Indiana University
School of Public and Environmental Affairs

SPEA V506: Statistical Analysis for Effective Decision-Making Lecture Notes

Developed by Professor Barry Rubin

PART ONE: DESCRIPTIVE STATISTICS

PRELIMINARY CONCEPTS

I. Functions of Statistics

A. Description - Summarizing information through the computation of measures such as percentages, means, medians, measures of dispersion, etc.

B. Inference - Statistics can be used to infer causality, but this must be done very carefully and cautiously. The process is chiefly the inference of properties of a population based on a sampling procedure.

1. A *population* is the collection of all possible individuals, entities, objects, or measurements of interest for a particular investigation. A *sample* is any portion or subset of the population.

2. A *statistic* is any measurable characteristic of a sample. A *parameter* is any measurable characteristic of a population. Statistical analysis utilizes statistics from representative samples to infer the parameters of an entire population.

II. Measurement and Classification

A. The use of statistical methods requires a means of classifying data by type. There are four such classification "scales."

1. Nominal Scale
   a. classification based on characteristics or attributes with no hierarchy or quantities implied;
   b. example - local governments can be classified by form of government: city manager - city council, strong mayor - weak council, a weak mayor - strong council;
   c. nothing is implied about which category is "better" or "higher" in a hierarchy;
   d. this is the most elementary form of measurement.

2. Ordinal Scale
   a. classification designed to provide a hierarchy or continuum along which cases or observations can be ordered;
   b. we can use relationships like greater than or less than;
   c. this scale does not provide any information as to "how much."

3. Interval Scale
   a. requires the use of some common standard of measurement so that we can indicate exact distance between observations or cases;
   b. implies quantity;
   c. for example, IQ is measured on an interval scale.
4. Ratio Scale
   a. highest level;
   b. requires that a nonarbitrary zero-point be identified;

c. can make statements like "*X* is twice as large as *Y*";
d. mathematical operators can be applied to data that conform to a ratio scale;
e. derives its name from the fact that meaningful ratios can be computed.

B. Variables

1. For any one case or observation (i.e., the subject of an investigation), we usually observe a variety of characteristics or attributes that enable us to categorize the case based on these characteristics.

2. If an attribute has the same value for all cases under observation, it is termed a *constant*. If different values are evident for different cases, it is a *variable*.

3. One useful definition of a variable is:
"...a characteristic that may distinguish one item from another or, more precisely, a dimension along which items may differ from each other." (Olson)

4. Variables that are based on a ratio scale can be divided into:
a. *continuous variables*, which can take any value along the scale, or;
b. *discrete variables*, which can assume only integer values (i.e., not fractions).

5. If more than one characteristic is recorded as part of the data obtained for each observation or case, the data are described as *multivariate*. If two characteristics are recorded for each observation, then the data are described as *bivariate*. If only one characteristic or variable is derived, the data are described as *univariate*.

III. Notation

A. General Notation

1. Letters such as *x*, *y*, or *z* are used to identify variables.

2. Specific values of a variable for an individual observation are denoted using subscripts, as in x_i, which represents the value of the i^{th} observation for *x*.

3. Upper-case *N* is used to indicate the number of observations in a population, whereas lower-case *n* gives the number of observations in a sample.

B. The Summation Operator

1. One of the most frequent mathematical operations that is necessary in understanding and using statistics is repetitive addition or summation. This operation is represented by the Greek capital letter sigma "Σ" which is called the summation operator. This operator as interpreted as meaning, "take the sum of the following elements". The summation operator also may include summation limits in the form of the beginning and ending point for subscripted variables (although these may be left out if these limits are fairly obvious in context).

2. The following examples illustrate the use of Σ and some of its major properties:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n$$

$$\sum_{i=1}^{4} y = y + y + y + y$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$$

$$\sum_{i=1}^{n} (x_i + y_i)^2 = \sum (x_i^2 + 2\, x_i y_i + y_i^2) = \sum x_i^2 + 2 \sum x_i\ y_i\ + \sum y_i^2 \neq\ \sum x_i^2 + \sum y_i^2$$

This last formula illustrates the property that if *k* is a constant:

$$\sum kx_i = \ kx_1 + kx_2 + \ldots + kx_n\ =\ k \sum x_i$$

Other properties involving summations are:

$$\sum x_i + \ k\ \neq \sum (x_i + k)$$

$$\sum x_i y_i \neq\ (\sum x_i)(\sum y_i)$$

$$\sum x_i^2 \neq\ (\sum x_i)^2$$

FREQUENCY DISTRIBUTIONS

I. The Concept of a Frequency Distribution

A. Definition

1. "A *frequency distribution* of a variable is a description of the frequencies with which various mutually exclusive and exhaustive categories of observation occur." (Olson) *Mutually exclusive* "means that no observation can be in more than one category," and *exhaustive* "means that the categories include all the observations."

2. A frequency distribution indicates how many observations on a characteristic or attribute of an entity (the thing about which we want to collect information) fall within each of a number of categories. The categories are generally constructed to encompass the entire range of observation values and to be equal in size.

3. A *relative frequency distribution* or a *percentage distribution* includes a third column which provides the percentage of the total number of observations represented by the number of observations in the category (i.e., the number of observations in the category divided by the total number of observations, this quotient then being multiplied by 100).

4. A *cumulative frequency distribution* indicates the number of observations or cases which are less than (or greater than) a particular value. A *cumulative percentage distribution* can be obtained by computing the percent of total observations for each frequency in the cumulative frequency distribution.

B. Constructing a Frequency Distribution

1. A frequency distribution is generally portrayed via a tabular or graphic representation of the data. This allows summarization of much of the information contained in the population or sample.

2. To construct a frequency distribution for quantitative data (interval or ratio scale), the number or size of the categories (or classes) is first identified. Based on the category limits or range, the categories are then laid out in either increasing or decreasing order. Finally, the observations are distributed to each category based on their values, with either a graphic or numeric tally being kept of how many observations fall within each category.

3. A qualitative (or nominal) frequency distribution for non-quantitative data is constructed as above, but the order of the categories is irrelevant.

4. The choice of category size is based on the purpose for which the distribution will be used and the nature of the data. The intent is to make the data easy to understand. Using a large number of categories may not provide enough summarization, whereas too few categories may obscure some of the information contained in the data. A good rule of thumb is to use between 5 and 15 equal-sized categories.

5. Another suggestion is to make the lower category limit of each category a multiple of the interval size. For example, if the category interval or size is 10, then the category limits should run from 0-10, 10-20, 20-30, etc. This aids readability.

6. Sturges' rule provides a formula for an estimate of the number of categories, and is $c = 3.3 (\log n) + 1$, where c is the number of categories and log n is the base 10 logarithm of the sample size.

II. Tabular and Graphic Display of Frequency Distributions

A. Frequency Tables - A tabular display of a frequency distribution generally consists of identifying the category limits in the first column, followed in the second column by the number of observations falling within the specified limits.

B. Graphic Representations of Frequency Distributions

1. A *histogram* provides a visual display of a frequency distribution via a bar graph. The horizontal axis depicts the category ranges for the variable, whereas the vertical axis represents frequency. A rectangle of the appropriate height is then generally used to display the frequency for each category (note that, for proper interpretation, this dictates using categories of equal size).

2. A *frequency polygon* is similar to a histogram, except that the midpoint at the top of each rectangle is connected by a line segment. A cumulative distribution presented in this form is termed an *ogive*.

3. *Bar charts* are used for graphic representation of qualitative data. These are differentiated from histograms by leaving a small space between each rectangle, thus indicating that the categories are not contiguous.

4. *Pie charts* can also be used to represent qualitative frequency distributions, but relative frequencies (percentages) need to be computed to determine the proportion of the "pie" that should be allocated to each category.

5. Care needs to be exercised in using graphic representations of frequency distributions to keep them from being misleading.

C. Characteristics of frequency distributions especially apparent in graphic form include *dispersion* (how tightly the distribution is clustered around a single value), *symmetry* (each half of the distribution being a mirror image) or *skewness* (an unbalanced distribution), and *kurtosis* (peakedness).

CENTRAL TENDENCY AND DISPERSION

I. The Concept of Central Tendency

A. For frequency distributions of interval or ratio scaled variables, there are important summary measures which provide an indicator of the central tendency of the distribution. The term *central tendency* refers to the most typical or most likely value for the characteristic for which the distribution is derived. There are three general measures of central tendency: the mode, median, and arithmetic mean.

B. General Measures of Central Tendency

1. Mode - this is the most frequently occurring value or score of a variable.

   a. The mode does not have to be centrally located in the distribution, and may not be unique. It can be computed for categorical or ordinal data, as well as for interval or ratio scale data.

   b. There is no formula for computing the mode. It is derived simply by inspecting the distribution.

2. Median - the median is defined as the value of the variable that divides the number of observations in half. It is the score for which half of the observations in the distribution are larger and half smaller.

   a. The median is also termed a *positional measure* or *measure of relative location* since it locates the position of some typical case relative to the other cases.

3. Arithmetic Mean

   a. The mean for a sample is written as $\bar{x}$ and is defined as the sum of the scores or values of a variable divided by the total number of cases.

   b. The mean for a population is written as μ (the Greek letter *mu*).

   c. The formulas for the sample and population mean are, respectively:[1]

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_4}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n} \qquad \textbf{(1.1)}$$

$$\mu_x = \frac{x_1 + x_2 + x_3 + \ldots + x_4}{N} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum x}{N} \qquad \textbf{(1.2)}$$

[1] Marchal, Mason, and Wathan use a capital *X* (instead of a lower case *x*) in their formulas to represent a particular value.)

4. Properties of the Mean

   a. The sum of the deviations (differences) of each score from the mean is zero:

$$\sum(\bar{x} - x_i) = (\bar{x} - x_1) + (\bar{x} - x_2) + \ldots + (\bar{x} - x_n) = \sum \bar{x} - \sum x_i = n\bar{x} - \sum x_i$$

$$= n\frac{\sum x_i}{n} - \sum x_i = \sum x_i - \sum x_i = 0$$

$$\sum(x_i - \bar{x}) = \sum(\bar{x} - x_i) = 0 \qquad \textbf{(1.3)}$$

   b. The sum of the squared deviations of each score or value from the mean is less than the sum of the squared deviations about any other number:

$$\sum(x_i - \bar{x})^2 = \text{minimum} \qquad \textbf{(1.4)}$$

   Let $\tilde{x}$ be any other number. If we consider the deviations of x_i around this number, and then subtract the true mean, we get:

$$\tilde{x} - x_i = (\tilde{x} - \bar{x}) - (x_i - \bar{x})$$

$$(\tilde{x} - x_i)^2 = (\tilde{x} - \bar{x})^2 - 2(\tilde{x} - \bar{x})(x_i - \bar{x}) + (x_i - \bar{x})^2$$

$$\sum(\tilde{x} - x_i)^2 = \sum(\tilde{x} - \bar{x})^2 - \sum 2(\tilde{x} - \bar{x})(x_i - \bar{x}) + \sum(x_i - \bar{x})^2$$

$$= \sum(\tilde{x} - \bar{x})^2 - 2(\tilde{x} - \bar{x})\sum(x_i - \bar{x}) + \sum(x_i - \bar{x})^2$$

$$= \sum(\tilde{x} - \bar{x})^2 + \sum(x_i - \bar{x})^2 \text{ since } \sum(x_i - \bar{x}) = 0$$

$$= n(\tilde{x} - \bar{x})^2 + \sum(x_i - \bar{x})^2$$

   In this last equation, the first term will always be positive since it is squared. Because the second term is the sum of the squared deviations about the true mean, any number ($\tilde{x}$) other than the mean appearing in the first term results in a larger value than will the true mean. This is called the *property of least squares*.

5. Calculation of a weighted mean is sometimes useful. A *weighted mean* is derived by taking the means of subsamples or groups, weighting (multiplying) each of these subsample or group means by the number of observations in the subsample or group, summing these values, and then dividing the result by the total number of observations. The formula for a weighted mean is:

$$\bar{x}_{\text{weighted}} = \frac{\sum_{j=1}^{k}(f_j \bar{x}_j)}{n}$$

where k represents the number of groups, f_j represents the number of observations in group j (i.e., the weights), $\bar{x}_j$ represents the group mean, and n is the total number of observations from all groups. The weighted mean is useful when data is presented in histograms or frequency distributions, and the raw data are not available. This is a common occurrence with government reports or other secondary sources of data.

C. Comparison of Mean and Median

1. The mean uses more information than the median - each score or value is used rather than its relative position.
2. Extreme values can have substantial effects on the mean whereas the median will be affected only slightly, if at all.
3. The mean is more stable than the median from sample to sample.
4. The mean is easily manipulated algebraically.
5. The mean requires interval scale data while the median can be used with ordinal scales.
6. Whenever there are considerably more extreme cases in one direction of a distribution (i.e., the distribution is highly skewed), use the median.
7. Often it is best to use both the mean and the median.

II. Other Positional Measures

A. *Quartiles* divide the observations into four quarters containing equal numbers of observations. The value of the first quartile indicates that 1/4 of the scores or observations are smaller.

B. *Deciles* divide the observations into ten groups containing equal numbers of observations. The value of the first decile indicates that 1/10 of the scores or observations are smaller.

C. *Percentiles* or *Centiles* divide the distribution into one hundred ordered segments containing equal numbers of observations. The first percentile value is interpreted analogously to those above.

III. The Concept and Measurement of Dispersion

A. Definition of Dispersion

1. While the mean or median tells us quite a bit about the central tendency of a distribution, they do not give any information concerning the other major characteristic of a distribution. This is the spread or dispersion of the distribution around its center.

2. *Dispersion* is defined as the variability or amount of fluctuation of the observations or cases around the mean of their distribution. It refers to the dispersal, spread, or scattering of the cases around the mean. A *measure of dispersion* is a single number that summarizes this characteristic for the distribution.

3. There are four basic measures of dispersion. They are the range, quartile deviation, mean deviation, and variance/standard deviation.

B. Range

1. The *range* is defined as the difference between the extremes of the distribution (i.e., the highest and lowest scores or values of the variable).

2. The range is generally represented as either the difference between the extremes or by giving the two extreme values.

3. Although the range is simple to derive and is useful in its own right, it is very unstable as a measure of dispersion, since one outlier can dramatically alter its value. Moreover, since the range relies on only two observation values, it uses very little of the information contained in the data.

C. Quartile Deviation

1. The *quartile deviation* is a more useful form of the range that is defined as half the distance between the first and third quartiles. Algebraically, this is:

$$Q = \frac{Q_3 - Q_1}{2}$$

where Q_3 and Q_1 are the third and first quartiles, respectively.

2. By this definition, half of the observations fall within the *interquartile range* defined by $(Q_3 - Q_1)$.

3. The quartile deviation is far more stable and less sensitive to extremes than the range. However, it still does not use most of the information available within the data.

D. Mean Deviation

1. When central location or central tendency is measured via the mean, variability or dispersion would be most logically measured by a statistic based on the distances of each observation's values from the mean. This difference for a single observation or case is termed its *deviation*, which for the i^{th} observation, is written as $(x_i - \bar{x})$.

   a. The most natural summary measure of these deviations might appear to be their sum. However, we can't use the sum of the deviations about the mean, for we have seen that this always equals zero [i.e., $\sum (x_i - \bar{x}) = 0$].

b. In order to utilize these deviations, we need to prohibit negative values. Although one way to accomplish this would be to drop out values smaller than the mean, this procedure would reduce by half the information about the distribution we have at our disposal.

c. There are two general methods for replacing negative deviations with positive ones, while still maintaining the information about distance from the mean for each observation. One method is to take the sum of the absolute values of the deviations from the mean. The other is to take the sum of the squared values of the deviations from the mean. The first leads to a measure of dispersion called the mean deviation. The second leads to the variance and standard deviation.

2. The *mean deviation* or *mean absolute deviation* is defined as the sum of the absolute values of the deviations from the mean of a distribution. Algebraically, it is:

$$\frac{\sum | x_i - \bar{x} |}{n} = \text{mean deviation} \qquad \textbf{(1.5)}$$

3. The mean deviation may be thought of as indicating how much, on the average, the scores deviate or vary from the mean.

4. Although the mean deviation appears to have a fairly straightforward interpretation, it suffers from two limitations -- it is not easily manipulated algebraically, and its relationship to other statistical devices and applications (such as the normal distribution) is not clear cut.

IV. The Variance and Standard Deviation

A. The Variance

1. If we use the method described above of squaring the deviations, we can replace the negative deviations with positive values. This yields the *variance*, which is defined as the sum of the squared deviations about the mean. It is symbolically represented for a population by the symbol σ^2 (sigma squared), and for a sample by the symbol s^2.

   a. The formula for the population variance is:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \qquad \textbf{(1.6)}$$

   b. The formula for the sample variance is:[2]

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad \textbf{(1.7)}$$

[2]Note that the denominator for the sample variance contains $(n - 1)$ rather than n. Subtracting one from the sample

2. Conceptually, the variance is simply the average squared deviation (or difference) of each observation from the mean. It depicts how much variation is found within the population or sample for the specified variable.

3. The variance is one of the most important concepts associated with statistical inference, and one which is repeatedly used to derive other statistics and to enable a large number of statistical methods. Understanding its meaning is crucial to every aspect of statistical analysis.

B. The Standard Deviation

1. One problem with the variance is that it is in units which are the square of those found in the original data. This makes an intuitive interpretation of the variance somewhat difficult. To correct for this problem, we take the square root of the variance, thereby converting this measure of dispersion back into the same units as the data.

2. The *standard deviation* is defined as the square root of the variance.

   a. The formula for the population standard deviation is:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \qquad \textbf{(1.8)}$$

   b. The formula for the sample standard deviation is:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \qquad \textbf{(1.9)}$$

3. The greater the spread of a distribution about the mean, the greater will be the standard deviation. However, as with the mean, extreme values give rise to extreme deviations, which in turn have substantial impacts on the value of the standard deviation (as a result of squaring the deviations). When a few extreme cases can give misleading results, it may be better to use the median and quartile deviation. Otherwise, the standard deviation is the best indicator of a distribution's dispersion.

