. Compute the variance of the prediction error. The estimated variance of the predicted value is:

$$s_p^2 = SSE [1 + (1 / n) + (x_1 - \bar{x}) / ((n - 1) s_x^2)]$$

where, x_1 is the value of the independent variable used to predict the dependent variable; \bar{x} is the estimated mean of x_1 ; SSE is of the estimated equation; s_x^2 is the variance of the independent variable.

3. Choose a significance level; determine from a t table the critical value of $t_{\alpha/2, n-2}$.
4. Compute the $(1 - \alpha) \%$ prediction interval, namely: $\hat{y} \pm t_{\alpha/2, n-2} \sqrt{s_p^2}$. This will be the range of values for any predicted value.

The limitations of regression analysis. Limitations include:

- 1) A change in regression relations over time resulting in a good fit during the tested period but a poor fit in another historical period or future period;
- 2) Widespread knowledge of the regression analysis may result in people reacting to the analysis, generating a change in the regression relationship. For example, a stock's low P to value may be bid up to a higher Price on news of the initial low ratio.
- 3) If the classical assumptions are violated or are not valid, then the regression results will not be valid.

Multiple Regressions. For linear models having more than one independent variable ($k \geq 1$), the probabilistic model becomes:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

and the estimated model becomes:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki}$$

There are a total of $k+1$ β 's to estimate. We now interpret β_j to be the change in y , caused by a one-unit increase in x_j , holding everything else constant.

Effect on error and prediction. Adding another x cannot increase the SSE. Recall that OLS chooses the variable coefficients that make SSE as small as possible. When $k=2$, OLS has the option of ignoring x_2 by setting $\hat{\beta}_2 = 0$. If OLS does this SSE will be the same with $k=2$ as it was with $k=1$. Since OLS always has this option, it need never choose $\hat{\beta}$'s that yield a larger value for SSE. In theory, increasing k may leave SSE unchanged. In practice, SSE will almost certainly decrease.

Since SSE cannot increase with multiple independent variables, the SSR cannot decrease, since $TSS = SSE + SSR$. In theory, adding another x may leave SSR unchanged. In practice, it will almost certainly increase SSR.

Adding another x cannot decrease r^2 . Since $r^2 = SSR / TSS$, adding a variable will \geq SSR. Since the total SS does not change, r^2 will either increase or be unchanged. In practice, adding another x will almost certainly increase SSR. While in theory, adding another x may leave r^2 unchanged, in practice, adding another x will almost certainly increase r^2 . Typically, in a multiple regression framework, the coefficient of determination is represented as R^2 (and called the Multiple Coefficient of Determination) rather than r^2 .

Calculating MSE with $k \geq 1$. Adding independent variables only serves to increase the size of the denominator with $n-2$ now becoming $n - (k + 1)$.

$$S_\varepsilon^2 = SSE / (n - (k + 1))$$

When S_ε^2 is unknown, adding independent variables will not increase SSE but will serve to increase the denominator. Thus, S_ε^2 will decrease to some extent with the addition of variables. If the additional x is very useful in explaining the variation in y , SSE will decrease a great deal, and S_ε^2 will decrease. If the additional x is not very useful in explaining the variation in y , SSE will decrease a small amount, and S_ε^2 will increase.

Hypothesis testing with multiple variables. To invoke the CLT, the residual must be close to Normal or the sample size must be large. Testing $H_0: \beta_j = \theta_0$ with $k \geq 1$.

Alternative Hypothesis:

$$H_A: \beta_j > \theta_0$$

$$H_A: \beta_j < \theta_0$$

$$H_A: \beta_j \neq \theta_0$$

The Test Statistic, assuming S_ε^2 is unknown, will be:

$$t = B_j^\wedge - \theta_0 / S_{B_j^\wedge}$$

The probability distribution will be: $B_j^\wedge \pm t_{\alpha, n-(k+1)} S_{B_j^\wedge}$

Rejection Regions:

$$\text{If } H_A: \beta_j > \theta_0, \text{ Reject if } t > t_{\alpha, n-(k+1)}$$

$$\text{If } H_A: \beta_j < \theta_0, \text{ Reject if } t < -t_{\alpha, n-(k+1)}$$

$$\text{If } H_A: \beta_j \neq \theta_0, \text{ Reject if } |t| > t_{\alpha, n-(k+1)}$$

When $k=1$, this is identical to the previous framework. If a large sample size is used to invoke the CLT, it is logically inconsistent to treat df as small, and use the t distribution, as it is not meaningfully different from the z .

Adjusted R^2 . This is the R^2 after the addition of variables. While R^2 cannot decrease with the addition of variables, the adjusted R^2 can decrease or increase when variables are added.

$$R_a^2 = 1 - S_\varepsilon^2 / S_y^2$$

Adding a useful x is likely to increase R_a^2 , while adding a relatively useless x is likely to decrease R_a^2 .

Factors Affecting deviation of B_j^\wedge . Small deviations in B^\wedge are preferred, because they will tend to cause smaller errors of estimation, narrow the confidence intervals, and give higher power for hypothesis tests. The variance of the estimator B^\wedge will decrease as the dispersion of x_j increases; as the sample size increases; and as the variance of the residual decreases. σ_{B^\wedge} will increase as the sample covariance between x_j and other independent variables increases. This is called multicollinearity.

The Relationship Between k and df . Adding another x will increase k , and will decrease $n - (k + 1)$ degrees of freedom used for confidence intervals for the B_i estimates. This will increase the critical values used for inference, widen confidence intervals, and lower the power of hypothesis tests (without affecting the significance level). Researchers sometimes feel that each additional independent variable “costs a degree a freedom.”

Omitted-Variable Bias. Often, independent variables are left out of regression models either because of a lack of data, or because of a lack of understanding and knowledge by the researcher that the variable is important in explaining the dependent variable. If the omitted variable is correlated with any other x 's that are included, the estimated coefficients of the included B_i^{\wedge} will then be biased, violating the BLUE nature of the OLS method. $E(B^{\wedge})$ is no longer B , but instead becomes:

$$E(B_i^{\wedge}) = B_i + (\text{Cov}(x_i, x_j) / \text{Var}(x_j)) B_j$$

Including another x into the equation has both positive and negative connotations. On the plus side, it will usually increase R^2 , it eliminates the possibility of omitted-variable bias (for that variable), and it may decrease the deviation of the residual, S_e . However, on the negative side, it may increase S_e , it will decrease the degrees of freedom, and may introduce multicollinearity problems into the process. Texts suggest to not base the decision on hypothesis tests, as this introduces pre-test estimation bias, but to consider theory, previous research and knowledge of the subject matter.

Perfect Collinearity Among Independent Variables. OLS cannot form coefficient estimates if there is a perfect linear relationship among any of the x variables. A linear relationship is one in which one of the x 's can be exactly calculated as a linear function of other x variables. Trying to perform such a regression will result in an error message from the computer.

Dummy Variables. A dummy variable is a variable that only takes on the values 0 and 1. It is a way of incorporating qualitative independent variables in a regression model.

To incorporate a two-level qualitative variable (eg., male/female, white/non-white, union/non-union), define x to take on the value of "1" for one of the levels, and "0" for the other level. The β coefficient for a dummy variable then measures, "The change in y caused by switching from the 0 level, to the 1 level, holding all else constant." In essence, the dummy variable allows the two groups to have different intercepts. Sometimes a dummy variable is termed an Intercept Dummy.

In order to use dummy variables to incorporate a multi-level qualitative independent variable, first define a dummy variable for each level. Then included all but one of these dummy variables in the regression. The excluded level is the comparison level, or comparison group.

Texts refer to a Dummy Variable Trap, where a dummy variable is included for each level. This will create perfect collinearity among the x variables, making it impossible for OLS to estimate the coefficients. The solution is to exclude one of the dummy variables. So, if a dummy variable is established for race, and four types of race are included (white, black, Asian, American Indian), as well as "all other" for all other races, the catch-all dummy variable for "all other" is often excluded to avoid the dummy variable trap.

Interaction Terms. When one variable's (x_1 's) effect on y depends upon the value of another variable (x_2) this relationship can be captured with an interaction term, x_1x_2 . Interacting a continuous variable with a dummy variable allows the slope to be different across the groups. These variables are often called "slope dummies."

The function form of the linear equation. The various x_i 's of the equation can be non-linear in nature (i.e. squared, cubed, square root, log, lognormal, sin, etc), so long as the entire function is linear in the B 's. It is quite common for the independent variables to include a quadratic. The decision regarding the functional form should be based upon theory, previous research and knowledge of the subject matter.

The standard interpretation of the β coefficients does not hold in a non-linear framework. The change in y caused by a one unit increase in x_1 , holding all else constant, is not β_1 , but is the first derivative of x_1 . For example, in a case where several forms of x_1 are being analyzed in the same estimated equation (i.e. $y^{\wedge} = \beta_1^{\wedge} x_1 + \beta_2^{\wedge} x_2 + \beta_3^{\wedge} x_1^2 + \beta_4^{\wedge} x_2^2 + \beta_5^{\wedge} x_1 x_2$), the change in y caused by one unit of x_1 will be x_1 's first derivative (i.e. $\beta_1 + 2\beta_3 x_{1i} + \beta_5 x_{2i}$).