4. An alternative formula for the standard deviation, that is easier to compute with a calculator, can be derived as follows:

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

size provides an *unbiased* estimate of the population variance, which is a desirable property for any estimator. The reason for this will be addressed during our discussion of statistical inference.

$$= \sqrt{\frac{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + \left(\frac{\sum x_i}{n}\right)^2 n}{n-1}} = \sqrt{\frac{\sum x_i^2 - 2\frac{(\sum x_i)^2}{n} + \frac{(\sum x_i)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} \qquad \textbf{(1.10)}$$

C. Standard Scores

1. The primary purpose of standard scores is to represent a variable in terms of the number of standard deviation (s.d.) units between each of its observation values and its mean. Thus, the data is then standardized and devoid of any of the original units of measurement. When data are represented in standard scores, they will have a mean of zero and standard deviation of one.

2. A *standard score* is defined as the deviation of an observation divided by the standard deviation. It is represented by *z* and often called a "*z*-score." The formulas for standard scores, for sample and population data respectively, are:

$$z_i = \frac{x_i - \bar{x}}{s} \text{ and } z_i = \frac{x_i - \mu}{\sigma}$$

D. Example of calculation of the mean, variance, standard deviation, and standard scores for a sample consisting of four observations (1, 2, 8, 5):

$$\bar{x} = \frac{1+2+8+5}{4} = 4$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(1-4)^2 + (2-4)^2 + (8-4)^2 + (5-4)^2}{4-1} = 10$$

$$s = \sqrt{10} = 3.162$$

$$z_i = \frac{1-4}{3.162} = 0.949; \; z_2 = \frac{2-4}{3.162} = -0.633$$

$$z_3 = \frac{8-4}{3.162} = 1.265; \; z_4 = \frac{5-4}{3.162} = 0.316$$

E. *Chebyshev's Theorem* provides a method for interpreting the standard deviation. It states that for <u>any</u> set of quantitative data, a minimum of $100(1 - 1/k^2)$ percent of the data will lie within *k* standard deviations of the mean, assuming that $k \geq 1$. A better method of interpretation is derived from the *Bienayme'-Chebyshev Inequality*. This rule of thumb for symmetric, unimodal distributions states that approximately 68% of the observations will be found within a distance of one standard deviation from the

mean, 95% will be found within two standard deviations from the mean, and over 99% will be found within three standard deviations from the mean. (For a normal distribution, this rule of thumb becomes an exact property, and is sometimes referred to as the *Empirical Rule*.)

F. Another measure of variation related to the standard deviation is the *Coefficient of Variation*. This is generally denoted by CV, and is defined as the standard deviation expressed as a percentage of the mean, or CV = 100 ($s / \bar{x}$). By dividing the standard deviation by the mean, the CV controls for the type of measurement unit, and allows comparing variation between two sets of data that use different units of measurement.

PROBABILITY THEORY

I. Statistical Definition of Probability

A. The Relationship between Probability and Statistics

1. A major aspect of statistical inference is determining the degree of uncertainty involved when attempting to generalize data from a sample to a population. Probability theory is used to measure this uncertainty.

2. Specifically, probability theory provides an evaluation of the likelihood of drawing a particular sample from a given population. Based on this likelihood, it is possible to quantitatively assess the degree of confidence we can place on the ability of sample statistics to represent the respective population parameters.

B. The Concept of Relative Frequency

1. An *experiment* in probability theory is defined as a procedure or process for generating and observing outcomes. Each repetition of the experiment is termed a *trial*. An example of such an experiment is tossing a coin 100 times, with each toss representing a trial.

2. The set of all possible outcomes of an experiment is termed the *sample space*.

3. An *event* is defined as an outcome or set of outcomes for an experiment (i.e., tossing a coin). When the event occurs it is termed a *success*, when the event does not occur, it is termed a *failure*.

4. *Relative frequency* is defined as the number of times an event occurs as a proportion of all trials in the experiment. If in tossing the coin one hundred times (100 trials), 46 heads occur, the relative frequency of heads is 46/100 or 46 percent.

5. An event can be further broken down to a *simple event*, which is not decomposable (e.g., the occurrence of a head in one flip), or a *compound event*, which is a combination of single events (e.g., the occurrence of heads in two consecutive tosses or trials).

C. Definition of Probability

1. In order to arrive at a mathematical definition of probability, we need to explore the idea of a limit or a limiting value. Many functions exist where, as we increase or decrease the value of the abscissa, the ordinate will approach a particular value. In order to define the probability of an event, we must look at the limiting value of the relative frequency.

2. If, as we increase the number of trials for an experiment, the relative frequency of successes to total trials approaches a particular value, we define this value as the *probability* of the event. Thus, the probability of an event represents the likelihood or chance that the event will occur.

3. In mathematical terms:

$$P(E) = \lim_{n \to \infty} \frac{x}{n}$$

where E = an event for a given experiment
n = the number of trials
x = number of occurrences of the event (i.e., successes)

4. We can often arrive at an approximation of a limit or probability for an event given a finite number of cases by plotting the relative frequency as n increases.

5. When we talk about a single event, (for example, the probability of selecting an Ace from a full deck of playing cards), we are really asking what the limiting value of the relative frequency is for that event over the "long run."

D. Independence, Mutual Exclusion, and Conditional Probabilities

1. Two events are *independent* if the occurrence of any one event does not depend on the occurrence of the other event.

2. Two events are *mutually exclusive* if they cannot occur simultaneously in the same experiment.

3. For example, if in an attempt to draw two Aces, two cards are drawn from a deck with the first card replaced before the second is drawn, the events are <u>independent</u>. On the other hand, if we do not replace the first card before attempting to draw the second, the probability of obtaining an Ace on the second draw will be influenced by the first draw. In other words, "with replacement," the probability of drawing an Ace on the second draw will be the same as that of drawing an Ace on the first. "Without replacement," the probability of drawing an Ace on the second draw will be <u>dependent</u> on what occurred during the first draw.

4. If the occurrence of an event B depends on the occurrence of another event A, we say that B is "conditional" upon A. This is the concept of *conditional probability*, indicated by $P(B \mid A)$. This is read as the probability of B given A.

5. In terms of conditional probabilities, two events A and B are independent if $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$. Two events A and B are mutually exclusive if $P(A \mid B) = P(B \mid A) = 0$, since the occurrence of one precludes the other.

6. Conditional probability can be expressed as the ratio of a *joint probability* (two or more events occurring together) to the simple probability of the initial event: $P(B \mid A) = P(A \cap B) / P(A)$, assuming that $P(A) \neq 0$.

II. Mathematical Properties of Probabilities

A. The maximum and minimum values for the probability of an event are 1 and 0, or $0 \leq P(E) \leq 1$.

B. Addition Rule or Sum Rule

1. For a set of mutually exclusive events *A, B, C, . . . K*, the probability of getting either *A* or *B* or *C . . .* or *K* is equal to the sum of all the individual probabilities, or

$$P(A \cup B \cup C \ldots \cup K) = P(A) + P(B) + P(C) + \ldots + P(K)$$

For example, the probability of drawing a face card (*J, Q, K*) from a standard deck of 52 cards = 4/52 + 4/52 + 4/52 = 3/13

2. A more general addition rule can be stated if the assumption of mutually exclusive events is relaxed, implying that we are looking for the probability of either event *A* occurring or event *B* occurring or both *A* and *B* occurring:

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

For example, the probability of drawing a red card or an Ace is:

$$P(\text{Ace}) = 4/52,\ P(\text{Red}) = 26/52,\ P(\text{Red Ace}) = 2/52$$
$$P(\text{Ace} \cup \text{Red}) = 4/52 + 26/52 - 2/52 = 28/52 = .54$$

C. Multiplication Rule or Product Rule

1. The multiplication rule for independent events *A, B, C, . . . K* is:

$$P(A \cap B \cap C \ldots \cap K) = P(A)\,P(B)\,P(C) \ldots P(K)$$

For example, the probability of drawing an Ace, King, and then a Spade with replacement is: $P(\text{Ace} \cap \text{King} \cap \text{Spade}) = (4/52)(4/52)(13/52) = .00148$

2. The multiplication rule for dependent events *A* and *B* is:

$$P(A \cap B) = P(A)\,P(B \mid A)$$

or for dependent events *A, B,* and *C* is:

$$P(A \cap B \cap C) = P(A)\,P(B \mid A)\,P[C \mid (A \cap B)]$$

For example, the probability of drawing two Aces without replacement is: $P(\text{Ace} \cap \text{Ace}) = (4/52)(3/51) = .0045$

whereas the probability of drawing four Aces without replacement is:
$P(\text{Ace} \cap \text{Ace} \cap \text{Ace} \cap \text{Ace}) = (4/52)(3/51)(2/50)(1/49) = (3.69)(10^{-6})$

III. Permutations & Combinations

A. Whenever we look for events occurring in a specified order, the order of events is known as a *permutation*. The examples in the preceding section each represent a single permutation of the possible outcomes. If we are not concerned with order, the group of events is simply termed a *combination*. When the order in which events occur is unimportant, the calculation of probabilities can get more complicated. The fundamental theorem of counting can be used to develop a rule for evaluating such probabilities.

B. The *Fundamental Theorem of Counting* states that, "If an event can occur in *m* ways, and if after it has occurred, a second event can follow it and occur in any one of *n* ways, then the two events can occur together, in the order stated, in (*m*)(*n*) different ways (. . . can be extended to any number of events)." (Weimer)

C. For any *n* distinct events, there are *n*! (*n* factorial) possible permutations. For example, the number of permutations of three cards such as Ace, King, and Queen is *AKQ*, *QAK*, *AQK*, *KQA*, *KAQ*, *QKA* or *n*! = 3! = (3)(2)(1) = 6.

D. General formulas for the number of permutations of *n* objects taken *n at a time*, and the number of permutations of *n* objects taken *r at a time*, are respectively:

$$ {}_nP_n = n! \qquad {}_nP_r = \frac{n!}{(n-r)!} $$

E. The number of combinations of *n* objects taken *r* at a time is given by the formula for the *binomial coefficient*, which is:

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n\ -\ r)!} $$

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

I. Random Variables

A. Definition

1. A *random variable* is defined as "a function that assigns a numerical value to each outcome in an experiment whose results depend to some extent on chance" (Olson), or less precisely as "a rule (or function) that assigns unique real numbers to each outcome in a sample space of an experiment." (Weimer)

2. The most distinguishing characteristic of a random variable (RV) is the probabilistic component associated with the process of determining its value. Any such process is termed *stochastic*, thus leading to the alternative name of *stochastic variable* for an RV.

B. The Concept of a Random Variable

1. Another way of distinguishing an RV from a non-random variable is that the specific outcome of a trial associated with the RV is not known or cannot be determined exactly in advance. Instead, information concerning the relative frequency of each outcome is theoretically or empirically calculable, allowing each outcome to be assigned a probability of occurrence if we were to have the requisite relative frequency data.

2. For example, it is the responsibility of state environmental management agencies to keep track of the quality of community drinking water supplies. Assuming that such a state agency wants to determine the overall quality of drinking water within its jurisdiction, the agency could take a *random sample* (a sample in which each member of the population, i.e., each water supply, has an equal chance of being selected) of drinking water from all community water supplies. The sampling process, due to its probabilistic nature, dictates that the water quality variable will be an RV. If, on the other hand, we had determined the exact water quality level for every community water supply in the state for a given point in time, the water quality variable would no longer be random.

3. Other examples of non-random variables would include whether or not each student in this class studied last night, the height of each student in the class, the hair color of each student who is a SPEA major, etc. But it is important to note that whether or not a variable is an RV depends on the respective population. If we are going to use data collected from this class as representative of the entire population of first year graduate students in SPEA, and each first year student had an equal chance of being enrolled in this class, then "whether or not you studied last night," height, and hair color each become RVs. Thus, a variable's random or non-random nature depends on the purpose to which it will be put. Perhaps Hamlet should have asked the question, "To be or not to be a random variable?"

4. The probabilistic component of an RV may come from a number of different types of processes -- not just from sampling. It is often the case that we either do not fully understand a process that generates outcomes, or cannot include all of the

factors that contribute to those outcomes in our analysis. A common example is in weather forecasting, in which we neither fully understand the process nor could include all the myriad factors that determine local weather. Thus, predicting the weather is a probabilistic process, with the resulting variables of temperature, precipitation, humidity, and barometric pressure all being random. Many forecasting processes have this characteristic, even in situations in which the entire population for the current time period is at hand. If we attempt to predict whether or not students in this class will study tonight, even if we know whether you studied last night, we would certainly be in the realm of RVs.

C. Discrete versus Continuous Random Variables

1. If a random variable can assume only integer values (including negative integers), it is termed a *discrete random variable*. If it can assume decimal values, it is termed a *continuous random variable*.

2. For example, any RV involving measurement (height, weight, length, area, etc.) will be continuous, since it can take on non-integer values. On the other hand, any RV involving "the number of" entities associated with a process or characteristic will be discrete.

II. Probability Distributions and Functions

A. Probability Distributions and Random Variables

1. The values of a random variable can be described in terms of a probability distribution. A *probability distribution* is a relative (percentage) frequency distribution indicating the probability of occurrence for each value of an RV.

2. For a discrete RV, a table of these probabilities, termed *a probability distribution table* or a *probability table*, can be constructed to show the discrete probability distribution. Such a table shows all possible values of the RV (i.e., the entire sample space) and the probabilities associated with their occurrence.

3. For example, a local police department interested in assessing the number of fatalities that are likely to occur in traffic accidents over the next Memorial Day weekend might construct the following table from relative frequency data obtained from previous years:

x	$f(x)$
0	0.15
1	0.25
2	0.30
3	0.20
4 or more	0.10

where $f(x)$ is the relative frequency (or probability) of taking on a particular value x. Notice that the probability in such tables must add up to 1.0, since all possible outcomes for the RV are present.

B. Probability Functions

1. We can assign probabilities to the values of an RV via a mathematical rule or *probability function* instead of (or in addition to) a table. Such a probability function provides a mathematical formula for calculating the probability associated with each outcome for the RV in the sample space. This is represented mathematically by $f(x)$.

2. For example, $f(x) = (2 + x) / 3$ for $x = -1, 0$ yields the values 1/3 and 2/3 for all values of the random variable x, and is a probability function.

3. Probability functions or tables are often depicted in graphic form. For discrete RVs, these graphs are composed of vertical line segments whose height above the x-axis indicates the probability for the associated value of the RV. All of the heights must sum to unity.

4. A *probability density function* is a function of a continuous random variable that portrays the probability of occurrence for values of the variable in defined intervals as the area under the curve delineated by the interval. The total area under the curve is defined as equal to one. In other words, "the probability that on a given trial the random variable x will take on a value between any two values x_i and x_j is equal to the proportion that the area under the density function's curve between x_i and x_j is of the total area under the curve."

5. A *distribution function* or *cumulative probability density function* shows the cumulative probability up to specified values of a random variable. The cumulative distribution function is often indicated by $F(x_i)$, and gives the probability of occurrence for a value of the random variable x that is less than or equal to x_i.

III. Expected Value

A. Given k possible outcomes, $x_1, x_2, \ldots x_k$, for an event and the probability for any outcome x_i as equal to $f(x_i)$, then the *expected value* of the random variable x is defined as:

$$E(x) = \sum_{i=1}^{k} x_i f(x_i) \qquad \textbf{(1.11)}$$

or $E(x) = \sum xP(x)$ (Lind et al., 2002).

B. For a random sample of a population where $1/N$ is the probability of an element of the population being selected, this formula becomes:

$$E(x) = \sum_{i=1}^{N} x_i \frac{1}{N} = \sum_{i=1}^{N} \frac{x_i}{N} = \mu$$

Thus, the expected value of a random variable is its mean. In other words, the *mean of a probability distribution* is defined as the expected value of the random variable:

$$E(x) = \mu$$

C. The *variance of a probability distribution* of a random variable can be shown to be the expected value of the squared deviation around the mean:

$$E(x-\mu)^2 = \sigma^2$$

D. The *standard deviation of a probability distribution* or a random variable can be shown to be the positive square root of the variance.

IV. The Binomial Probability Distribution

A. Purpose - Binomial experiments give rise to the first theoretical distribution that we encounter associated with inferential statistics. Important in its own right as a method for assigning probabilities to experiments whose trials have only two possible outcomes, the binomial distribution also illustrates how theoretical distributions are used. Moreover, derivation and description of the discrete binomial distribution provides a needed precursor to a discussion of continuous distributions.

B. Derivation and Definition

1. A *binomial experiment* is defined as an experiment that consists of n identical independent trials, where each trial has only two possible outcomes (commonly identified as "success" or "failure") and the probability of success for each trial is the same.

2. If we define π as the probability of success for a trial in a binomial experiment, $1 - \pi$ as the probability of failure for a trial, and we assume that there are n independent trials, then the probability of obtaining x successes and $(n - x)$ failures <u>in a specified order</u> is found via the multiplication rule for probabilities as:

$$\underbrace{\pi \text{ times } \pi \text{ times } \pi \ldots \pi}_{x \text{ terms}} \; \underbrace{(1-\pi) \text{ times } (1-\pi) \text{ times } (1-\pi) \ldots (1-\pi)}_{(1-\pi) \text{ terms}} = \pi^x (1-\pi)^{(n-x)}$$

3. Based on the rule for combinations, the probability of obtaining n successes <u>in any order</u> can be found by:

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

where C represents the number of combinations of n items taken x at a time. Thus, for any given outcome of the experiment, the formula for the probability of that outcome is:

$$P(x) = \binom{n}{x} \pi^x (1-\pi)^{(n-x)} = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{(n-x)} \qquad \textbf{(1.12)}$$

or $P(x) = {}_nC_x \pi^x (1-\pi)^{(n-x)}$ (Lind et al., 2002).

where π is the probability of success of a trial and $1 - \pi$ is the probability of failure of a trial.

Probability =	No. of ways of getting *x* successes	times	Probability of *x* successes of any given way

This formula associates probabilities with each given outcome of the experiment, and thus is consistent with the definition of a probability function. It gives rise to a probability distribution termed the *binomial distribution*.

C. Given formula 1.12 for the binomial probability function, a probability distribution table or graph can be constructed for any specific combination of the three parameters *x*, *n*, and π.