Log forms can generate an importance type of meaning and analysis. If both the y and x variables are estimated as a log (i.e. $\ln(y)$; $\ln(x)$), then the x coefficient is an elasticity. If $\ln(y)$ and x exists, then B^{\wedge} represents the $\% \Delta y / \Delta x$. If y and $\ln(x)$ is evaluated, then B^{\wedge} represents $\Delta y / \% \Delta x$.

Restricted Least Squares. This involves the estimation of a model's parameters by minimizing SSE subject to one or more restrictions across one or more parameters. For example, if $\beta_1 + \beta_2 = 1$, a restriction has been imposed on the regression equation such that the fit will not longer be done via OLS, but by restricted least squares. Enforcing a restriction cannot decrease SSE.

Multiple parameters and their testing. When a researcher wishes to test hypotheses regarding more than one parameter /restrictions, it is inappropriate to perform multiple tests using one data set. Instead a joint hypothesis test must be performed. This is where the F test, discussed in prior sections, becomes quite useful. Since the F statistic is used to compare across two populations, it is ideal for joint testing.

$$F = (SSE_R - SSE_U) / q / (SSE_U / (n - (k+1)))$$

Where, SSE_R is from restricted model; SSE_U is from unrestricted model; q is number of restrictions; $k+1$ is number of parameters of unrestricted model. If the restriction is valid, SSE_R will be close to SSE_U , and F will be close to 0. Reject if $F^{\wedge} > F_{\alpha, n-(k+1)}$.

The Chow Test. Suppose it is believed that all parameters, including the intercept, are affected by the value of the dummy variable D_i . The probabilistic model uses $k+1$ interaction terms for a total of $2(k+1)$ parameters:

$$y_i = \beta_0 + \gamma_0 D_i + \beta_1 x_{1i} + \gamma_1 D_i x_{1i} + \dots + \beta_k x_{ki} + \gamma_k D_i x_{ki} + \epsilon_i$$

This implies that there are actually two regimes:

$$D_i=1: y_i = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)x_{1i} + \dots + (\beta_k + \gamma_k)x_{ki} + \varepsilon_i$$

$$D_i=0: y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i.$$

Testing that the regimes are equivalent is a test of $H_0: \gamma_0 = \gamma_1 = \dots = \gamma_k = 0$; H_A : at least one $\gamma_j \neq 0$. This can be tested using the F test. SSE_U can be found by estimating the general model above, or by estimating separate equations for the two regimes, and adding the two SSE values.

Times Series and Forecasting

A time series is a collection of data recorded over a period of time. There are four types: a trend, a cyclical variation, a seasonal variation, and irregular variation. With seasonal trends, the LT is a smoothed direction of a time series. With a cyclical, the change to the time series is over a period of time longer than one year. With a seasonal, the change is over a season, and the trend tends to repeat over a season. In irregulars, an episodic is an unpredictable but identifiable variation, while a residual is unpredictable and unidentifiable. This would be associated with random events. Irregulars cannot be projected into the future.

Linear trends are for straight line projections. It uses the same formula as noted before of $Y' = a + b * t$. The least squares method computes for the best fit along a regression line, and the trend can be used to estimate into the future. This would be for a LT projection, for example. A moving average method is used with seasonal variations, and it merely smoothes out the fluctuation in data by a moving arithmetic mean through the time series. This should follow a fairly linear line and have a definite rhythmic pattern of fluctuations. The cyclical and irregular activities can sometimes be removed entirely with a moving average, leaving a nice straight projection.

Non-Linear Trends would be used for non-linear changes. A logarithmic formula is used: $\log Y' = \log a + \log b * t$. Basically, the same Y intercept formula of a linear trend is used, except that it is moved into a log equation, resulting in a curved line.

Sales data can “Deaseasonalize” or isolate the seasonal changes from the normal growth change to revenues over a period of time. Seasonal Indexes can be created to track a quarterly or monthly pattern. Typically, a seasonal fluctuation can be measured by comparing the season to a moving average method. This can eliminate a cyclical or irregular component. The first step is to determine the 4 quarter moving average. Then, the moving average should be centered – this is done by averaging the last 2 moving average periods. The specific seasonal period for each quarter is then computed by the Sales of the quarter divided by the centered MA. This then generates the ratio of the original value to the MA, which contains the only the seasonally specific information. Then organize the seasonal info into a table sorted by quarter, and develop a mean of year over year quarterly seasonal ratio to the MA. This produces the seasonal index, and it can be compared to the current quarter to determine how the current quarter is doing.

The resulting sales data is seasonally adjusted with the seasonal fluctuations removed. This is done by taking the sales and dividing into the seasonal index to equal the deseasonalized sales figures.

Forecast of deseasonalized data. This is determined by the least squares method using the deseasonalized historical data. This is then projected into the future with the adjusted sales figures being in the trend.

Forecasts may not be accurate for several reasons: There may be a failure to examine the assumptions; there may be limited expertise; a lack of imagination; a neglect of the constraints; excessive optimism; reliance on mechanical extrapolation; premature closure; and over specification.

Basic Statistical Concepts Regarding Investments

An investment is typically defined as a current commitment of money for a period of time in order to derive future payments as compensation for the following: 1) the time the funds have been committed; 2) the expected rate of inflation; 3) the uncertainty of future payments. For investors to accept a “pure” or nominal rate of interest that only compensates for the deferral of consumption into the future, the stream of future payments must be very stable and not subject to volatility at the future date. The nominal rate of interest is often considered the risk free rate of return (RFRR or r_f). If there is price volatility of the investment, the investors will then consider this to be an investment risk, and will most likely demand a risk premium, which is an additional return added to the nominal interest rate. The investor is trading a known dollar today for an expected future stream of payments tomorrow, and the investor will demand compensation for the time, the expected rate of inflation, and any risk associated with the flow of the future income stream.

Time Value of Money. The value of an investment is chiefly a function of time. Cash today, if conserved, saved, and invested will more worth considerably more in the future. It is a question of current consumption versus current savings and then consumption deferred until tomorrow.

Cash inflow is a fancy name for expenses, while cash outflow is the cash receipt.

TVM calculations are dependent upon interest rates, also known as the discount rate. In a certain world, the interest rate would be the RFRR, such as the US T Bill. In an uncertain world, an inflation premium must be accounted for which factors in expected levels of inflation. Further, the interest rate must account for the default risk premium on the asset being reviewed. The opportunity cost (or discount rate) at equilibrium of supply and demand for money would incorporate these factors into the market pricing of interest rate for any one asset. Thus, the interest rate at equilibrium, would be approximately equal to the RFRR, an inflation premium that adds in expected inflation levels, and a default risk premium of the asset.

The stated annual interest rate is the quoted rate on an annual basis, r_s . The stated interest rate, or quoted interest rate, does not account for compounding, and is thus referred to as the “simple” interest rate. The stated annual interest rate (or the quoted interest rate) = monthly interest rate * 12. Ex: a quoted annual interest rate may be 8%. The quoted or stated monthly rate would then be 0.67%. The frequency of compounding is number of compounding periods per year, m . Thus, $r_s = r * m$, or $r = r_s / m$.

The effective annual rate (EAR) is also known as annual percentage rate, or APR. The effective interest rate accounts for the compounding of interest.

$$\text{EAR} = (1 + r_s / m)^m - 1 \text{ or}$$
$$\text{EAR} = (1 + \text{periodic interest rate})^m - 1.$$

For example, if the annual interest rate is 8% and it is compounded semi-annually, then the EAR = 8.16%. In terms of continuous compounding, the formula is: $EAR = e^{rs} - 1$.

Future value is the amount of cash compounded in the future from the present value of cash, assuming a certain stated rate of interest. The present value is the value today of cash flow into the future. In other words, a stated sum of cash at a future time is worth a discounted amount into the present time. This is all because of the power of compounding returns – interest being earned on top of interest generates a larger amount of cash in the future. The opportunity cost is the rate of return of alternative investments that can generate interest or income bearing amounts. The fair or equilibrium value is the value at which investors are indifferent about an investment. The amount is so “fair” that an investor could either buy or sell at the stated value and still feel that he or she achieved a reasonable valuation for the investment.

Perpetuities are a stream of equal payments expected to last forever, or into perpetuity. The most famous example of a perpetuity is the Consul bond, or perpetuity bond. Consuls were issued by the British government in 1815 to consolidate past war debts into a general perpetual bond issue. The equation is: $PV = PMT / i$. As interest rates increase. The present cash value of a perpetuity decreases, due to a higher rate of return generating a correspondingly greater amount of money able to cover the same monthly or annual payment stream. The PV of a Perpetuity occurs where r is the interest rate (in decimal points), and assuming an annual compounding, a constant interest rate, r , and a perpetuity into the future:

$$PV = FV / r, \text{ or}$$
$$PV = FV / 1 + r;$$

The calculation of PV and FV of perpetuities can be applied to more general forms of PDV CF equations. Building up the equations from a perpetuity and starting with a single sum of money, for a specific number of future periods (no longer a perpetuity), N :

$$PV = FV / (1+r)^N, \text{ or}$$
$$PV = FV (1+r)^{-N}$$

Stating for FV, the equations can be rearranged as:

$$FV = PV (1 + r), \text{ or}$$
$$FV = PV (1 + r)^N$$

For compounding with other than annual periods, just divide the annualized interest rate by the number of periods of compounding per year, and then multiply the number of years by the number of periods per year to generate the number of total periods. A few terms are in order at this point. The nominal interest rate is a quoted or stated rate, and is stated in terms of an annual rate of interest. The effective annual rate (EAR or EFF) is that rate generated over periods to produce a future value. Typically, the effective rate is

higher than the nominal rate for compounding with less than annual periods. This is due to interest being earned on interest generated from earlier in the year. The effective rate therefore demonstrates the power of compounding interest. The annual percentage rate (APR) is the annual rate of the periodic interest rates. The APR merely sums all of the periodic rates to an annual amount. For example, credit cards with a 1.5% monthly periodic rate will carry 18% APR. The effective rate on this amount would be 19.6%, however.

Where there is more than compounding period per year, m , and using the stated annual interest from above, r_s , we can say that $r = r_s / m$, and substitute from there. Also, the future period is broken into less than annual periods so that we can compound over smaller periods. Thus, the forward period is stated as $m * n$.

$$FV = PV (1 + r_s / m)^{m * N}$$

and:

$$PV = FV (1 + r_s / m)^{-m * N}$$

With continuous compounding, $FV = PV e^{r_s * N}$.

Annuities are a series of payments of equal amounts over fixed intervals generating certain levels of income or cash. An ordinary annuity, or deferred annuity, has payments occur at the end of each period of time for the payment. An annuity due requires payments at the beginning of the period. (Remember to set for BGN or END on the payment key before the calculation occurs for the value of the annuity). FV_{An} is the future value of the annuity over n periods. Annuities Due generates larger values because compounding will start at the beginning of each period, thus providing slightly larger values over the stretch of time.

An ordinary annuity is a series of CF payments, A , due at the end of each period (indexed at $t = 1$). An annuity due is a series of CF payments due at the start of the period (indexed at $t = 0$). This is the case of leases, etc. Annuities can involve both even CF and uneven CF. Perpetuities, discussed briefly above, are simply CF's that go on forever, in perpetuity.

Starting with the perpetuity of annuity, $PV = A / r$, we have seen that FV / PV calculations generated from there are generally of the mold: $FV = PV (1 + r)^N$. Expanding from there, the general annuity formula for use with ordinary annuities and even cash flows and is:

$$FV = A [(1+r)^N - 1] / r,$$

and

$$PV = A [(1 - (1 / (1+r)^N)) / r]$$

Then, an annuity due with even cash flows will simply add the first annuity payment into the equation:

$$PV = A + A [((1+r)^N - 1) / r].$$

Uneven Cash Flows are a series of cash flows that varies in amount. Annuities and perpetuities, conversely, typically involve regular periodic payments of equal amounts. Basically, the series of uneven cash flows are brought back to present value, and then the results of the series are summed. Future value can be ascertained in the same manner, only in reverse, by taking the present value of the cash into the future on each of the uneven amounts and then summing the results of the entire series. Interest can likewise be solved, for any given present and future value, although for uneven flows, the IRR key must be computed (this is the internal rate of return).

For uneven cash flows, the most direct route would be to calculate the future value of each cash flow one at a time and then summing. The calculator has speed keys that asks for each period's CF, and then sums from there. The equation is:

$$PV = \sum [A (1 + r)^{-t}]$$

Fractional Time Periods compound or discount over fractional periods of time. The dollar amount is solved by computing the annualized amount, and then dividing by the fractional time period as a percentage of the entire year.

For mortgages, tuition loans, or other similar loan structures, the formulas above can be used in the context of time indexing, with the general form of equation being:

$$PV_0 = FV_t (1+r)^{-N}$$

Amortized loans are repaid in equal payments over the life of the loan. An amortization schedule will be normally generated, giving the required payment date, payment amount, interest, and principal. For an amortization schedule involving mortgages, auto loans, college tuition payment schedules, etc, and starting with the PV of an annuity, $PV = A [((1 - (1 / (1+r)^N)) / r)]$, we expand further to include multiple compounding periods per year:

$$PV = A [(1 - (1 / (1+ rs / m)^{m*N})) / (rs / m)].$$

This generates a monthly amortization schedule for compounding on a monthly, or m times per year, basis. The time index comes into play when there are uneven cash flows at various time periods.

NPV Equations. Net Present Value concepts are important in the field of Corporate Finance. Net present value of an investment is PV of cash inflows minus PV of cash outflows.

$$NPV = \sum \text{net CF}_t / (1+r)^t,$$

where r is the discount rate, and CF is the net CF at time t . N is the projected life of the investment, and NPV is summed for all periods of N . The net CF is inflows – outflows for each period t . So, the inflows and outflows in each period must be netted, and then brought back to PV.