1. Using the binomial function to derive a graph or table provides a *theoretical binomial distribution*. Generating a graph or table from relative frequency data obtained from repeating a binomial experiment yields an *empirical binomial distribution*. As the number of repetitions of the experiment increase, the empirical binomial distribution will approach the theoretical distribution. Thus, if the number of repetitions is large enough, we can use the theoretical distribution to accurately represent the empirical one. Both of these distributions assign a probability to each possible value of the random variable *X* (the number of successes in *n* trials).

2. Appendix A of Lind, Marchal, and Wathan provides distributions for various combinations of these parameters. This table is generated from formula 1.12, and can be used in place of the formula whenever the appropriate parameters appear in the table. Otherwise, the formula must be used to derive the distribution. To generate a graph of a binomial distribution, the value of *x* is represented on the *x*-axis and the value of $P(x)$ is represented by the *y*-axis.

D. The values of *n* and π determine the shape of the binomial distribution graph. From the respective formulas for a random variable, the mean, variance, and standard deviation of a binomial distribution can be derived. These are, respectively:

$$\mu = n\pi \qquad \sigma^2 = n\pi(1-\pi) \qquad \sigma = \sqrt{n\pi(1-\pi)}$$

E. Example - Use the binomial function to find the probabilities of obtaining *x* heads in 10 coin flips where $x = 0, 1, 2, \ldots, 10$:

$$P(0) = \frac{10!}{0!(10-0)!}(.5)^0(.5)^{10} = (1)(1)(.000977) = .001$$

$$P(5) = \frac{10!}{5!(10-5)!}(.5)^5(.5)^5 = .246094 = .246$$

$$P(10) = \frac{10!}{10!(10-10)!}(.5)^{10}(.5)^0 = .000977 = .001$$

No. of Heads	$P(x)$	No. of Heads	$P(x)$
0	.001	6	.205
1	.010	7	.117
2	.044	8	.044
3	.117	9	.010
4	.205	10	.001
5	.246		

THE NORMAL DISTRIBUTION

I. Continuous Probability Distributions

A. A *continuous probability distribution* provides the probabilities associated with outcomes of a continuous random variable. By definition, the number of possible outcomes for a continuous variable is theoretically infinite, and limited practically only by the precision of the measurement instrument. Thus, any single value or point within such a distribution can be represented with an infinite number of decimal places. The probability of occurrence of any specific value (e.g., 1.2375869847756868...) will be equal to zero. Thus, probabilities for continuous distributions are generally determined within specific intervals.

B. The mathematical function or formula used to determine the probabilities associated with a continuous random variable is termed a *probability density function* (pdf), since the likelihood of occurrence of a range of values is determined by the density or area under the curve defined by the function.

C. There are many continuous probability distributions which are used in statistical inference. Some of the most common ones are the normal distribution, Student's *t*-distribution, the *F* distribution, the χ^2 distribution, and the exponential distribution.

II. Introduction to the Normal Distribution

A. Concept

1. Assume we measure the height of 50 adult men and then derive a frequency distribution in the form of a histogram -- the frequencies would pile up in the neighborhood of the mean and fall off as we moved toward either side, although this distribution would be somewhat irregular. Increasing the number of cases to 200 would add more regularity to the distribution and provide greater symmetry. With 10,000 cases, the distribution would become very regular, symmetric, and centered about the mean.

2. If we then connect the midpoints and smooth out the graph, we will have a figure which approaches the theoretical probability distribution termed the normal distribution.

3. The use of this type of curve as a probability distribution is credited primarily to Johann Gauss, who lived between 1777 and 1855. Thus, it is also called the *Gaussian distribution* in honor of his work.

B. Definition of the Normal Distribution

1. The *normal distribution* is a continuous probability distribution for the random variable *x* which is based on the following probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-(x-\mu)^2/2\sigma^2} \qquad \textbf{(1.13)}$$

where:

$f(x)$ = is the probability density function for the random variable x,
π = 3.14159...,
e = 2.71828...,
μ = the mean (or expected value) of the random variable x, and
σ^2 = the variance of x

C. Characteristics

1. The curve is bell shaped.

2. The theoretical normal curve never actually touches the x-axis or baseline but instead approaches it asymptotically.

3. The mode, median, and mean all occur at the same point, implying that the normal distribution is symmetric around the mean.

4. The area between the curve and the x-axis will always be equal to 1.

5. There is not any one normal curve but a family of normal curves. A normal curve is fully defined when both its mean and variance are specified.

6. For <u>any</u> normal curve, the area between the mean and any ordinate (observation) which is specified as a distance from the mean in terms of standard deviation units, is constant.

D. Characteristic 6 above leads to another property of normal curves, which is an exact description of the proportion of any normal curve that can be found within a certain number of standard deviation units away from the mean. This is often termed the *empirical rule*, and can be stated as:

1. 68.27% of the area under the curve will be found within one standard deviation of the mean (i.e., $\mu \pm \sigma$).

2. 95.45% of the area under the curve will be found within two standard deviation units of the mean (i.e., $\mu \pm 2\sigma$).

3. 99.73% of the area under the curve will be found within three standard deviation units of the mean (i.e., $\mu \pm 3\sigma$).

III. Calculating Probabilities (Areas) Under the Normal Distribution

A. If we apply the standard score or z-score transformation (i.e., dividing each deviation by the standard deviation of the distribution) to a normal curve, the result is termed the *standard form of the normal distribution* or just the *standard normal distribution*. In this form, the equation for the normal curve becomes:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad \textbf{(1.14)}$$

indicating that the curve will have a mean of 0 and a standard deviation of 1. There is only one such standard normal curve, since it is fully specified by the values of the mean and standard deviation. The *z*-scores for this curve are termed *standard normal deviates*.

B. Area Relationships of the Standard Normal Distribution

The area under the standard normal distribution between any two values of *z* is extremely important, since this area represents probabilities. The proportion of the distribution involved is computed via transforming the observation values into standardized scores and using a table of areas under the standard normal distribution.

C. To determine the proportion of the area (i.e., probability) demarcated by a specific *z*-score for the standard normal distribution, we could theoretically use formula 1.14. Plugging in a specific value of *z* would yield the height of the curve above the *x*-axis at this point. However, this is not very useful, since probabilities for a continuous distribution need to be calculated across an interval. In order to do so, we could take the integral of the curve in equation 1.14 over the specified interval. Since this is a non-trivial exercise, it is convenient to have these values already calculated in the form of a table of "Areas Under the Normal Curve" in the back flyleaf of Lind, Marchal, and Wathen. This table gives the proportion of the area of the standard normal curve between any *z*-score and the mean.

D. For example, to determine the probability of a *z*-score > 1.85, the value in the table for $z = 1.85$ is 0.4678. Since this represents the proportion of the curve between 1.85 and 0, we would need to subtract this from 0.5 to determine this probability. Thus, $P(z > 1.85) = 0.5 - 0.4678 = 0.0322$. We would expect to find a *z*-score greater than 1.85 only 3.22 percent of the time.

To determine the probability of obtaining a *z*-score between –1.33 and 1.33, we would look up the value in the table for $z = 1.33$, which is 0.4082. Since this represents the proportion of the area between 0 and 1.33, and since the curve is symmetric, the area between –1.33 and 0 is also 0.4082, implying that the probability of obtaining a *z*-score in this range is 0.8164.

PART TWO: STATISTICAL INFERENCE

SAMPLING METHODS AND SAMPLING DISTRIBUTIONS

I. Sampling

A. It is generally not possible or practical to obtain data on an entire population. Thus, a sample of the population is required from which we can generalize to the entire population. The process for obtaining the sample is termed *sampling*.

B. The most common form of sampling is *simple random sampling*. The key concept associated with this method of obtaining the sample is the assurance that each entity within the population has an equal chance of being selected for inclusion in the sample. According to Olson, "A *random sample* is one in which every member of the population of relevant observations has equal probability of being selected. More precisely, a *simple random sample* of *n* observational units is one that is chosen in such a way that every possible combination of *n* observational units that could be drawn from the population has equal probability of being selected." A table of random numbers is generally used to obtain such a sample.

C. Several other types of random samples are also often used. These are systematic sampling, stratified random sampling, and cluster sampling.

1. A *systematic sample* is a random sample for which a rule or system is utilized to select the observations to be included in the sample, generally from a list of members of the population.

2. A *stratified random sample* "involves separating the population into non-overlapping groups, called *strata*, and then selecting a simple random sample from each stratum. The information from the collection of simple random samples would then constitute the overall sample." (Weimer) However, to make the sample representative, the number of cases selected from each strata should be proportional to the percentage of the strata in the entire population. Stratified sampling is generally used to make sure that every strata is represented in the sample. We would expect each strata to have characteristics different from the other stratas, implying that most of the variation in the data will be between strata, with less variation within strata.

3. *Cluster sampling* uses a multistage sampling procedure where the population is so large or dispersed as to make a simple random sample too costly. The population is divided into a large number of groups, with a subset of the groups selected via random sampling. A single-stage cluster sample would then include the entire population within all of the sampled groups. A two-stage cluster sample would include a second round of simple random sampling to generate a sample of the population within each sampled group. In cluster sampling, the expectation is that each cluster would have characteristics that were very similar to those of any other cluster, implying that most of the variation in the data will be within clusters rather than between clusters.

D. Once a random sample is obtained, it can be used to generate estimates of the population parameters. Remember that a *parameter* provides a numerical summary about the characteristics of a population, whereas a *statistic* provides a numerical summary about the characteristics of a sample. These include the mean, median, mode, variance, and standard deviation of the sample.

E. *Sampling error* is introduced by any sampling procedure, and is defined as the difference between a sample statistic and its corresponding parameter in the population. Note that multistage cluster sampling can add substantially to the amount of sampling error present in the data, since at each stage of the selection, there is a risk of acquiring a non-representative sample.

F. If the sample from which a statistic has been obtained is random, it can provide an *estimate* of the respective population parameter. One definition of an *estimator* is that it is a statistic that is used to provide an estimate of the population parameter.

G. The terms "accuracy" and "precision" relate to how well a statistic performs as an estimate of the respective population parameter.

1. The *accuracy* of an estimate indicates the lack of bias or systematic error in the statistic's representation of the population parameter. A good estimate in this sense is termed *unbiased*.

2. The *precision* of an estimate indicates the lack of variability or random error in the statistic's representation of the population parameter. A good estimate in this sense is one that minimizes random error, i.e., has minimum variability about the actual parameter given repeated sampling.

II. The Sampling Distribution

A. The best way to comprehend the concept of a sampling distribution is to envision a population from which we take repeated samples. For each sample, we calculate a statistic (for example, the mean or standard deviation) and generate a frequency distribution for the statistic. We then determine the relative frequency (i.e., probability) of each value of the statistic, and then plot these values on a graph. In this way, we assign a probability to each value that the statistic can have. This gives us an *empirical sampling distribution*.

B. Definition

1. Since all statistics are generated via a sampling process that has an associated random element, *a statistic is a random variable which will have an associated probability distribution*.

2. One definition of a *sampling distribution* is "the distribution of values taken by the statistic in a large number of samples from the same population" (Moore and McCabe). Weimer defines the *sampling distribution of a statistic* as "the distribution of all possible values of the statistic computed from samples of the same size."

3. It is possible to derive a *theoretical sampling distribution* for a statistic, which provides information about the probability of occurrence of various values of the statistic derived mathematically from theoretical principles (i.e., without repeated sampling from the population). These theoretical distributions form the basis of statistical inference, and include the normal distribution, Student's *t*-distribution, etc.

C. Characteristics - As in the case of any distribution, we can summarize the character of a sampling distribution through the use of measures of location and dispersion.

1. The mean, variance, and standard deviation of a sampling distribution are generally written using Greek letters with a subscript indicating the respective statistic. For example, these quantities for the sample mean are:

$$\mu_{\bar{x}} \qquad \sigma^2_{\bar{x}} \qquad \sigma_{\bar{x}}$$

2. The standard deviation of a sampling distribution is termed the *standard error* of the statistic or estimate. It represents the degree of variability of the value of the statistic from one sample to another, and is thus a way of summarizing the precision of the statistic. Statistics with low standard errors are termed *efficient*, while those that give the least amount of variability among all possible estimators for the specific population parameter are termed *minimum variance*.

3. The center of an *unbiased* statistic's sampling distribution is equal to the actual value of the respective parameter. This represents the estimate's accuracy.

III. Sampling Distribution of the Mean and the Central Limit Theorem

A. Given a random sample of n observations from a population of size N with mean μ and standard deviation σ, then the sampling distribution of the sample mean $\bar{x}$ will have a mean equal to μ and a standard error (i.e., standard deviation) equal to $\sigma/\sqrt{n}$.

1. This indicates that the sample mean is an unbiased estimate of the population mean, and that the variation in the estimate of the population standard deviation can be reduced by increasing the sample size (n).

2. Thus, the precision or efficiency of the estimate of σ depends on the sample size and not the size of the population (assuming N is at least 10 times larger than n).

3. Note that the definition of the standard error as $\sigma/\sqrt{n}$ assumes that the population is either sampled with replacement, or is sufficiently large so that the lack of independence resulting from sampling without replacement is small enough to be irrelevant. The latter will occur whenever the sample size is less than five percent of the population size. Otherwise, the following formula for the standard error of the mean, which contains a correction factor for sampling without replacement, should be used:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

B. Given a random sample of n observations from a *normally distributed* population of size *N* with mean μ and standard deviation σ, then the sampling distribution of the sample mean ($\bar{x}$) will be *normally distributed* with a mean equal to μ and a standard error equal to $\sigma/\sqrt{n}$.

1. The implication of this statement is that the table of the normal distribution can be used to determine the probability of occurrence of certain values for the sample mean. This explicitly allows for a determination of the probability of an error of any specified amount occurring in the estimate of the population mean.

2. An interesting aspect of the sampling distribution for the sample mean is that the sample mean will have a standard error which is smaller than that of the sample median or mode. This is one reason why the sample mean is used in statistics much more extensively than the sample median or mode.

C. The Central Limit Theorem

1. The *Central Limit Theorem* (CLT), also called the *Normal Approximation Rule*, extends point B above to any sample where the sample size is large. This theorem can be stated as:

   As the sample size *n* for a random sample increases, the sampling distribution of the sample mean ($\bar{x}$) will approach a normal distribution with a mean equal to the population mean μ and a standard error equal to $\sigma/\sqrt{n}$.

2. The rule of thumb for the CLT indicates that a sample size of 30 will be sufficient for the sampling distribution of the mean to be approximated by the normal distribution. Thus, for any sample of size 30 or more, the CLT allows use of the table of the normal distribution to determine the probability of obtaining any specific value of $\bar{x}$ for our estimate of μ.

IV. The Binomial Distribution, the Normal Distribution, and Proportions

A. The Normal Distribution as an Approximation of the Binomial Distribution

1. As the number of observations in a sample for a binomial random variable gets large, the binomial distribution empirically approaches the shape and assumes the properties of a normal distribution.

2. Recalling that the mean, variance, and standard deviation for a binomial random variable are respectively given by:

$$\mu = n\pi \qquad \sigma^2 = n\pi\,(1-\pi) \qquad \sigma = \sqrt{n\pi\,(1-\pi)}$$

where π represents the probability of the binomial event occurring in any one trial and $(1-\pi)$ represents the probability that the event does not occur in any one trial, then *a normal distribution with mean μ and standard deviation σ can be used to approximate the binomial distribution if* $n\pi \geq 5$ *and* $n(1-\pi) \geq 5$.

B. The ability to use the normal distribution to approximate the binomial is very useful for dealing with statistical inference related to population proportions. A population proportion generally represents the percentage of a population that has a specific characteristic vs. not having that characteristic. Since the presence or absence of the characteristic is a binomial event, the binomial distribution is appropriate for assigning probabilities to proportional outcomes. A proportion is arrived at by simply dividing each element in the sampling distribution by the sample size *n*. This transformation just re-scales the distribution, and does not change its shape. So if the sampling distribution of the number of success of a binomial random variable is approximately normal with the above mean and standard deviation, then the *sampling distribution of the sample proportion p* (which represents the number of successes divided by the sample size *n*) will be approximately normal with a mean and standard error of:

$$\mu_p = \frac{n\pi}{n} = \pi \qquad \sigma_p = \frac{\sqrt{n\pi\,(1-\pi)}}{n} = \sqrt{\frac{\pi\,(1-\pi)}{n}}$$

C. Once the validity of using the normal distribution is established, *z* scores can be computed using the following formula:

$$z = \frac{p-\pi}{\sigma_p} = \frac{p-\pi}{\sqrt{\pi\,(1-\pi)/n}}$$

[Note that the same correction factor for small samples (i.e., for $n > .05\,N$) without replacement should be utilized for the standard error of *p*.]

POINT AND INTERVAL ESTIMATION

I. Introduction to Statistical Inference and Estimation

A. The primary purposes of statistical inference are:

- to determine, to the best extent possible, population parameters by making inferences from sample statistics; and
- to test hypotheses about the population.

B. A major aspect of inference is obtaining an estimate of a population parameter via the sample data. There are two types of estimates:

1. "A *point estimate* is a single numerical value calculated from sample data and taken to be indicative of the value of the population parameter." (Olson)

   a. Point estimates, by themselves, convey nothing about the precision of the estimate. Thus, the convention is to include the standard error along with the estimate. This is defined as $\sigma/\sqrt{n}$. But we do not know σ, since it is a characteristic of the population. The best guess we have for σ is the sample standard deviation s, which gives us the *estimated standard error* for the mean of a random sample as:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \textbf{(2.1)}$$

   b. A common method for presenting point estimates is to provide the estimate with the standard error, separated by "±".

2. "An *interval estimate* is a set of adjacent numerical values (or an interval on the real number line) determined from sample data and accompanied by a statement of the probability that such an interval includes the population parameter." (Olson)

   a. The interval itself is termed a *confidence interval*, the endpoints of the interval are termed the *confidence limits*, and the probability that the interval contains the population parameter is termed the *confidence level*.

   b. There are two general confidence levels used in statistical inference: 95% and 99%. Occasionally, some research will use a 90% confidence level.

II. Point Estimate of μ

The sample mean is an unbiased estimate of the population mean. This is the best statistic to use in estimating μ. As in all types of point estimates, the standard error should be provided along with the point estimate, $\bar{x}$.