The internal rate of return (IRR) is the discount rate that makes the $NPV = 0$. $NPV = CF_0 + CF_1 / (1+IRR)^1 + \dots + CF_N / (1+IRR)^N$. The project is accepted where the $IRR >$ opportunity cost of capital (i.e. the WACC). Without a calculator or a spreadsheet doing the iteration, the IRR has to be sequentially increased or decreased until $NPV = 0$.

Comparisons can be made between NPV and IRR. NPV uses the discount rate, which is external to the model, and is typically based on market information. The IRR is completely endogenous, and does not depend upon any outside information. The IRR will generate a single number in terms of a rate of return (and then the decision rule results when $IRR \geq WACC$), whereas the NPV will base the decision rule upon whether the PV of the net CF is positive or negative.

NPV and IRR will give the same decision rule when competing projects are independent of each other. However, when multiple projects are not independent, the IRR decision may suffer, as a result. Also, when there is limited funding available, IRR might not lead to a maximization of shareholder wealth, where the scale of the projects differ – the IRR of a smaller project might be higher, for example, even where the net wealth build-up would be smaller (due to the smaller size of the project with a higher IRR). Further, the timing of the CF may generate a higher IRR even though NPV is lower than another project. The NPV incorporates market opportunity cost of capital in the discount rate, and thus should be used where NPV and IRR conflict. IRR may produce one neat and simple rate of return, but the maximization of net wealth may be greater by using NPV.

Holding Period Returns and Yields. For purposes of portfolio performance measurement and evaluation, the IRR is called the money or dollar-weighted rate of return because it accounts for the timing and amount of dollar flows in and out of the portfolio. This method will generate a differing rate of return, depending upon CF timing.

To calculate the money-weighted and time-weighted rates of return in a portfolio, dollar weighted returns are calculated from the IRR equation, which can also be thought of as $PV_{out} = PV_{in}$, with the example being CF going out initially and in period 1, while CF comes in over several periods. $CF_{out} + CF_1 / (1+r) = CF / (1+r) + CF_N / (1+r)^N$. This type of return is also known as the arithmetic rate of return.

The time-weighted rate of return is not affected by cash inflows and outflows. It measures the compound rate of growth of \$1 invested in a portfolio over a stated period of time. Portfolio managers will use the time-weighted rate of return for their performance evaluation, since the timing of CF inflows and outflows will not affect the rate of return. CF timing usually is beyond the control of the investment manager, and thus the manager's performance evaluation should not be based upon the dollar-weighted rate of return.

Time weighting is geometric rate of return, and is more appropriate for long-term portfolio performance measurements. The time-weighted return is the geometric mean of two or more holding periods, and mirrors the compound rate of growth. The equation is done in two parts. First, each period's return is calculated separately as: $r_i = (MVE - MVB) / MVB$, which is the market value ending – market value beginning divided by MVB. This is sometimes described as holding period return (HPR or HPY, for holding period yield). Then, the return for each year is calculated at a geometric / compounded rate: $(1+r_1) * (1+r_2) * \dots * (1+r_n) - 1 = r$.

The geometric mean can be seen as the multiplication of all holding period returns to the 1/n power minus 1. The equation is $GM = (HPR1 * HPR2)^{1/n} - 1$. The geometric mean is superior to the arithmetic mean, at times, because it is based on the ending value and not the beginning value. The arithmetic mean can be biased upward, especially for volatile equities. When annual rate vary widely, the GM will always be below the arithmetic mean. For portfolios, the mean historical rate of return is the weighted average of the rates of return of the individual investments of the portfolio. The weighting is the relative beginning values of each investment.

The Holding Period Return (HPR), inclusive of dividend payments, is: $HPR_t = (P_1 - P_{t-1} + D_t) / P_{t-1}$. Total return has an aspect of time to it (monthly, yearly, etc). There is also no current unit attached to it. Also, note that this is the same general equation used for T Bills holding period returns.

T Bills are quoted on a bank discount basis, rather than a price basis. Holding period return (HPR) for a T bill is the return that an investor has on the bill, if held to maturity. The general form of the equation is $HPR = (P_1 - P_0 + D_1) / P_0$, where P1 is the initial price, P0 is the price at maturity, and D1 is the interest or dividend paid at maturity and any accrued interest before maturity (as is the case with coupon bearing bonds). With a T bill, no interest is earned, so $HPR = (P_1 - P_0) / P_0$.

The bank discount yield is: $r_{BD} = (D / F) * (360 / t)$, where r_{BD} is the bank discount yield, D is the dollar discount (equal to the difference between the face value of the T Bill and its purchase price), F is the face value of the bill, and t is the # of days until maturity. D is also referred to as the dollar discount from par. The T bill is considered a pure discount instrument.

Effective annual yield (EAY) is the compounded version of HPR, and accounts for interest on interest. $EAY = (1 + HPY)^{365/t} - 1$. The general rule is that the bank discount rate will be less than the effective annual yield, because of compounding of interest with EAY. We have seen the effective annual yield in the last chapter being expressed as the APR, or effective annual return (EAR), which was $EAR = (1 + r_s / m)^m - 1$ or $EAR = (1 + \text{periodic interest rate})^m - 1$. Note the similarities in equations, since essentially the yield or return is being measured in effective annual terms.

The money market yield is comparable to yield quotes on interest bearing money market instruments. $r_{MM} = (360 * r_{BD}) / (360 - t * r_{BD})$. The mm yield does not require knowing the T bill price. Generally, the mm yield is larger than the bank discount rate.

The yield to maturity on a bond ignores compounding, and is the yield quoted in the bond markets. It is annualized, and is multiplied by 2 because bonds typically pay interest semi-annually. $YTM = 2 [(M / P)^{1/N} - 1]$. M is the maturity or face value, P is bond price, and N is the number of semiannual periods to maturity.

For continuously compounded rates of return, it is important to know that most stock return data is lognormally distributed. A discrete compounded rate of return is based upon specific data points. The continuously compounded rate of return becomes a continuous probability distribution. If a stock's continuously compounded rate of return is normally distributed, the future stock price will be lognormally distributed. A series of stock prices (with specific and discrete rate of returns attached to it) can still be lognormal even where the continuously compounded rate of return is not normal, due to the central limit theorem as to large numbers assuming a Normal distribution.

$r_{t, t+1} = \ln S_{t+1} / S_t = \ln (1 + R_{t, t+1})$, where r is the continuously compounded rate of return for period t to t+1, S_t is the stock price in period t, and R is the rate of return in period t to period t+1. This can be expressed in words as: the continuous rate of return is the natural log (or lognormal return relative) of 1 + the holding period return, with the HPY = ending price (S_1) divided by the beginning price (S_0 , or price relative). **HPY is thus $= S_{t+1} / S_t$** . Note that the returns are assumed to be IID. A shorter equation to the continuously compounded rate of return = $\ln 1 + HPR$, or $\ln (EV / BV)$, where EV is the ending value of a stock price, and BV is the beginning value.

Risk and Return Measurements for use in Portfolio Analysis. The calculation of many investment related items are essentially the probability-weighted expected values for the various financial variables.

For examples, covariance is a measure of co-movement between two variables, and is: $Cov (R_i R_j) = E [R_i - ER_i) (R_j - ER_j)]$. This is the probability weighted average of the cross-product of each variable's deviation from its own expected value. Note that $Cov (R_i R_j) = Cov (R_j R_i)$. Covariance can be negative if the returns of one asset are higher than its expected value while the returns of another asset are lower than the expected value. A covariance of 0 indicates that the returns of the assets are unrelated. Covariance will be positive if returns of both assets of a two-asset portfolio, for example are both above their respective expected returns.

Correlation is also a measure of co-movement between two variables, but is expressed as a ratio, and thus eliminates some of the problems with a stated value of variance (say 156.65. What is the significance of such a number?). Perfect correlation of 1.00 to -1.00 for perfectly negative correlation. Correlation is the degree to which the movements of two assets are related, or correlated, together. $\rho (R_i R_j) = Cov (R_i R_j) / \sigma (R_i) \sigma (R_j)$, subject to $-1 \leq \rho \leq 1$. Correlation then is Covariance / the cross product of the asset's

standard deviation. Correlation of < 1.00 will reduce risk through diversification, although normally, a negative correlation is sought out for diversification purposes, with $\rho < 0$.

While this introduction into correlations of assets is more extensively covered in portfolio theory, we can obtain a better understanding of the terms if we briefly consider the matter. As the number of holdings in a portfolio increases, the importance of covariance increases, while the variance (and the square root of the variance) becomes less important. Covariance between assets comes to dominate individual asset's variances (and standard deviation) in a portfolio. The covariance of an asset with itself is simply the asset's own variance. And, as correlation decreases, diversification increases because the covariance of assets will decrease, thus enabling a portfolio's returns to not deviate much in response to the movements of individual asset's returns.