III. Confidence Interval for μ When σ is Known

A. To construct a confidence interval or interval estimate for μ, a decision is first made as to the tolerable risk (probability) of such an interval not including the parameter. The interval is then computed by setting the confidence limits a certain multiple of standard error units on each side of the point estimate. Assuming that the sampling distribution of $\bar{x}$ is normal, then *z*-scores will also be normally distributed with known probabilities of occurrence. Thus, the number of standard error units on each side of the point estimate is based on the *z*-score corresponding to the acceptable probability of the interval not including the parameter.

B. For example, suppose we draw a random sample of 100 cases from a population with a known standard deviation of 14.2 units. The sample mean is 93.5. We wish to compute a confidence interval for μ so that 99% of the time that we compute such an interval, it will contain the actual value of μ. Another way of stating this is that we are willing to be incorrect in our assertion that the parameter lies in the interval one time out of a hundred.

1. To compute a 99% confidence interval for μ, we must determine the *z*-score which corresponds to a confidence level of 99%. The appropriate *z*-score must result in 99% of the area under the normal distribution being contained in the central part of the curve, and 1% being contained in the tails. Since the 1% must be equally distributed in both tails, 0.5% must be contained in each tail.

2. α is used to represent the total area in the tails of the distribution, and is termed the *significance level*. The remaining area, that under the central part of the curve, will then be $(1 - \alpha)$, and is equal to the confidence level. To determine the appropriate value of *z* to use, compute $\alpha/2$, **subtract this from .5**, and look up this value in the table of the normal distribution. In this example, $\alpha/2 = .01/2 = .005$ **and .5-.005 = .495**, which corresponds to a *z*-score of 2.58. In general, the formula for a confidence interval is then:

$$\bar{x} - z_{\alpha/2}\,\sigma_{\bar{x}} < \mu < \bar{x} + z_{\alpha/2}\,\sigma_{\bar{x}} \quad \text{with a } 1-\alpha \text{ confidence level} \qquad \textbf{(2.2)}$$

3. $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 14.2/\sqrt{100} = 1.42$. Thus, the confidence interval for this example is:

$$93.5 - 2.58\ (1.42) < \mu < 93.5 + 2.58\ (1.42) =$$

$$89.84 < \mu < 97.16 \quad \text{with a 99\% confidence level}$$

IV. Confidence Interval for μ When σ is Unknown

A. The ability to construct a confidence interval for μ rests on the assumption that we can use the normal distribution for determining probabilities. When σ is known, we assume that *z*-scores are normally distributed. Since the sampling distribution of $\bar{x}$ is Gaussian (assuming either a normal population or a sample size > 30), and $\bar{x}$ is the only random variable contained in the expression for the confidence interval, this assumption is

justified. However, if σ is not known (as is generally the case), we must estimate it with the sample standard deviation *s*. The estimated standard error would be:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\left(\frac{\sqrt{\sum (x - \bar{x})^2}}{n-1} \right)}{\sqrt{n}} \qquad \textbf{(2.3)}$$

B. The formula for the *z*-score would also need to be changed to reflect the substitution of the estimated standard error for the actual standard error, so that the new formula would be:

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \qquad \textbf{(2.4)}$$

C. Student's *t* distribution

1. Even if the sampling distribution of $\bar{x}$ is Gaussian, equation 2.4 would no longer have a normal distribution. Unlike the formula for *z*-scores, the denominator of equation 2.4 includes a second source of variation -- namely the sample standard deviation. This statistic, like the sample mean, is a random variable (which is subject to sampling error). This formula now contains two sources of variation in the form of both $\bar{x}$ and *s*. And we <u>cannot</u> claim that the sampling distribution of *s* is normal.

2. The expressions contained in equation 2.4 define the *t*-statistic. The sampling distribution of this statistic was investigated by W.A. Gossett, who wrote under the pseudonym "Student." Hence, it is known as *Student's t-distribution*. The *t-statistic* is formally defined as the ratio of a normally distributed point estimate minus the parameter to the estimated standard error of the point estimate.

3. The quantity $(n - 1)$ in equation 2.3 is termed the *number of degrees of freedom*, and indicates the number of values that are unconstrained in calculating the statistic. It represents the amount of information contained in the sample that can be used to obtain the estimate. Since calculation of s requires that we first know $\bar{x}$, one degree of freedom is used up.

4. The distribution of the *t*-statistic differs from the normal distribution primarily when $(n - 1)$ is small. As the sample size increases, the precision of the estimate of the standard deviation improves (i.e., it gets closer to the value of the parameter σ), and the *t*-distribution will approach that of the normal distribution. However, when $(n - 1)$ is small, there is potentially greater variability in the *t*-statistic and its distribution contains more area in the tails and less in the center. It is thus thinner in the center and thicker in the tails.

5. In order to utilize the *t*-statistic as indicated above, it is necessary to assume a normal population. The CLT cannot be used in this instance, since it requires that we know σ exactly -- not via an estimate. Thus, *a normal distribution of the population must be assumed to use the t distribution*. Luckily, the *t* distribution has been shown to be relatively insensitive to moderate departures from the normality

assumption, implying that confidence intervals computed with the *t*-statistic are generally accurate even with these violations. Such a statistic or distribution that can withstand moderate departures from its assumptions is termed *robust* with respect to those assumptions.

D. Determining a Confidence Interval Using the *t* distribution

1. The formula for a confidence interval for μ when σ is estimated by s (i.e., when the *t*-statistic must be used), is:

$$\bar{x} - t_{\alpha/2}\, s_{\bar{x}} \;<\; \mu \;<\; \bar{x} + t_{\alpha/2}\, s_{\bar{x}} \quad \text{with a } 1-\alpha \text{ confidence level} \qquad \textbf{(2.5)}$$

where α is again the significance level.

2. The back flyleaf of Lind, Marchal, and Wathen provides critical values for the *t* distribution based on the confidence level (or level of significance) selected and the number of degrees of freedom ($n - 1$) in the sample. For a large sample (greater than 120 observations), the *t* distribution approaches the normal distribution so closely that it is virtually identical, and a table for the normal distribution can be used.

3. Assuming that σ was not known in the above example but $s = 15.8$, the confidence interval for μ would then be:

$$93.5 - 2.66\,(15.8/\sqrt{100}) \;<\; \mu \;<\; 93.5 + 2.66\,(15.8/\sqrt{100}) =$$

$$89.30 \;<\; \mu \;<\; 97.70 \quad \text{with a 99\% confidence level}$$

4. Note that if the exact number of degrees of freedom is not found in a table of the *t*-distribution, it is best to use the next lowest figure for degrees of freedom. This results in a larger *t*-value and a more conservative (i.e., larger) confidence interval.

V. Point and Interval Estimation for Population Proportions

A. As in the case of other parameters, we need to derive both point and interval estimates of population proportions.

B. A proportion derived from a random sample is an unbiased point estimate of the respective population proportion. In other words, $E(p) = \pi$, where p represents the estimated value of the population proportion π.

C. Confidence Interval for π

1. An interval estimate of a population proportion is based on the normal approximation to the binomial distribution. Given that $n\pi \geq 5$, and $n(1 - \pi) \geq 5$, and the sample size is no more than 5% of the population size, then the sampling

distribution for the sample proportion p will be approximately normal with mean and standard deviation respectively equal to:

$$\mu_p = \frac{n\pi}{n} = \pi \quad \text{and} \quad \sigma_p = \frac{\sqrt{n\,\pi\,(1-\pi)}}{n} = \sqrt{\frac{\pi\,(1-\pi)}{n}} \qquad \textbf{(2.6)}$$

The standard deviation σ_p is the *standard error of the proportion*.

2. A z-score can be put together for π in the same fashion as was done for μ. If p is normally distributed with a mean π and standard deviation σ_p, then:

$$\frac{p-\pi}{\sigma_p} \sim N(0,1)$$

3. The confidence interval estimate for p will then be:

$$p - z_{\alpha/2}\sigma_p \;<\; \pi \;<\; p + z_{\alpha/2}\,\sigma_p \quad \text{with confidence level} \; 1-\alpha \qquad \textbf{(2.7)}$$

4. But again we do not know the value of σ_p. However, the sample proportions p and $1 - p$, substituted in the second part of equation 2.6, provide a satisfactory estimate of the standard error, which is then indicated as either s_p or $\hat{\sigma}_p$.

5. The *estimated margin of error* for a proportion is $E_{\alpha/2} = z_{\alpha/2}s_p$. This is the information that is generally given to describe the precision of survey results assuming that a confidence or significance level is also reported. Note that it is possible to calculate both the maximum margin of error for a sample proportion at a specified confidence limit, and the size of a random sample required to estimate a population proportion with a given margin of error and confidence level.

6. Once again, if $n > .05\,N$ and sampling is done without replacement, then s_p is multiplied by the square root of the quantity $\{(N - n) / (N - 1)\}$ as an adjustment, termed the *finite-population correction factor*.

VI. Determining the Sample Size

A. Given the above formulas for confidence intervals, it is possible to use the *error of estimate* to derive the sample size required for a given confidence level assuming that an estimate of σ is at hand.

HYPOTHESIS TESTING: INTRODUCTION

I. Statistical Hypotheses and Hypothesis Testing

A. A hypothesis is a statement about a future event, or an event the outcome of which is unknown at the time of the prediction, set forth in such a way that it can be rejected or not rejected. A *statistical hypothesis* is a claim about a population that can be tested with data obtained from a random sample.

1. The *null hypothesis* states that there is no statistically significant difference between certain population values. It is a statement of "no effect" or "no difference." The *alternative hypothesis* states that there is a statistically significant difference between these population values.

2. "If we find that the sample data are sufficiently unlikely to have occurred in the situation proposed by the null hypothesis, we reject the null hypothesis, which says that the population values are equal, and conclude instead that the population values are different from each other. This is the case in which the sample difference is said to be *statistically significant*. If we find that the observed sample data are reasonably likely to occur in the situation proposed by the null hypothesis, we do not reject the null hypothesis, and we attribute the sample difference to chance. This is the case in which the sample difference is said to be not statistically significant." (Olson)

3. Note that **statistical significance does not imply practical significance**. It does not tell us about the magnitude or importance of differences, but only whether or not certain sample differences can occur frequently by chance.

B. There is a risk of error in testing hypotheses when working with a sample. Probability theory allows us to set the constraints within which a hypothesis can be rejected in terms of the likelihood of error involved in a test.

C. The error of rejecting the null hypothesis when it is true is termed a *Type I error*. The error of failing to reject the null hypothesis when it is false (also called the fallacy of affirming the consequent) is termed a *Type II error*.

D. These errors result from our desire to generalize beyond the limits of our sample to the entire population. Probability theory and sampling distributions allow us to evaluate the likelihood of these errors.

E. Hypothesis testing proceeds by eliminating hypotheses:

1. In order to establish the validity of a theory, we generally look at the consequences which would follow if it were true. If these consequences do not hold, i.e., if they are false, we can reject the validity of the proposed theory.

2. But if the consequences are true, we can only say that *the theory may be true* or *that it cannot be rejected*. The only way to determine if a theory is absolutely true is to demonstrate that there is no other alternative theory, which is generally not

possible. A theory as to a causal sequence of events is tested by deriving hypotheses (consequences), which can then be tested.

II. Steps in Hypothesis Testing

A. *Specify the Model and Hypotheses* - in order to use a sampling distribution for statistical testing, assumptions have to be made about both the population and the procedures by which the sample is derived. These assumptions can be dichotomized by the degree of acceptance given to them by the researcher. Assumptions which are given a substantial degree of acceptance form what is termed *the model*, whereas the assumption or assumptions carrying the greatest degree of uncertainty comprise the hypothesis or hypotheses that are to be tested.

1. Since statistical tests do not allow selective rejection of assumptions, it is important that the hypothesis be the only assumption which is really in doubt -- this is one criterion for the selection of appropriate tests.

2. Statistical testing generally works in reverse: a hypothesis is set up which the researcher is interested in rejecting. This is the null hypothesis, indicated as H_0. It generally posits the lack of a relationship or no difference between groups. The alternative hypothesis is indicated by H_1, and generally posits the existence of the relationship or the nature of expected differences.

B. *Identify the Appropriate Sampling Distribution* - mathematical reasoning is used to associate a sampling distribution with the test statistic.

1. This distribution indicates how likely each of the possible outcomes are if the assumptions are correct. Then a decision rule is set up to indicate under which outcomes the hypothesis can be rejected and what risk is involved.

2. Using such a decision rule will result in making an erroneous judgment (a Type I or Type II error) occasionally, but in the long run we can expect to make the correct decisions as to rejection most of the time.

C. *Determine a Significance Level and Critical Region*

1. The limits defining which alternatives require rejection of the assumptions are derived from the sampling distribution and comprise the *critical region*. Possible outcomes from the experiment are divided into two categories: those falling within the critical region which will permit rejection and those that do not.

2. A critical region is defined when the researcher

   a. determines the acceptable risk for making Type I and Type II errors, and

   b. selects either a one-tailed or two-tailed test (i.e., whether or not the critical region will include both the left and right tails of the sampling distribution or just one of these tails).

3. The probability of making a Type I error (rejecting a null hypothesis when it is true) is the *significance level* of a test.

4. For any given test, there is an inverse relationship between the probability of making a Type I error and that of making a Type II error. As the significance level (probability of a Type I error) of a hypothesis test decreases, the probability of a Type II error increases. The probability of a Type II error is generally represented as β, with the *power* of a test being defined as $(1 - \beta)$. Thus, power represents the ability of a hypothesis testing procedure to reject a false hypothesis, which we would like to maximize. But because of the inverse relationship between Type I and Type II errors, this is somewhat difficult. The probability of both types of errors can be reduced by increasing the sample size or by reducing the variability of the observations (generally by reducing measurement error or using procedures such as paired samples).

5. A significance level should be chosen to make it as difficult as possible to obtain the results which are expected. In social and natural science research, common significance levels are the .05 or .01 level, and sometimes even the .001 level.

6. If we can predict the direction of a relationship, one-tailed tests are preferable to two-tailed tests at the same significance level. It is often possible to predict direction on the basis of previous studies or theory.

D. *Compute the Test Statistic* - Compute a test statistic which has the specified sampling distribution. This is the value that will be compared to the critical region of the sampling distribution.

E. *Execute the Test* - Decide to reject or not reject the null hypothesis based on the specified significance level. If the test statistic falls within the critical region, the null hypothesis is rejected with a known probability of a Type I error. If it does not fall within the critical region, the null hypothesis is not rejected and the possibility of a Type II error is accepted.

F. *Compute the P-Value* (optional) - The *p*-value is the smallest probability for which the null hypothesis can be rejected and is generally provided by statistical software. While computer software is generally necessary to get an exact *p*-value, this can be approximated using the statistical distribution tables and finding the significance level that comes closest to the computed value of the test statistic.

HYPOTHESIS TESTING: POPULATION MEANS AND T-STATISTICS

I. Tests about Means

A. There are three types of tests concerning population means:

- single-sample test for a hypothesized value of the mean,
- independent-samples test of the hypothesized difference between two means,
- paired-samples test of the hypothesized difference between two means.

B. The test statistic and distribution used in each of these three cases is either the standard normal deviate (*z*-score) and the normal distribution for cases when σ is known, or the *t*-statistic and Student's *t* distribution for cases when σ is unknown.

C. The *critical region* is derived from the distribution in the same manner that confidence limits are derived. The size of the critical region reflects the significance level α (the acceptable probability of making a Type I error).

a. A value for the test statistic is chosen so as to establish a critical region that contains α percent of the area of the distribution, with $(1 - \alpha)$ percent outside the critical region. The value of the test statistic that establishes the critical region is termed the *critical value.*

b. As in confidence interval estimates, a two-sided test requires the critical region to be equally distributed in both tails. Thus, the critical value for the test statistic is chosen so that $\alpha/2$ percent of the distribution is contained in each tail.

c. A one-sided test implies that we are interested in either H_1 greater than the value specified in the null hypothesis or H_1 less than the value specified in the null hypothesis. In this case, the critical region is distributed entirely in one tail. Thus, the critical value for the test statistic is chosen so that all α percent of the distribution is contained in the appropriate tail.

D. *If the observed or computed value of the test statistic is greater than the critical value for a right-tailed test, or less than the critical value for a left-tailed test (i.e., the test statistic lies in the critical region), the null hypothesis is rejected with a known probability of making a Type I error*.

II. Single-Sample Test for a Specific Value of the Mean

A. The null hypothesis H_0 for this type of test specifies a specific value for the population mean. The alternative hypothesis H_1 specifies that the population mean is not equal to this value. For example:

1. $H_0: \mu = 10.0$ and $H_1: \mu \neq 10$ for a two-sided or two-tailed test
2. $H_0: \mu \leq 35.7$ and $H_1: \mu > 35.7$ for a one-sided or one-tailed test

B. Single-Sample Test Where σ is Known

1. The test statistic is the standard normal deviate or *z*-score. The rationale behind its use is identical to that for establishing confidence intervals for the population mean when σ is known. As in the case of confidence intervals, the normal distribution is used to assign probabilities to the occurrence of various values of the sample mean. In this case, we determine the probability of the sample mean occurring by chance if the population mean actually had the hypothesized value. The statistic and its distribution are:

$$\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

2. Example:

   Suppose that we want to check on the adequacy of sampling procedures used in a local survey, and we suspect that middle and upper income families were not sampled correctly (i.e., had a larger or smaller probability of appearing in the sample than lower income families). Recent census data, which involves the entire population, shows that μ = \$7,500 and σ = \$1,500 for middle and upper income families. Assuming that our sample of 100 middle and upper income families had $\bar{x}$ = \$7,900, can we conclude that the sample was indeed random?

   a. Specify the Model and Hypotheses

      i. Normally distributed sampling distribution for the sample mean

      ii. Model: μ = \$7,500, σ = 1,500.

      iii. Random sampling implies that $\bar{x} = \mu$ = \$7,500, thus

      H_0: μ = \$7,500 H_1: $\mu \neq$ \$7,500

   b. Identify the Appropriate Sampling Distribution

      Since n = 100 is greater than 30, the CLT indicates that the sampling distribution for $\bar{x}$ will be normally distributed. Since we know σ, the normal distribution is the appropriate sampling distribution for this situation.

   c. Determine a Significance Level and Critical Region

      i. Generally decide between .05 or .01 level of significance - in this case, select .05 as the significance level α.

      ii. Decide on one-tailed or two-tailed test - since we are interested in determining if the sample is biased in either direction, a two-sided test is appropriate.

      iii. Derive the critical region for the distribution - for the normal distribution, a two-sided test at α = .05 implies that the critical value for *z* must be selected so that 2.5% (.025) of the area of the normal distribution lies in each tail.