Expected value is the probability weighted average of all possible outcomes of the random variable. The ER incorporates the basic risk of an investment and the uncertainty that an investment will earn its expected rate of return. $E(X) = \sum P(x_i) * x_i$. The expected return of a portfolio is just the sum of the weighted average of the expected returns on the component securities. $E(R_p) = \sum [w_i E(R_i)]$, where $\sum w_i = 1.00$, also expressed as $E(R_p) = \sum E(w_i R_i)$. Note a shortcut: if assets are uncorrelated (i.e., $\rho = 0$), $E(XY) = E(X)E(Y)$. Another shortcut for expected value is: $\sum E(w_i R_i) = w_i E(R_i) + \dots w_n E(R_n)$, with the weighted average of $n * E R_n$ being summed.

The variance of a random variable is the expected value of the squared deviations from its expected value, and is: $\sigma^2(X) = \sum [(X - E(X))^2]$. In lay terms, the equation states that the variance is the sum of the squared deviation of expected value (as defined above). Zero variance indicates zero dispersion. Variance must be zero to positive, because it involves the sum of the squares. Variance is a measure of dispersion about the mean, of course. Increasing variance indicates increasing dispersion about the mean. Also, $\sigma^2(wR) = w^2 a \sigma^2(R)$, and that $\sigma^2(a + R) = \sigma^2(R)$.

In terms of portfolio variance, $\sigma^2(R_p) = E [R_p - E(R_p)]^2$. This is also $= \sum_{i=1 \text{ to } n} \sum_{j=1 \text{ to } n} w_i w_j \text{Cov}(R_i R_j)$, or abbreviated further as: $\sigma^2(R_p) = \sum \sum w_i w_j \text{Cov}(R_i R_j)$. Note: an expanded version of the equation, for two assets, is: For two assets, it is: $\sigma^2(R_p) = w^2 a \sigma^2(R_a) + w^2 b \sigma^2(R_b) + 2w_a w_b \text{Cov}(R_a R_b)$.

Standard deviation is simply the positive square root of the variance, $\sigma = \sqrt{\sigma^2(R_p)}$. Standard deviation also measures dispersion about the mean, but does so in the same units as the variable being measured, whereas variance measures dispersion in units squared. Standard deviation of a portfolio is the basic measure of risk for the CAPM and MPT.

Covariance in a forecasting, forward-looking sense requires the used of joint probabilities. The joint probability function of random variables gives the probability of the joint occurrence of the random variables, and is essentially the summation of the probability of all cross-product combinations of deviations of returns from their expected

returns. $Cov(R_i, R_j) = \sum_i \sum_j P(R_i, R_j) (R_i - ER_i) (R_j - ER_j)$. When random variables are independent, the joint probability function is the product of the individual probability function of the random variables. $P(X, Y) = P(X) P(Y)$, if X and Y are independent.

Many portfolio evaluation measures have been developed over the years. Perhaps the most famous is the Sharpe Ratio. This is a relative measure of portfolio risk, this time against return. It is $Sharpe = (r_p - r_{frr}) / \sigma_p$, where r_p is the mean return of a portfolio, r_{frr} is the mean return of the RFRR. Basically, the Sharpe ratio measures the excess return beyond that of the rfrr as compared with the standard deviation of the portfolio. It asks the question of: what is the extra return from a portfolio for the extra risk of a portfolio?

Shortfall risk is important in portfolio analysis. This is that risk that the portfolio will fall in value below the minimum acceptable level over some time horizon. Roy's safety-first criteria states that the optimal portfolio minimizes the probability that portfolio return will fall below the threshold level. The objective is to minimize $P(R_p < R_t)$. The optimal portfolio will maximize the safety-first ratio of: $SFR\ ratio = (E(R_p) - R_t) / \sigma_p$. Thus, shortfall risk is essentially described in terms of standard deviation units. This is similar to the Sharpe ratio, discussed above. The highest Sharpe Ratio will minimize the probability that the portfolio return will be less than the RFRR.

A Monte Carlo simulation is useful in simulating complex financial problems. Basically, random numbers can be generated on variables to be tested, and then outcomes under various trials can be analyzed, resulting in a probability distribution. Monte Carlo can be used to simulate a proposed policy without actual implementation, and thus is ideal for "what if" situations.

An historic simulation uses historical data for the simulation, instead of pure random numbers. Historical simulation uses repeated sampling from historical time series. It cannot do "what if" scenarios since it cannot go outside of the range of historical data points, and can only reflect the risks represented in the historical data sample.

Both types of simulations are heavily used in actuarial work, pension liability estimation, option pricing, and VAR, as probability ranges can be generated prior to the implementation of any policy or financial plan. The limitations of both Monte Carlo and historical simulations include utilizing only statistical efforts, without any analytical methods (unlike Black-Scholes), and not being able to generate precise estimates (such as an exact option price) but only a range of probabilities.