The value for z which has $\alpha/2 = .025$ is 1.96. Thus, $z_{.05} = \pm 1.96$ defines the critical region.

d. Compute the Test Statistic

i. Compute the observed value for the statistic:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

ii. For the example, this is:

$$z = \frac{7900 - 7500}{1500 / \sqrt{100}} = 400 / 150 = 2.667$$

e. Execute the Test

i. This decision is based on whether the calculated or observed value of the test statistic z lies in the critical (rejection) region. Reject H_0 if z lies in this region, but also report the exact significance level (p-level) at which the rejection occurs.

ii. For the example, since $z = 2.667 > 1.96$, we can reject H_0: $\mu = 7500$ at the .05 level of significance. Since we know that μ is in fact equal to 7500, we would reject the implied hypothesis of the sample being random.

f. Calculate the *P*-Value

i. The exact p-level is .0076.

3. The relationship between confidence intervals and hypothesis testing is clear from this result, in that the confidence interval for the estimate of μ from the sample is:

$$7900 - (-1.96)\,(150) \; < \; \mu \; < \; 7900 + (1.96)\,(150) \; =$$

$$7606 \; < \; \mu \; < \; 8194 \qquad \text{with a .95 confidence level}$$

As can be seen here, if the null hypothesis is rejected, the hypothesized value of μ will *not* lie within the calculated confidence interval.

C. Single Sample t Test Where σ is Unknown

1. As with the derivation of confidence intervals, the more general case for hypothesis testing is when σ is unknown. This situation is identical to that for deriving confidence intervals for μ. Thus, the appropriate distribution is Student's t distribution and the test statistic is the t-statistic, as in:

$$\frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t\,(n-1\ df)$$

2. Example:

   Suppose a random sample of size 50 has a sample mean of 10.5 and a sample standard deviation of 2.2. Test the hypothesis that μ is 10.0 using a two-tailed test at $\alpha = .01$.

   a. Specify the Model and Hypotheses

      i. Normally distributed sampling distribution for the sample mean
      Random sampling
      Normal population

      ii. H_0: $\mu = 10.0$ H_1: $\mu \neq 10.0$

   b. Identify the Appropriate Sampling Distribution

      Since $n = 50$ is greater than 30, the CLT indicates that the sampling distribution for $\bar{x}$ will be normally distributed. Since we do not know σ, Student's t distribution is appropriate for this situation.

   c. Determine a Significance Level and Critical Region

      i. .01 level of significance

      ii. Decide on one-tailed or two-tailed test: since we are not given any a priori information about the potential direction of H_1, we need to use a two-sided test.

      iii. Derive the critical region for the distribution: for the t distribution, a two-sided test at $\alpha = .01$ implies that the critical value for t must be selected so that 0.5% (.005) of the area of the t distribution lies in each tail, which means we would use the column labeled $t_{.005}$ in the table of the t-distribution along with df = 49. Since there is no row in the table with precisely this value, use the row showing 40 df. Thus, the critical value of t is 2.704. The critical region is defined as $t_{.005} = \pm 2.704$.

   d. Compute the Test Statistic

      i. Compute the observed value for the statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

      ii. For the example, this is:

$$t = \frac{10.5 - 10.0}{2.2/\sqrt{50}} = 0.5/0.311 = 1.607$$

   e. Execute the Test

For the example, since $t = 1.607 < 2.704$, we cannot reject H_0: $\mu = 10.0$ at the .01 level of significance.

3. The confidence interval for μ would be:

$$10.5 - (2.704)\,(0.311) < \mu < 10.5 + (2.704)\,(0.311) =$$

$$9.66 < \mu < 11.34 \text{ with a .99 confidence level}$$

The confidence interval including the hypothesized value of μ is consistent with failing to reject the null hypothesis.

III. Single-Sample Hypothesis Tests Involving Proportions

A. Hypothesis tests about binomial population proportions are, like confidence intervals, based on the normal approximation to the binomial distribution. Once again, given that $n\,\pi \geq 5$, and $n(1-\pi) \geq 5$, and the sample size is no more than 5% of the population size, then the sampling distribution for the sample proportion p will be approximately normal with mean and standard deviation respectively equal to:

$$\mu_p = n\pi\,/n = \pi \quad \text{and} \quad \sigma_p = \frac{\sqrt{n\pi\,(1-\pi)}}{n} = \sqrt{\frac{\pi\,(1-\pi)}{n}} \qquad \textbf{(2.8)}$$

B. If p is normally distributed with a mean π and standard error σ_p, then:

$$\frac{p-\pi}{\sigma_p} \sim N(0,1)$$

Even though we do not know the value of $\sigma_{\hat{p}}$, the hypothesized values of the population proportions π and $1 - \pi$, substituted in the second part of equation 2.8, provide a satisfactory estimate of the standard error, which is then indicated as either s_p or $\hat{\sigma}_p$. *Z*-scores can then be calculated and hypotheses tested in the usual manner.

HYPOTHESIS TESTING: DIFFERENCES BETWEEN PARAMETERS

I. Introduction

A. We are often interested in determining whether differences exist between two populations. If all elements of the populations were available, such determinations would be made by comparing their parameters. Since most populations are not fully available for study, we use the techniques of statistical inference to compare statistics obtained from random samples taken from the populations. These comparisons are generally based on sample means, variances, and proportions.

B. The type of inferential statistics utilized for the comparisons depends on whether the two samples are independent or dependent (i.e., related to each other). As Weimer states, "If the selection of sample data from one population is unrelated to the selection of sample data from the other population, the samples are called *independent samples*. If the samples are chosen in such a way that each measurement in one sample can be naturally paired with a measurement in the other sample, the samples are called *dependent samples*. Each piece of data results from some source. A source is anything, a person or an object, that produces a piece of data. If two measurements result from the same source, then the measurements can be thought of as being paired. As a result, two samples resulting from the same set of sources are dependent."

C. Independent samples may be obtained either by taking separate random samples from the two populations, or by dichotomizing a single random sample taken from a combined population based on the characteristic or variable that is hypothesized to differentiate the two populations.

II. Testing the Difference between Two Means from Independent Samples

A. An *extension of the CLT* provides the means to begin developing the test procedures for independent samples.

1. If independent random samples of sizes n_1 and n_2 are drawn from populations that are distributed normally, and n_1 and n_2 are each > 30, then the sampling distribution of the difference between the two sample means $\bar{x}_1$ and $\bar{x}_2$ will be:

$$(\bar{x}_1 - \bar{x}_2) \sim N \left(\mu_1 - \mu_2 , \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)$$

2. As n_1 and n_2 become large enough (each > 30), then the CLT implies that the sampling distribution of the difference between these two means will be Gaussian, no matter what distribution is found in the population.

3. The obvious test statistic is the standard normal deviate based on the difference between the sample means, with H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$. Thus, the statistic would be:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \qquad \text{since } H_0 : \mu_1 = \mu_2 \tag{2.9}$$

4. The variances of the two samples are additive in that there is a separate degree of variability in each sample which must be taken into account in determining the overall test statistic.

B. Independent-Samples *t* test for Equal Variances

1. While equation 2.9 provides a test statistic for situations in which σ_1^2 and σ_2^2 are known, this situation is virtually non-existent. It would seem reasonable to assume that the *t* test could be generalized to cover this situation, as in the case of single sample tests, by using the sample variances to estimate σ_1^2 and σ_2^2. However, using both s_1^2 and s_2^2 in the denominator of 2.9 does *not* generate a statistic with a *t* distribution. The formula for a *t* statistic requires that there is only one source of variation in the denominator, i.e., a single estimate for the standard error. This calls from some additional assumptions before we can utilize a *t* test for this problem.

2. One simplifying assumption that is often made is that the populations from which the two independent samples are taken have the same variance. This implies that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. But there are still two estimates of this variance, s_1^2 and s_2^2, rather than just one estimate. To obtain the single estimate, we pool the two estimates of the variance to obtain a test statistic that has a *t* distribution:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_{pooled}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t \ (n_1 + n_2 - 2 \ df) \tag{2.10}$$

3. In order to *pool* the variance estimates from the separate samples, we take a weighted average of these estimates to obtain the common estimated variance s^2. This implies that the first sample variance needs to be weighted (multiplied) by the degrees of freedom associated with that sample variance, and then added to the second sample variance which has been weighted by its degrees of freedom. This sum is then divided by the combined degrees of freedom:

$$s_{pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{2.11}$$

Substituting the formula for the pooled variance from 2.11 for s^2 in 2.10 results in the test statistic:

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_{pooled}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t \ (n_1 + n_2 - 2 \ df) \tag{2.12}$$

4. Note that this *t* statistic is robust against small or moderate violations of the assumption of equal variance (termed homogeneity of variance) between the two populations, especially if the samples are of approximately equal size. In fact, for equal sized samples, the variance of one sample can be up to three times the variance of the other without severely interfering with the accuracy of the test.

5. Normality assumptions for the populations are necessary to use the *t* test when n_1 or n_2 are less than 30. As in other applications of the *t* distribution, the *t* statistic is fairly robust to these assumptions.

C. Example: Independent Samples *t* test

Suppose we are interested in determining if there is a bias either towards or against urban counties in the distribution of federal grants-in-aid. We take two independent samples -- one of urban counties and one of rural counties -- and compute the level of federal grants-in-aid per capita. For the urban counties, $n_1 = 25$, $\bar{x}_1 = \$160$, and $s_1 = \$12$. For the rural counties, $n_2 = 25$, $\bar{x}_2 = \$130$, and $s_2 = \$8$. To determine if there is a statistically significant difference between these two sample means, we carry out the following steps:

1. Model: Independent Random Samples
   Normally Distributed Populations
   $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but are unobservable
   H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$

2. Since σ is unknown, must use the *t* distribution.

3. $\alpha = .01$, use a two-sided test

4. Compute the test statistic:

$$t = \frac{(160-130)}{\sqrt{\left(\frac{(25-1)12^2 + (25-1)8^2}{25+25-2}\right)\left(\frac{1}{25}+\frac{1}{25}\right)}} = \frac{30}{2.884} = 10.40$$

5. With 48 df, the table of the *t* distribution (using 40 df) indicates $t_{.005}(40) = 2.704$ Thus, H_0 of no difference between the mean per capita federal grants-in-aid between urban and rural counties can be rejected at the .01 level of significance. Thus, we can conclude that the sample data is consistent with a bias in favor of urban counties. A confidence interval for the difference between the means can be derived in the usual manner as:

$$30 - (2.704)\,(2.884) < \mu_1 - \mu_2 < 30 + (2.704)\,(2.884) =$$

$$22.20 < \mu_1 - \mu_2 < 37.80 \quad \text{with a 99\% confidence level}$$

Once again, note the relationship between hypothesis testing and confidence intervals. When the null hypothesis is rejected, the hypothesized value of the

parameter (in this case the difference between μ_1 and μ_2) will not be encompassed by the interval.

D. Independent-Samples *t* test for Unequal Variances

1. When the size of the two samples is very different, and when the variances are not equal, the following *t´* statistic will have an approximate *t* distribution with ($n_1 + n_2$) df:

$$t' = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t\,(n_1 + n_2 - 2\ df) \qquad \textbf{(2.13)}$$

2. It is common practice to use this *t´* statistic (often called the *separate-variance t test*) to test hypotheses concerning population differences when the assumption of constant variance is unwarranted. The true *p*-value or significance level will always be equal to or less than that associated with 2.13. Thus, the error associated with the *t´* statistic will always be on the side of a more conservative test. As long as the sample sizes are similar, or are large, this error will also be quite small.

III. Testing the Difference between Two Means from Dependent (Paired) Samples

A. *Before and after Testing* or *Paired-Samples Tests* represent a situation in which the samples are dependent, having been intentionally matched to each other. This situation can be thought of as a mechanism for improving the standard error for the estimate of the hypothesized parameters, thus reducing the probability of a Type I error. The improvement in the standard error arises because the pairing or matching of the observations essentially excludes many extraneous variables which otherwise might influence the observation values in random ways.

B. Since the samples are not independent, we cannot use the independent-samples tests. Instead, the observations are thought of as coming from a single sample where the observation value is defined as the difference, in the value of the variable, between each member of a matched pair. This difference is termed the *difference score* and is computed as:

$$d_i = x_{1i} - x_{2i} \quad \text{for } i = 1, 2, ..., n$$

C. If the population is normally distributed, we can derive a statistic with a *t* distribution (assuming that sigma is unknown) that is termed the *paired-samples t test*:

$$\frac{\bar{d}}{s_d/\sqrt{n}} \sim t\ (n-1\ df) \qquad \textbf{(2.14)}$$

The null hypothesis is, of course, that the sample mean for the difference between the pairs is 0, i.e., H_0: $\mu_d = 0$

D. An interval estimate can be derived in the usual manner.

E. Example: Paired Samples *t* test

A random sample of 10 departments in a large city is selected to determine the impact of direct participation of employees in the city's strategic planning process. Two staff members are identified for each department that are matched on every relevant characteristic, except that one of the staff members was involved as an active participant in the organization's strategic planning process, whereas the other was not. The number of times that strategic objectives appear in each employee's writing over the course of a year is measured, with the following results:

Department	Participant Employee	Non-Participant Employee	d_i
1	263	87	176
2	317	208	109
3	185	156	29
4	162	97	65
5	411	58	353
6	132	108	24
7	39	13	26
8	163	151	12
9	108	96	12
10	16	11	5

Given this sample data, can we conclude that direct participation in the strategic planning process resulted in greater use of strategic planning outcomes by staff within city departments?

1. Model: Random Sample
   Normally Distributed Population
   H_0: $\mu_d \leq 0$, H_1: $\mu_d > 0$

2. Since σ^2 is unknown, must use t distribution

3. $\alpha = .05$, use a one-sided test since we only care if participation has a positive impact

4. Compute the test statistic:

$$\bar{d}_i = 81.1, \quad s_d = 109.71, \quad n = 10$$

$$\frac{81.1}{109.71/\sqrt{10}} = \frac{81.1}{34.69} = 2.338 \sim t(9\,df)$$

5. With 9 df (one-tailed), the table of the t distribution indicates $t_{.05}(9) = 1.833$. Thus, H_0 of no difference between the employees can be rejected at the .05 level of significance. We would conclude that active participation in the strategic planning process appears to result in greater use of strategic planning outcomes.

ESTIMATION AND HYPOTHESIS TESTING:
STANDARD DEVIATIONS/VARIANCES

I. Point Estimates of the Population Variance and Standard Deviation

A. As in the case of estimating the population mean with the sample mean, an estimate of the population variance can be obtained via the sample variance. The expected value of the sample variance (i.e., the mean of its sampling distribution) is in fact the population variance, thus indicating that the sample variance is an unbiased estimator of the population variance. In other words:

$$E(s^2) = \sigma^2$$

B. Although it would seem that we could extend this idea and use the sample standard deviation as an unbiased estimate of the population standard deviation, this is not the case. The sample standard deviation can be shown to yield a biased estimate of the population standard deviation (i.e., $E(s) \neq \sigma$). An intuitive illustration of this can be seen by calculating the mean of a set of s^2-values and the mean of the corresponding set of s-values. Taking the positive square root of the former will not yield the latter. However, this bias is not very large, and the general practice is to use s as an estimate of σ.

II. Interval Estimates of the Population Variance and Standard Deviation

A. The statistic that is used to associate probabilities with the sampling distribution of the sample variance is constructed by comparing the sample variance (weighted by the respective degrees of freedom) to the hypothesized value of the population variance. This statistic has a *Chi-Square Distribution* (χ^2) with $(n - 1)$ degrees of freedom. It is:

$$\frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2\ (n-1\ df) \qquad \textbf{(2.15)}$$

B. The χ^2 distribution appears to be quite different from that of the normal and t distributions. Its values are all positive and it is not symmetric until the number of degrees of freedom is greater than 30.

1. Like the t distribution, the shape of the χ^2 distribution depends on the number of degrees of freedom and assumes a normal population (it is very sensitive to this). Appendix I of Lind, Marchal, and Wathen provides the χ^2 distribution for df $\leq$ 30.

2. The χ^2 distribution does become Gaussian in shape when df > 30. If an adjustment is made for the fact that the χ^2 distribution lies entirely to the right of the origin, when df > 30 the normal distribution can be used. The χ^2 values are calculated via the formula:

$$\tfrac{1}{2}\left(\pm z_{\alpha/2} + \sqrt{2(df)-1}\right)^2 \qquad \text{where } \alpha \text{ is the significance level}$$

3. The χ^2 distribution is directly related to the normal distribution, in that it can be generated by squaring the values of normally distributed z-scores.

4. Since the χ^2 distribution is not symmetric around the origin, a two-sided H_1 requires the critical values $\chi^2_{1-\alpha/2}(n-1 \text{ df})$ and $\chi^2_{\alpha/2}(n-1 \text{ df})$, a one-sided "less than" H_1 requires the critical value $\chi^2_{1-\alpha}(n-1 \text{ df})$, and a one-sided "greater than" H_1 requires the critical value $\chi^2_{\alpha}(n-1 \text{ df})$.

C. The process of deriving confidence interval estimates for σ^2 is analogous to that for μ, with the χ^2 statistic being used instead of *z* or *t* statistics.

1. Assuming a random sample from a normal population, the *confidence interval for the population variance* is defined as:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \quad \text{with confidence level } 1-\alpha \qquad \textbf{(2.16)}$$

2. And the *confidence interval for the population standard deviation* is defined as:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}} \quad \text{with confidence level } 1-\alpha \qquad \textbf{(2.17)}$$

3. There is one major difference between these confidence intervals and those used for the population mean. In this case, the term representing the margin of error, $(n-1)/\chi^2$, multiplies the point estimate of the parameter. In the confidence interval for the population mean, the term representing the margin of error is added to and subtracted from the point estimate of the parameter. The difference arises from the χ^2 statistic carrying out the comparison between the point estimate and the value of the parameter via a ratio instead of the difference ($\overline{x} - \mu$) that is used in *z* or *t* statistics.

4. The ability to make inferences about variances and standard deviations using the χ^2 statistic is highly dependent on the assumption of a normal population. Unlike the *t* statistic, <u>slight variations from this normality assumption can have substantial effects on the probability levels thought to be associated with a sampling distribution</u>.

D. Example: Provided in Class

HYPOTHESIS TESTING: CONTINGENCY TABLES AND CHI SQUARE TESTS

I. Chi-Square Tests Involving Proportions

A. Although *z*-scores can be used for comparing binomial proportions from two populations, they are not generalizable to a situation in which there are more than two proportions (i.e., multinomial proportions). A test statistic that is more general is Pearson's X^2 statistic, which approximates the χ^2 distribution assuming a large enough sample. Since a χ^2 test with one degree of freedom is identical to a *z*-score, the χ^2 test can be used for single or multiple sample hypothesis tests, thus obviating the need for *z*-scores in most cases involving proportions.

B. *Pearson's X^2 Statistic* is defined as:

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \qquad \textbf{(2.18)}$$

where o_i = the observed frequency (the number of occurrences of a categorical variable in the sample) for category *k*, and
e_i = the expected frequency (the number of occurrences that would be expected for the categorical variable in this category by chance) for category *k*.

(Lind, Marchal, and Wathen use f_o and f_e to represent o_i and e_i, respectively.)

1. "If the expected frequencies [e_i] correctly reflect the population proportions [π_i], then ... the sampling distribution of Pearson's X^2 statistic is approximated by the χ^2 distribution with degrees of freedom equal to the number of categories *k* minus the number of independent sample-based restrictions used in the calculation of the expected frequencies." (Olson)

2. The only such restriction that we will deal with in the case of proportions is that the expected frequencies e_i must always add up to the number of observations. The number of degrees of freedom is computed as $(k - 1)$. This is the number of categories (not the number of observations) minus one.

3. The χ^2 distribution can be used to assign probabilities to calculated values of Pearson's X^2 statistic with a large sample size. There are two rules of thumb for making sure the sample size is large enough for the approximation to the χ^2 distribution to hold:

   a. The approximation is sufficiently accurate with 1 degree of freedom (i.e., two categories), if none of the expected frequencies e_i is less than 10.

   b. The approximation is sufficiently accurate with more than one degree of freedom if 80% or more of the expected frequencies e_i are at least 5 and none is less than 1.

C. Applying Pearson's X^2 Statistic: The Goodness-of-Fit Test

1. Pearson's X^2 statistic allows a comparison of sample proportions to hypothesized values of the proportions in the population. The hypothesized values are derived from the null hypothesis, which for *k* categories and specific hypothesized values of π indicated as π_{0i}, is written:

$$H_0 : \pi_i = \pi_{0i} \text{ and } H_1 : \pi \neq \pi_{0i} \text{ for } i = 1, 2, \ldots, k \text{ where } \sum_{i=1}^{k} \pi_{0i} = 1$$

2. To carry out this test, the observed frequencies for the occurrence of the variable are recorded for each of the *k* categories. The expected frequency for each category e_i is then derived based on the assumption that H_0 is true. This quantity is calculated by multiplying the sample size *n* by the hypothesized proportion π_{0i} for each category *i*. In symbols, this is: $e_i = n\pi_{0i}$ for $i = 1, 2, \ldots, k$

3. If the null hypothesis is false, the squared difference in the numerator of Pearson's X^2 will be large (i.e., the observed proportions will differ from the expected proportions), and the statistic will generally have a value greater than one.

    a. Even if the differences between o_i and e_i are negative, they will still contribute to a positive numerator. This means that, if the null hypothesis is false, Pearson's X^2 will always produce a value that falls in the upper tail of the χ^2 distribution.

    b. Thus, the critical value for a test with a two-sided H_1 at the α level of significance will be χ^2_α with $k - 1$ df. If the computed value of Pearson's X^2 is greater than this critical value, H_0 can be rejected at the stated level of significance.

D. Example: X^2 Goodness-of-Fit Test

Local governments can be classified as having several different structures. In a random sample of 160 local governments, the following proportions occurred:

Structure	Percentage of Govt's	Number of Govt's
1. strong mayor-weak council	17.5	28
2. weak mayor-strong council	31.9	51
3. strong mayor-strong-council	21.3	34
4. city manager	29.4	47
Totals	100.1	160

Are these results likely to have occurred by chance, or is there a preference among local governments for some structures over others?

1. Assumptions: Random Sample; none of the expected frequencies will be less than 5.

Hypotheses: The probability of occurrence for any of these categories by chance would be 0.25. Thus,

H_0: $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$ H_1: $\pi_i \neq 0.25$ for some i

2. Significance Level: use $\alpha = .05$, two-sided test, and the χ^2 distribution

3. Obtain expected frequencies:

Category	o_i	e_i	$o_i - e_i$	$(o_i - e_i)^2/e_i$
1	28	(160)(.25) = 40	–12	3.60
2	51	40	11	3.02
3	34	40	–6	0.90
4	47	40	7	1.23

4. Compute Pearson's X^2 as:

$$X^2 = 3.60 + 3.02 + 0.90 + 1.23 = 8.75$$

5. Decision: The critical value for $\chi^2_{.05}$ (3 df) = 7.815. Since 8.75 > 7.815, we can reject H_0 at the .05 level of significance. Our conclusion is that these results would not generally have occurred by chance, and that there appears to be some preference among cities for either the weak-mayor/strong council or the city manager structures.

II. Contingency Tables and Pearson's X^2

A. One extremely important application of Pearson's X^2 is a general procedure that provides a test for significant differences among categories of a single variable in two populations (termed a *chi-square test of homogeneity of populations*), or for cross-classifications of two variables from the same population (termed a *chi-square test of independence*).

B. Both of these tests utilize a tabular cross-classification of the sample data that is termed a *contingency table*. Such a table displays the observed frequencies (i.e., count data) for the various cross-combinations of the categories of the two variables (or populations) in its cells.

1. The last column on the right side of the table, and the bottom row of the table, contain the frequency distributions for each of the two variables.

2. The values of the cells in this last column are also the sum of the respective rows, whereas the values of the cells in the bottom row are also the sum of the respective columns. These sums are also termed *marginal frequencies*.

C. Example: 3 x 3 contingency table cross-classifying income group with degree of participation in community activities.

Income Group - Observed Frequencies

Participation	Lower	Middle	Upper	**Total**
Low	52	30	10	**92**
Med	16	28	22	**66**
High	18	35	37	**90**
Total	**86**	**93**	**69**	**248**

III. Testing for Significant Differences between Proportions

A. A contingency table is intended to address the question of whether the differences between the proportions of each column are statistically significant. Here the question is can we conclude, from this sample, that upper income groups exhibit a higher degree of community participation than do lower income groups?

B. In order to test this, pose a null hypothesis indicating there is no difference in community participation among the income groups, implying that the proportions of high, medium, and low participants are the same for each income group.

C. Under the assumptions of random sampling and a correct null hypothesis, we can derive the expected frequencies for each cell of the table based upon the totals. This is accomplished by applying the same proportions found in each row total (for example, 92/248) to the column totals (e.g., 86) to determine the proportion that would be expected to appear in each cell if H_0 was correct. In general terms, the *expected frequency* is defined as the product of these relevant marginal totals divided by the total sample size n. The formula is:

$$e_{a\,b} = \frac{m_a\, n_b}{n} \qquad \textbf{(2.19)}$$

where e_{ab} = the expected frequency for the cell in column a, row b,
m_a = the row total for row a, and
n_b = the column total for column b.

1. For the above example, the expected frequencies would be:

Income Group - Expected Frequencies

Participation	Lower	Middle	Upper	**Total**
Low	32	34	26	**92**
Med	23	25	18	**66**
High	31	34	25	**90**
Total	**86**	**93**	**69**	**248**

2. The expected frequencies can then be compared to the observed frequencies from the sample(s). If the differences between the respective cells are large, the null hypothesis can be rejected. The statistic that is used to indicate when these differences are large enough to conclude that they are unlikely to have occurred by chance is Pearson's X^2. The formula is the same as equation 2.18:

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \qquad \textbf{(2.20)}$$

where k = (k_A k_B), which is simply the number of rows times the number of columns, and is equal to the total number of cells in the body of the table (excluding the "Total" column and row).

3. The number of degrees of freedom for the statistic is the product of ($k_A - 1$) ($k_B - 1$), since there are a number of "sample-based restrictions" imposed on these calculations. The sum of each row of expected frequencies must equal the row total, which imposes k_A restrictions. The sum of each column of expected frequencies must also equal the column total, which (due to the k_A restrictions) only imposes an additional $k_B - 1$ restrictions. Thus, there are ($k_A + k_B - 1$) restrictions. The number of degrees of freedom is:

$$\text{df} = k_A\, k_B - (k_A + k_B - 1) = k_A\, k_B - k_A - k_B + 1 = (k_A - 1)\,(k_B - 1)$$

4. The null hypothesis is tested by the usual method of employing the sampling distribution of the statistic, the χ^2 distribution if the assumptions regarding sample size are met. If the value of the test statistic is greater than that expected by chance, the null hypothesis can be rejected at the appropriate level of significance.

D. Example: The X^2 Test for Independence

1. Assumptions: Independent Random Samples
   H_0: No differences among income groups with respect to community participation.
   H_1: Differences do exist among income group with respect to participation.

2. Significance Level: .01 level of significance.
   Use upper-tail of χ^2 distribution for one-sided test.

3. Compute the test statistic

   a. Obtain expected frequencies:

52	30	10	**92**
16	28	22	**66**
18	35	37	**90**
86	**93**	**69**	**248**

   b. Compute Pearson's X^2:

Cell	o_i	e_i	$o_i - e_i$	$(o_i - e_i)^2$	$(o_i - e_i)^2 / e_i$
a	52	31.9	20.1	404.01	12.665
b	30	34.5	– 4.5	20.25	0.587
c	10	25.6	–15.6	243.36	9.506
d	16	22.9	– 6.9	47.61	2.079
e	28	24.7	3.3	10.89	0.441
f	22	18.4	3.6	12.96	0.704
g	18	31.2	–13.2	174.24	5.585
h	35	33.7	1.3	1.69	0.050
i	37	25.0	12.0	144.00	5.760
Total	248	247.9			37.377

$X^2 = 37.377$; df = (3 – 1) (3 – 1) = 4; critical value $\chi^2_{.01}$ (4) = 13.277

4. Decision: Since $X^2 = 37.377 >> \chi^2$ .01, 4 = 13.277, reject H_0 at $\alpha = .01$

PART THREE
IDENTIFYING RELATIONSHIPS: ANALYSIS OF VARIANCE, CORRELATION, AND REGRESSION ANALYSIS

ANALYSIS OF VARIANCE

I. Introduction to Analysis of Variance

A. Analysis of Variance (ANOVA) is a general procedure for testing for differences among two or more population means. In this sense, it is an extension of the independent-samples *t* test. Its main use is to test for a relationship between a categorical or ordinal scale variable and an interval/ratio scale variable.

1. The basic procedure is to first divide the observations in a sample into several subsamples, based on the value of the categorical/ordinal scale variable. Each observation in a specific subsample will have the same value for this categorical/ordinal scale variable, and will also have a value for the interval/ratio scale variable.

2. Although the null hypothesis states that there is no difference between population means (as represented by the categories of the categorical variable and the respective subsamples), the test is carried out via comparison of variances instead of direct comparisons between the subsample means (as is done in the independent-samples *t* test).

   a. If the null hypothesis is true, the population means for each category (as represented by each subsample) will be equal to the overall mean. Thus, the subsample means should not differ markedly from the overall sample mean, any small differences being attributable to random effects. Thus, the basic concept behind the test is a comparison of these differences.

   b. Each of these differences is in fact a deviation from the overall mean, as defined previously. To compare the differences between subsample means, we first have to sum the deviations in such a way that positive and negative differences do not cancel each other out. As we have already seen, the best method for doing this is to sum the squared deviations from the respective means. If we divide each of these squared deviations by their respective number of degrees of freedom, they define an estimate of the population variance. This estimate is often termed the *between groups variance*, since it represents an estimate of the population variance (i.e., the deviations) between the categories. (Lind, Marchal, and Wathen refer to between groups variation as the *treatment variation*, in which the term "treatment" refers to the different populations that are being compared (Lind et al., 2002).)

   c. If the null hypothesis is true, and the population means are in fact equal, then the average variation within any subsample or category (i.e., the deviations of each observation within the subsample from the subsample mean) should not be markedly different from the variation between the categories or subsamples.

3. Thus, two estimates of the variance are computed, one based on all the observations combined and one based on a combination of the variances of each subsample. These two separate estimates of the population variance are compared via an *F* statistic. The variance estimate based on the combined observations is placed in the numerator and the variance estimate based on the combined subsample variances is placed in the denominator.

    a. If the null hypothesis is true, these two estimates of σ^2 will be similar and their ratio will be close to one, implying that the null hypothesis of no difference between the means cannot be rejected.

    b. If the null hypothesis is false, these estimates will differ markedly, and the ratio will be far in excess of one, allowing for rejection of the null hypothesis and the conclusion that the variation between the groups is much greater than the variation within the groups.

II. Decomposing Total Variation

A. *Variation* is formally defined as the sum of the squared deviations of observation values from the mean, or $\sum (x_i - \bar{x})^2$. With respect to ANOVA, this quantity represents total variation about the *grand mean*, which is the mean for the entire sample, and is also termed the *total sum of squares* (TSS) and is referred to as *SS total* in Lind, Marchal, and Wathen (Lind et al., 2012).

B. Notation

- variable *A* or Factor *A* = the categorical variable (this is also termed the independent or treatment variable)
- x = the dependent variable
- k represents the number of categories or subsamples for variable *A*
- n_a is the number of observations in the *a*th subsample or category of variable *A*
- $n = \Sigma\, n_a$ and is the total number of observations in the sample
- x_{ai} represents the *i*th observation in subsample or category *a*
- $T_{a.}$ represents the sum of the observations in the *a*th subsample or category, ($\Sigma\, x_{ai}$), the dot (.) indicating this is summed over all values of the subscript *i*
- $\bar{x}_{a.} = T_{a.} / n_a$ which is the mean of the observations in subsample or category *a*
- $T = \Sigma\, T_{a.} = \Sigma\Sigma\, x_{ai}$ which is the grand total of all the observations in the sample
- $\bar{x} = T / n$ which is the grand mean of all the observations in the sample

C. In order to derive two appropriate estimates of the population variance, it is first necessary to break the total variation in the sample, which is represented by the deviations about the grand mean, into two parts -- one which can provide an estimate of the population variance from the differences between the category totals, and one which can provide such an estimate from differences within the categories. If we begin with the deviation between a single observation x_{ai} and the grand mean, we can add zero in the form of ($\bar{x}_{a.} - \bar{x}_{a.}$) and square the result, to give:

$$x_{ai} - \bar{x} = (\bar{x}_{a.} - \bar{x}) + (x_{ai} - \bar{x}_{a.}) \quad \textbf{(3.1)}$$

$$(x_{ai} - \bar{x})^2 = [(\bar{x}_{a.} - \bar{x}) + (x_{ai} - \bar{x}_{a.})]^2 = (\bar{x}_{a.} - \bar{x})^2 + 2(\bar{x}_{a.} - \bar{x})(x_{ai} - \bar{x}_{a.}) + (x_{ai} - \bar{x}_{a.})^2$$

Summing both sides of this equation over *i* and *a* yields the sum of the squared deviations about the grand mean for the entire sample, which is the total variation:

$$\sum_a \sum_i (x_{ai} - \bar{x})^2 = \sum_a \sum_i (\bar{x}_{a.} - \bar{x})^2 + 2\sum_a \sum_i (\bar{x}_{a.} - \bar{x})(x_{ai} - \bar{x}_{a.}) + \sum_a \sum_i (x_{ai} - \bar{x}_{a.})^2 \quad \textbf{(3.2)}$$

However, the middle term of this expression is, in fact, equal to zero. If we look just at the inner summation, which is the sum over all values of *i* within the category, this term is:

$$\sum_i (\bar{x}_{a.} - \bar{x})(x_{ai} - \bar{x}_{a.})$$

But within any one category or subsample, the quantity $(\bar{x}_{a.} - \bar{x})$ is constant. Thus, we can move this term to the left of the summation symbol, giving the following expression:

$$(\bar{x}_{a.} - \bar{x}) \sum_i (x_{ai} - \bar{x}_{a.})$$

This leaves a constant times the sum of the deviations of each observation within a category from the category mean. Recalling that $\sum (x_i - \bar{x})$ is equal to zero implies that this sum for each category will also be equal to zero. Thus, the entire middle term in equation 3.2 drops out, leaving:

$$\sum\sum (x_{ai} - \bar{x})^2 = \underbrace{\sum\sum (\bar{x}_{a.} - \bar{x})^2}_{\text{BSS}} + \underbrace{\sum\sum (x_{ai} - \bar{x}_{a.})^2}_{\text{WSS}} \quad \textbf{(3.3)}$$

D. Equation 3.3 says that the total sum of the squared deviations about the grand mean can be decomposed into two components: (*n* times the sum of the squared deviations of each category mean about the grand mean) + (the sum, for all categories, of the sum of the squared deviations of each observation in category *a* about its category mean).

1. The first term on the right side of equation 3.3 is commonly termed the *between groups sum of squares* (BSS) since it is obtained by summing across each category or "group." The second term on the right side of 3.3 is commonly termed the *within groups sum of squares* (WSS) since it is obtained by summing within each category or group. Please note that the BSS is the equivalent of *Sum of Squares Treatment* (SST), and WSS is the equivalent of *Sum of Squares Error* (SSE) in the Lind, Marchal, and Wathen text.

2. Conceptually, the between groups sum of squares (sum of squares treatment in the text) represents the *explained variation* (referred to as the *treatment variation* in the text), since this is the variation in the sample that is accounted for by the

categorization derived from the values of the independent variable. The within groups sum of squares (sum of squares error in the text) represents the *unexplained variation* (referred to as the *random variation* in the text), since this variation is not accounted for by the categorization.

3. The interpretation of these two quantities in terms of explained and unexplained variation provides a key perspective on how ANOVA works. The explained (or treatment) variation will be large relative to the total variation whenever the category means differ markedly. This implies that the categorical variable accounts for the majority of the variation in the dependent variable within the sample. In this case, the ratio of a variance estimate based on the BSS (or SST) to that based on the WSS (or SSE) would be large, allowing rejection of the null hypothesis. If the category means do not differ substantially, then the variation within the categories will not be markedly different from the variation between the categories, and this ratio will be close to one, preventing rejection of the null hypothesis.

4. Alternative names for the between groups sum of squares include the *explained sum of squares*, the *variation between samples*, the *variation due to Factor A*, and the *sum of squares due to Factor A*. Alternative nomenclature for the within groups sum of squares includes the *residual sum of squares*, the *variation within samples*, the *error variation*, and the *sum of squares due to replications within levels of Factor A* (but I certainly do not know anyone who would admit to using this last form).

E. Deriving Alternative Estimates for the Population Variance

1. The estimate of σ^2 that goes in the numerator of the F statistic is based on the between groups sum of squares adjusted for the respective degrees of freedom, whereas the estimate that goes in the denominator of the F statistic is based on the within groups sum of squares adjusted for its degrees of freedom.

2. The degrees of freedom associated with the BSS (or SST) are derived from the fact that each category mean is treated as an observation. But to compute the deviations between the category means we first have to estimate the grand mean $\overline{x}$, which uses up one degree of freedom. Since there are k categories, the number of degrees of freedom associated with the BSS is $(k - 1)$.

3. The degrees of freedom associated with the WSS (or SSE) are derived from the fact that the deviations within each category require the category mean to first be estimated. Thus, there are $(n_a - 1)$ degrees of freedom within a category and

$$\sum_a \sum_i (n_a - 1) = \sum_i^{(n\,df)} n_a - \sum_a^{(k\,df)} 1 = (n - k)$$

$n - k$ degrees of freedom for these deviations within the entire sample.

4. Note that the degrees of freedom for the total variation in the sample equals the within groups degrees of freedom plus the between groups degrees of freedom, or:

$$n-1 = (n-k)+(k-1)$$
$$df_{\text{total}} = df_{\text{WSS}} + df_{\text{BSS}}$$

5. Thus, the estimates for the population variance based on the WSS and the BSS are, respectively:

$$s^2_{\text{WSS}} = \frac{\sum\sum(x_{ai} - \bar{x}_{a.})^2}{n-k} = \frac{\text{WSS}}{n-k} \quad \text{and} \quad s^2_{\text{BSS}} = \frac{\sum\sum(x_{a.} - \bar{x})^2}{k-1} = \frac{\text{BSS}}{k-1} \qquad \textbf{(3.4)}$$

These estimates of the variance are also respectively called the *within sample mean square* ($\text{MS}_{R(A)}$) and the *mean square for Factor A* (MS_A). The *R* subscript in the former indicates that it is based on random variability after Factor *A* is accounted for. Lind, Marchal, and Wathen refer to these estimates of the variance as the mean square error (MSE) and the mean square for treatments (MST), respectively.

III. The Test Statistic

A. The null hypothesis is H_0: $\mu_1 = \mu_2 = ... = \mu_k$ and the alternative hypothesis is H_1: not H_0.

B. This test is analogous to the independent-samples *t* test for comparing two means. However, as in the case of testing a single hypothesized value for σ^2, the comparison between the two independent sample variances (s_1^2 and s_2^2) is carried out via a ratio rather than a difference. The sampling distribution for this ratio has been named the *F distribution*, with its properties worked out by R.A. Fisher (for whom it is named). Assuming normally distributed populations, and a null hypothesis of the form H_0: $\sigma_1^2 = \sigma_2^2$, the test statistic is:

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1,\ n_2 - 1\ df)$$

This is termed the *independent-samples F test of the difference between two variances*. If H_0 is true, then the *F* distribution is the sampling distribution of the ratio of two independent estimates of the same Gaussian-population variance.

C. The *F* distribution, like the χ^2 distribution, is both asymmetric (skewed right) and always positive (since the ratio compares two squared values). It also takes on different shapes, depending on the number of degrees of freedom. But unlike either the *t* or χ^2 distributions, the *F* distribution's shape varies based on the degrees of freedom in the numerator ($n_1 - 1$) and in the denominator ($n_2 - 1$).

1. The *F* distribution is related to the normal distribution, for as the number of degrees of freedom increases, the *F* distribution becomes Gaussian in shape.

2. An *F* distribution with one degree of freedom in the numerator is directly related to the *t* distribution, in that: $F_\alpha(1, r \text{ df}) = [t_{\alpha/2}(r \text{ df})]^2$

D. Like the single-sample χ^2 test for variances, the *F* test is also very sensitive to the assumption of a normal population.

E. Since the *F* statistic allows a comparison of the variances from two independent samples, it is the logical test statistic to use for ANOVA. In fact, Ronald Fisher, for whom the *F* statistic is named, is credited with developing the ANOVA technique.

1. Assuming that the estimates of σ^2 are obtained from independent random samples, that the underlying populations are normally distributed, and that the variances for the categories are equal, then this statistic will have an *F* distribution with $(k-1,\ n-k\ \text{df})$:

$$\frac{s^2_{\text{BSS}}}{s^2_{\text{WSS}}} = \frac{\text{MS}_A}{\text{MS}_{R(A)}} \sim F(k-1,\ n-k\ df) \qquad \textbf{(3.5)}$$

2. If the underlying populations are normally distributed, the subsamples derived via the categorization on the independent variable will be independent.

F. Critical values of the *F* distribution are obtained from a table of the *F* distribution. However, the nature of this application of the *F* statistic makes obtaining critical values analogous to the procedure used for Pearsons' X^2.

1. Since the squared deviations will always be positive, the only situation that will allow us to reject the null hypothesis occurs when the numerator is considerably larger than the denominator. This implies that only the right tail of the distribution is necessary for hypothesis tests. Once again we have a non-directional alternative hypothesis, with a null hypothesis that can only be rejected if the calculated value of the F statistic falls in the right tail of the *F* distribution. Thus, the significance level is always α -- do <u>not</u> divide α by 2.

2. The relationship between ANOVA for a categorical variable with two categories and the independent-samples *t* test is partly apparent from the fact that an *F* with $(1,\ n-1\ \text{df})$ will be equal to the square of a *t* statistic with $(n-1\ \text{df})$ at the same significance level.

G. Computational Formulas

1. Although formulas 3.3 and 3.4 provide the theoretical basis for ANOVA, they are not the most convenient computational forms for deriving the BSS, WSS, and the total sum of squares (TSS).

2. The *computational formula for the TSS* is:

$$\begin{aligned}\text{TSS} = \sum\sum (x_{ai} - \bar{x})^2 &= \sum\sum x^2_{ai} - \frac{\left(\sum\sum x_{ai}\right)^2}{n} \\ &= \sum\sum x^2_{ai} - \frac{T^2}{n}\end{aligned} \qquad \textbf{(3.6)}$$

The equivalent computational formula in the Lind, Marchal, and Wathen text for the TSS (referred to as the SS total) is:

$$\text{SS total} = \sum X^2 - \frac{(\sum X)^2}{n}$$

where $\sum X^2$ is equal to sum of the squared X values and $(\sum X)^2$ is equal to the sum of the X values squared (Lind et al., 2002).

3. The *computational formula for the BSS* is:

$$\begin{aligned} \text{BSS} &= \sum\sum(\bar{x}_{a.} - \bar{\bar{x}})^2 = \sum \frac{\left(\sum x_{ai}\right)^2}{n_a} - \frac{\left(\sum\sum x_{ai}\right)^2}{n} \\ &= \sum \frac{T_{a.}^2}{n_a} - \frac{T^2}{n} \end{aligned} \qquad \textbf{(3.7)}$$

The equivalent computational formula in the text for the BSS (referred to as the SST) is:

$$\text{SST} = \sum\left(\frac{T_c^2}{n_c}\right) - \frac{(\sum X)^2}{n}$$

where T_c is the column total for each treatment and n_c is the number of observations in each treatment (Lind et al., 2002).

4. The *computational formula for the WSS* is:

$$\begin{aligned} \text{WSS} &= \sum\sum(x_{ai} - \bar{x}_{a.})^2 = \sum\sum x_{ai}^2 - \sum \frac{\left(\sum x_{ai}^2\right)}{n_a} \\ &= \sum\sum x_{ai}^2 - \sum \frac{T_{a.}^2}{n_a} \end{aligned} \qquad \textbf{(3.8)}$$

The equivalent computational formula for the WSS (referred to as the SSE) using the notation in the Lind, Marchal, and Wathen text is:

$$\text{SSE} = \sum X^2 - \sum\left(\frac{T_c^2}{n_c}\right)$$

5. Since the TSS = BSS + WSS (or SS total = SST + SSE), only two of these formulas need to be used, with this identity then providing the last component.

6. We can also estimate the standard error for each of the category means. This is found from the formula:

$$s_{\bar{x}_a} = \sqrt{s_{\text{WSS}}^2 / n_a} = \sqrt{\text{MS}_{R(A)} / n_a}$$

H. Example of One-Way Analysis of Variance (murder rate data on following page)

Assumptions: Independent random samples; normality of underlying populations; and equal population variances.

Hypotheses: H_0: $\mu_1 = \mu_2 = \mu_3$ H_1: The three population means are not all equal.

Calculations:

TSS = $[(4.3)^2 + (2.8)^2 + (12.3)^2 + ... + (11.4)^2 + (1.9)^2] - [(161.0)^2 / 24]$
= $1453.575 – [25921 / 24]$ = $1453.575 – 1080.042$ = **373.533**

BSS = $\{[(68.6)^2 / 8] + [(44.8)^2 / 8] + [(47.6)^2 / 8]\} – [(161.0)^2 / 24]$
= $1122.345 – 1080.042$ = **42.303**

WSS = $[(4.3)^2 + (2.8)^2 + (12.3)^2 + ... + (11.4)^2 + (1.9)^2]$
$– \{[(68.6)^2 / 8] + [(44.8)^2 / 8] + [(47.6)^2 / 8]\}$
= $1453.575 – 1122.345$ = **331.230**

TSS = BSS + WSS = 42.303 + 331.230 = 373.533 is a way of checking your calculations.

s^2_{BSS} = **MS_A** = $42.303 / (3 – 1)$ = **21.152**
s^2_{WSS} = **$MS_{R(A)}$** = $331.230 / (24 – 3)$ = **15.773**[3]

$F_{(2, 21)}$ = **21.152 / 15.773 = 1.34** Critical value of $F_{(2, 21; .05)} = 3.47$

Conclusion: Since 1.34 < 3.47, the null hypothesis that all three means are equal, or of no relationship between community type and murder rate, cannot be rejected at the .05 level of significance.

Calculation of **estimated standard errors:** $s_1 = \sqrt{15.773/8} = \sqrt{1.972} = 1.404 = s_2 = s_3$, since in this case, $n_1 = n_2 = n_3 = 8$

[3] Lind, Marchal, and Mason equivalent notation: TSS=SS total, BSS=SST, WSS=SSE, MS_A=MST, $MS_{R(A)}$=MSE

.

Sample ANOVA Tables: Generic and Murder Rate Example

Categories of Factor A

	A_1	A_2	. . .	A_k	Total
Observation Values	$x_{1,1}$ $x_{1,2}$ . . $x_{1,n1}$	$x_{2,1}$ $x_{2,2}$ . . $x_{2,n2}$	x_{ai}	$x_{k,1}$ $x_{k,2}$ . . $x_{k,nk}$	
Sums	$\sum x_{1i}$ $T_{1.}$	$\sum x_{2i}$ $T_{2.}$		$\sum x_{ki}$ $T_{k.}$	$\sum\sum x_{ai}$ $\sum T_{a.} = T$
Means	$\bar{x}_{1.}$	$\bar{x}_{2.}$		$\bar{x}_{k.}$	$\bar{x}$
n's	n_1	n_2		n_k	n

ANOVA Murder Rate Data Categorized by Community Type

	Industrial Community	Trade Community	Recreational Community	Total
Murders Per 100,000 Population	4.3 2.8 12.3 16.3 5.9 7.7 9.1 10.2	5.1 6.2 1.8 9.5 4.1 3.6 11.2 3.3	12.5 3.1 1.6 6.2 3.8 7.1 11.4 1.9	
Sums	68.6	44.8	47.6	161.0
Means	8.58	5.60	5.95	6.71
n's	8	8	8	24

TSS = 373.533, BSS = 42.303, WSS = 331.235
$s^2_{BSS} = 21.152$, $s^2_{WSS} = 15.773$, $F = 1.34$

IV. Two-Way Analysis of Variance

A. ANOVA can be readily extended to two (or more) independent categorical variables, if the number of cases within each cell or subcategory formed by the intersection of the values of the two categorical tables is equal. If this condition is met, hypotheses concerning both independent variables can be tested by referring to row effects and column effects.

1. *Row effects* are derived from the categories of the second variable. These categories are listed on the left-side of the ANOVA data table. The second variable is generally called variable *B* or *Factor B.* (Lind, Marchal, and Wathen refer to the second treatment variable as the *blocking variable*.)

2. *Column effects* are derived from the categories of the first variable, or Factor *A*, and are the same as those of a one-way ANOVA. These categories are listed at the top of the ANOVA data table.

3. The combination of the two independent variables yields a matrix of cells similar in structure to a contingency table. However, the contents of each cell is a group of observations, with the number of observations in each group being the same for all cells.

B. Given the above assumption, a sum of squares can be computed for Factor *B* based on the deviations of the row means from the grand mean. In other words, a one-way ANOVA is carried out down the rows of the table, directly analogous to what is done across the columns of a one-way ANOVA.

1. The total variation in the data can then be broken down into three components:

$$TSS = BcSS + BrSS + WgSS$$

where BcSS = Between Column SS (BSS in one-way ANOVA),
BrSS = Between Row SS, (due to Factor *B*), and
WgSS = Within Group (Cell) Sum of Squares.

The Lind, Marchal, and Wathen text breaks the variation down into the components SS total, SST, SSE, and SSB, such that:

$$SSE = SS\ total - SST - SSB$$

where SS total and SST are calculated the same way they were using the one-way ANOVA formulas, SSE is obtained using the above formula, and SSB is the sum of squares blocks and is equal to:

$$SSB = \sum\left(\frac{B_t^2}{k}\right) - \frac{(\sum X)^2}{n}$$

B_t is the total for the block (or row) and k is the number of observations in each block (Lind et al., 2002).

2. A portion of the total variation that is not explained or accounted for by the column variable (Factor *A*) may, in this way, be explained by the row variable (Factor *B*). The variation that is left unexplained by both Factor *A* and *B* will be contained in the WgSS. This latter component represents the unexplained, residual, or error sum of squares.

3. Assuming that Factor *A* and Factor *B* do not interact in a way that influences values of the dependent variable (i.e, there is no interaction effect, as explained below), then three independent estimates of the population variance can be obtained, with two *F* statistics being used to test the respective null hypotheses. One *F* statistic would compare the estimated variance based on the BcSS to the estimate based on the WgSS (adjusted by the appropriate df), and the other would compare the estimate based on the BrSS to the WgSS. The respective null and alternative hypotheses are:

   $H_0(A)$: $\mu_{1.} = \mu_{2.} = ... = \mu_{a.}$ H_1: The $\mu_{a.}$ are not all equal for Factor *A*.

   $H_0(B)$: $\mu_{.1} = \mu_{.2} = ... = \mu_{.b}$ H_1: The $\mu_{.b}$ are not all equal for Factor *B*.

4. Each *F* statistic would test for a relationship between one of the factors and the dependent variable, while simultaneously *controlling for* the effects of the other factor. This is accomplished by removing the variation due to the *control variable* (the factor whose SS is not included in the test statistic) by extracting this variable's SS from the unexplained variation.

C. Two-Way ANOVA with Independent Samples of Equal Size

1. The above discussion assumed that there was no additional effect on the dependent variable created by Factors *A* and *B* working together. However, two-way ANOVA is generally carried out by including a potential interaction effect in the analysis. In fact, a major purpose of performing a two-way ANOVA is often to test such a null hypothesis of the absence of an interaction effect between the main factors. This type of ANOVA is termed *two-way analysis of variance with independent samples of equal size*.

2. An *interaction effect* is defined by Olson as:

   > "The extent to which the differences between the levels of one factor are different at different levels of another factor; the unique effect of the combination of two or more factors; an effect that cannot be accounted for by the sum of the effects of the separate factors."

   a. *Interaction* refers to Factors *A* and *B* working together to influence the level of the dependent variable in a manner different from their individual effects on this variable. An interaction effect implies that the effects of a factor on the dependent variable will differ depending on which category of the other factor is being considered.

b. For example, if we add "region" as Factor *B* to the Murder Rate example, an interaction effect would imply that the impact of "city type" on murder rate would be different in the northeast than it would in the southeast.

3. If an interaction effect between the factors is not found in the population, then a property termed *additivity* should hold within the sample. This property implies that the differences among the cell means between columns are the same for each row, and that the differences among cell means between rows are the same for each column.

   a. For example, in the following table of cell means, the change in row B_1 between the cell mean A_1B_1 and A_2B_1 is 5 units, and the change between A_2B_1 and A_3B_1 is 10 units. These differences are the same between the respective cells in rows B_2 and B_3. The same relationships are apparent with respect to the columns:

	A_1	A_2	A_3
B_1	5	10	20
B_2	10	15	25
B_3	25	30	40

   If there is no interaction effect arising from Factors *A* and *B*, the difference of means between these categories will be constant, as in the above example. If an interaction effect does exist in the population, this additivity principle will not be found in the sample data.

   b. The null hypothesis for an interaction effect is based on this additivity concept, in that each cell mean, in the presence of such an effect, would be derived solely from the grand mean plus a contribution from Factor *A* and one from Factor *B*. No other influence (such as the interaction of Factors *A* and *B*) is present. Thus, the null hypothesis can be stated as:

$$H_0(AB):\ \mu_{ab} = \mu + (\mu_{a.} - \mu) + (\mu_{.b} - \mu) \quad \text{for all combinations of } a \text{ and } b$$

   where μ_{ab} is the mean of the population of observations for category *a* of Factor *A* and category *b* of Factor *B*.

4. Additional notation for two-way ANOVA:

   - k_A represents the number of categories or subsamples for Factor *A*,
   - k_B represents the number of categories or subsamples for Factor *B*,
   - x_{abi} represents the *i*th observation in a cell formed by the intersection of category *a* of Factor *A* and category *b* of Factor *B* (termed "cell *ab*"),
   - n_b is the number of observations in the *b*th subsample or category of Factor *B*,
   - n_R is the number of observations (replications) contained in each cell,
   - $n = \Sigma\, n_a = \Sigma\, n_b = k_A k_B n_R$ and is the total number of observations in the sample,
   - $T_{a..}$ = the sum of the observations in column *a*,
   - $T_{.b.}$ = the sum of the observations in row *b*, ($\Sigma_b\, x_{abi}$), the dot (.) indicating this is summed over all values of the subscripts *a* and *i*,

- $T_{ab.}$ = the sum of the observations in cell *ab*,
- $\bar{x}_{a..} = T_{a..} / n_a$ which is the mean of the observations in subsample or category *a* (i.e., column *a*),
- $\bar{x}_{.b.} = T_{.b.} / n_b$ which is the mean of the observations in subsample or category *b* (i.e., column *b*),
- $\bar{x}_{ab.} = T_{ab.} / n_R$ which is the mean of the observations in cell *ab*,
- $T = \Sigma\Sigma\Sigma\, x_{abi}$ which is the grand total of all the observations in the sample,
- $\bar{x} = T / n$ which is the grand mean of all the observations in the sample

5. Analogous to the case of the one-way ANOVA, we need to decompose the TSS into four components -- one due to the main effects of Factor *A* (BcSS), one due to the main effects of Factor *B* (BrSS), one due to the interaction effects of Factors *A* and *B* (ISS), and one due to unexplained influences and random error (the residual or unexplained variation -- WgSS). The equation for this decomposition of a two-way ANOVA is:

$$\sum_a \sum_b \sum_i (x_{abi} - \bar{x})^2 = \sum_a \sum_b \sum_i (\bar{x}_{a..} - \bar{x})^2 + \sum_a \sum_b \sum_i (\bar{x}_{.b.} - \bar{x})^2 +$$

$$TSS \quad = \quad BcSS \quad + \quad BrSS \quad +$$

(3.9)

$$\sum_a \sum_b \sum_i (\bar{x}_{ab.} - \bar{x}_{a..} - \bar{x}_{.b.} + \bar{x})^2 + \sum_a \sum_b \sum_i (x_{abi} - \bar{x}_{ab.})^2$$

$$ISS \quad + \quad WgSS$$

$$\text{TSS} = \text{BcSS} + \text{BrSS} + \text{ISS} + \text{WgSS}$$

6. The degrees of freedom associated with each of these quantities are:

$$\text{TSS}_{df} = \text{BcSS}_{df} + \text{BrSS}_{df} + \text{ISS}_{df} + \text{WgSS}_{df}$$

$$n - 1 = (k_A - 1) + (k_B - 1) + (k_A - 1)(k_B - 1) + (n - k_A k_B)$$

$$= k_A + k_B - 2 + k_A k_B - k_B - k_A + 1 + n - k_A k_B = n - 1$$

7. The three null hypotheses to be tested are then:

$H_0(A)$: $\mu_{1.} = \mu_{2.} = ... = \mu_{a.}$ $\quad$ H_1: The $\mu_{a.}$ are not all equal for Factor *A*.

$H_0(B)$: $\mu_{.1} = \mu_{.2} = ... = \mu_{.b}$ $\quad$ H_1: The $\mu_{.b}$ are not all equal for Factor *B*.

$H_0(AB)$: $\mu_{ab} = \mu + (\mu_{a.} - \mu) + (\mu_{.b} - \mu) = \mu_{a.} + \mu_{.b} - \mu$ for all combinations of *a* and *b*

H_1: The population cell means (μ_{ab}) are not completely determined by the population marginal means ($\mu_{a.}$, $\mu_{.b}$)

8. Computational Formulas for Two-Way ANOVA with Independent Samples of Equal Size:

$$BcSS = SS_A = \left(\frac{1}{k_B n_R} \sum_{a=1}^{k_A} T_{a..}^2 \right) - \frac{T^2}{n} \quad \textbf{(3.10)}$$

$$BrSS = SS_B = \left(\frac{1}{k_A n_R} \sum_{b=1}^{k_B} T_{.b.}^2 \right) - \frac{T^2}{n} \quad \textbf{(3.11)}$$

$$ISS = SS_{AB} = \left(\frac{1}{n_R} \sum_{a=1}^{k_A} \sum_{b=1}^{k_B} T_{ab.}^2 \right) - BcSS - BrSS - \frac{T^2}{n} \quad \textbf{(3.12)}$$

$$WgSS = SS_{R(AB)} = \sum_{a=1}^{k_A} \sum_{b=1}^{k_B} \sum_{i=1}^{n_R} x_{abi}^2 - \frac{1}{n_R} \sum_{a=1}^{k_A} \sum_{b=1}^{k_B} T_{ab.}^2 \quad \textbf{(3.13)}$$

$$TSS = SS_{total} = \sum_{a=1}^{k_A} \sum_{b=1}^{k_B} \sum_{i=1}^{n_R} x_{abi}^2 - \frac{T^2}{n} \quad \textbf{(3.14)}$$

9. *F* Statistics for Two-Way ANOVA

   Main Effect of Factor *A*:

   [BcSS / ($k_A - 1$)] / [WgSS / ($n - k_A k_B$)] ~ F [($k_A - 1$); ($n - k_A k_B$)]

   Main Effect of Factor *B*:

   [BrSS / ($k_B - 1$)] / [WgSS / ($n - k_A k_B$)] ~ F [($k_B - 1$); ($n - k_A k_B$)]

   Interaction Effect of Factors *A* and *B*:

   [ISS / ($k_A - 1$)($k_B - 1$)] / [WgSS / ($n - k_A k_B$)] ~ F [($k_A - 1$)($k_B - 1$); ($n - k_A k_B$)]

10. The estimate of the standard error for any of the means in the two-way ANOVA is obtained in the same manner as for the one-way ANOVA. The estimate of the variance based on the WgSS ($MS_{R(AB)}$) is divided by the number of observations on which the mean is based. The square root of this result is estimate of the standard error.

D. Example of Two-Way Analysis of Variance

Two-Way ANOVA Murder Rate Data Categorized by City Type and Region

Region	Industrial Community	Trade Community	Recreational Community	Total
NE	4.3 5.9 2.8 7.7 $\Sigma x = 20.7$ $\overline{x} = 5.18$	5.1 3.6 1.8 3.3 $\Sigma x = 13.8$ $\overline{x} = 3.45$	3.1 3.8 1.6 1.9 $\Sigma x = 10.4$ $\overline{x} = 2.60$	$\Sigma x = 44.9$ $\overline{x} = 3.74$
SE	12.3 9.1 16.3 10.2 $\Sigma x = 47.9$ $\overline{x} = 11.98$	6.2 4.1 9.5 11.2 $\Sigma x = 31.0$ $\overline{x} = 7.75$	6.2 11.4 7.1 12.5 $\Sigma x = 37.2$ $\overline{x} = 9.30$	$\Sigma x = 116.1$ $\overline{x} = 9.68$
Sums	68.6	44.8	47.6	161.0
Means	8.58	5.60	5.95	6.71

Assumptions: Independent random samples; normality of underlying populations; and equal population variances.

Hypotheses: $H_0(A)$: $\mu_{1.} = \mu_{2.} = \mu_{3.}$ H_1: The $\mu_{a.}$ are not all equal for Factor *A*.

$H_0(B)$: $\mu_{.1} = \mu_{.2}$ H_1: The $\mu_{.b}$ are not all equal for Factor *B*.

$H_0(AB)$: $\mu_{ab} = \mu_{a.} + \mu_{.b} - \mu$ for all combinations of *a* and *b*

H_1: The population cell means (μ_{ab}) are not completely determined by the population marginal means ($\mu_{a.}$, $\mu_{.b}$)

Calculations:

TSS = $[(4.3)^2 + (2.8)^2 + (12.3)^2 + ... + (11.4)^2 + (12.5)^2] - [(161.0)^2 / 24]$
= **373.533**

BcSS = $\{[(68.6)^2 / 8] + [(44.8)^2 / 8] + [(47.6)^2 / 8]\} - [(161.0)^2 / 24]$
= **42.303**

BrSS = $\{[(44.9)^2 / 12] + [(116.1)^2 / 12]\} - [(161.0)^2 / 24]$
= **211.226**

ISS = $(1/4) [(20.7)^2 + (13.8)^2 + ... + (37.2)^2] - 1080.042 - 211.226 - 42.303$
= **8.014**

WgSS = $[(4.3)^2 + (2.8)^2 + (5.9)^2 + (7.7)^2 + (12.3)^2 + ... + (11.4)^2 + (12.5)^2]$
$- (1/4) [(20.7)^2 + (13.8)^2 + ... + (37.2)^2]$ = **111.990**

Since TSS = BcSS + BrSS + ISS + WgSS
= 42.303 + 211.226 + 8.014 + 111.990 = 373.533 checks calculations.

s^2_{BcSS} = **MS_A** = 42.303 / (3 – 1) = **21.152**

s^2_{BrSS} = **MS_B** = 211.226 / (2 – 1) = **211.226**

s^2_{ISS} = **$MS_{(AB)}$** = 8.014 / [(3 – 1)(2 – 1)] = **4.007**

s^2_{WgSS} = **$MS_{R(AB)}$** = 111.995 / [24 – (3)(2)] = **6.222**

Interaction Hypothesis Test

$F_{ISS\,(2,\,18)} = 4.007 / 6.222 = 0.64$ Critical value of $F_{(2,18;\,.05)} = 3.55$

Conclusion: Since 0.64 < 3.55, the null hypothesis that there is no interaction effect between community type and region cannot be rejected at the .05 level of significance.

Factor *A* (Community Type) Main Effects Hypothesis Test

$F_{A(2,\,18)} = 21.152 / 6.222 = 3.40$ Critical value of $F_{(2,18;\,.05)} = 3.55$

Conclusion: Since 3.40 < 3.55, the null hypothesis that there is no direct relationship between community type (Factor *A*) and murder rate, while controlling for region, cannot be rejected at the .05 level of significance.

Factor *B* (Region) Main Effects Hypothesis Test

$F_{B\,(1,\,18)} = 211.226 / 6.222 = 33.95$ Critical value of $F_{(1,18;\,.05)} = 4.41$

Conclusion: Since 33.95 > 4.41, the null hypothesis that there is no direct relationship between region (Factor *B*) and murder rate, while controlling for community type, can be rejected at the .05 level of significance.

Calculation of estimated standard errors:

$s_{ISS} = \sqrt{6.222/2} = 1.764$; $s_{WgSS} = \sqrt{6.222/1} = 2.494$; ... etc.

Suppose, though, that the observed value for the *F* statistic for the interaction hypothesis test had been greater than the critical value. If this had been the case, the conclusion would be that the null hypothesis of no interaction effect between region and community type with respect to murder rate could be rejected at the specified level of significance. This would imply that, in addition to any main effects that are detected, there would appear to be a joint effect of region and community type on murder rate. In other words, the effect of region on murder rate would appear to vary by community type, and/or the effect of community type on murder rate would appear to vary by region.

E. Two-Way ANOVA without Replication

1. If improving the precision of the ANOVA for Factor *A* is the main concern, and an explicit hypothesis test of Factor *B* is not of primary importance, a two-way ANOVA design termed *Two-Way ANOVA without Replication* can be used. This is the form of two-way ANOVA which is presented in Lind, Marchal, and Wathen.

2. The differences between this and the more general two-way ANOVA described above is that the data is *blocked* so that each cell contains only one observation, and that potential for an interaction effect is ignored. This is simply a special case (and a simpler case) of the more general two-way ANOVA with independent samples of equal size.

3. The purpose of using Factor *B* to create the blocked analysis is analogous to paired-sample *t* tests. In this case, the variation due to Factor *B* is removed or controlled for, thus reducing the noise in the system and improving the precision of the estimates for the means and standard errors associated with Factor *A*.

4. The BcSS (for Factor *A*'s main effects), the BrSS (for Factor *B*'s main effects), and the TSS are derived using formulas 3.10, 3.11, and 3.14, respectively. The error sum of squares (SSE) has a slightly different formula from that of the WgSS for the more generic two-way ANOVA. This is:

$$\text{SSE} = \sum\sum[x_{ab} - (\bar{x}_{a.} - \bar{x}) - (\bar{x}_{.b} - \bar{x}) - \bar{x}]^2 \qquad \textbf{(3.15)}$$

and has $(k_A - 1)(k_B - 1)$ degrees of freedom.

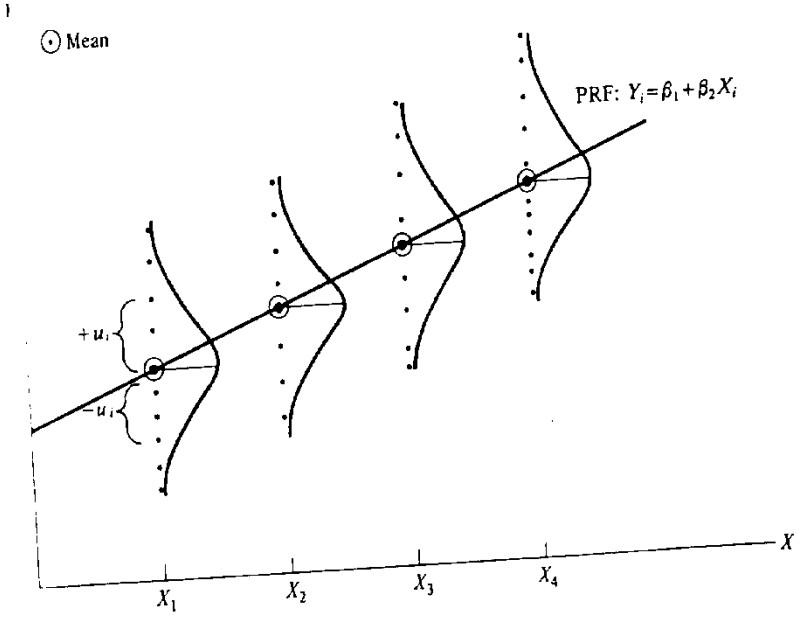
5. The *F* statistics are derived in the usual way. A test of the null hypothesis for Factor *B* may or may not be undertaken, although a large value for the estimate of the variance based on Factor *B* indicates that the blocking was successful.

BIVARIATE REGRESSION: ASSUMPTIONS, ESTIMATION, AND THE GAUSS-MARKOV THEOREM

I. The Regression Function

A. Introduction

1. Assume we have two variables, *X* and *Y*, for which *X* has a series of fixed values and *Y* takes on several different values for each *X* (e.g., *X* might be equal to years of education and *Y* equal to income for a population). If we plot the positions of the conditional means of *Y* and connect these points, we form a curve which describes the nature of the statistical relationship between *X* and *Y*:



The curve does not tell us exactly what *Y* will be for a given value of *X*, but it does tell us the expected value of *Y*, i.e., $[\mathrm{E}(Y|X)]$.

2. If the relationship described by the curve is linear, we can derive an equation which represents the curve mathematically. Since an equation for a line can be generated from its slope and intercept with the vertical axis, we have:

$$\mathrm{E}(Y|X) = \beta_1 + \beta_2 X_i \qquad \textbf{(3.16)}$$

where β_1 is the intercept and β_2 is the slope

3. We can write an exact equation (one without conditional expectations) by specifically accounting for the variation of each *Y* value around its conditional mean. This is accomplished by adding the stochastic "error term" to equation 3.16 and removing the expectation operator:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \textbf{(3.17)}$$

where β_1 is the intercept, β_2 is the slope, and u_i is the disturbance or error term resulting from measurement errors or outside influences on *Y* which are not accounted for by *X*.

This equation forms the basis for the structure of the *two variable classical linear regression model*. It is a *stochastic equation* for the use of u_i implies that for every value of *X* there is a probability distribution for *Y* (i.e., the value of *Y* can never be predicted exactly).

4. If each value of *u* is a result of many small causes and measurement errors, we can expect that these effects will generally be random and normally distributed, since there will be many more small deviations from conditional means than large deviations. We would also expect that, in the long run, these random disturbances would cancel each other out. Thus, the distribution of *u* would have a mean of zero [E (u_i) = 0]. In order to obtain unbiased estimates of the values of β_1 and β_2, it is necessary to make this last assumption as well as several others.

B. Assumptions of the Two Variable Linear Model

1. E (u_i) = 0	zero mean
2. Cov ($u_i\, u_j$) = 0 for $i \neq$ j	nonautoregression or no serial correlation
3. Var (u_i) = E (u_i^2) = σ^2	homoscedasticity
4. Cov (*X*, *u*) = 0	no relationship between X and the error term, or alternatively can use X is fixed
5. Correct specification	equation 3.17 is the Pop. Regression Function

These assumptions, together with equation 3.17, make up the two variable classical linear regression model. Assumption 2 implies that knowing one value of *u* will not allow prediction of any other values of *u*, whereas assumption 3 states that the variance of *u* is constant. Assumption 4 is a simplification over using $\mathrm{E}(Y|X)$; alternatively, *X* and *u* can be assumed to be independent. A hypothesis (or assumption) of linearity in the parameters is inherent in the form of equation 3.17.

II. Estimation and the Gauss-Markov Theorem

A. As noted below, the Gauss-Markov theorem states that, when the above assumptions are met, the best method for estimation of equation 3.17 is Ordinary Least Squares (OLS). This is true of either population or sample data. Since we generally use sample data, the values of *u* are unobservable (as previously stated).

B. The OLS procedure involves finding the unique line which has the property that the sum of the squares of the deviations (vertical distances on the scattergram) of the actual values of *Y* from this line is a *minimum*. This line is the sample regression function (SRF) or sample regression equation. If $\hat{Y}_i$ is the value of *Y* on the line associated with X_i, then:

$$\sum (Y_i - \hat{Y}_i) = 0 \quad \text{(a property of least squares)}$$

$$\frac{\sum (Y_i - \hat{Y}_i)^2}{n} = \text{minimum,}$$

since this last quantity is the variance of the actual *Y* values from the *Y* values on the line (the $\hat{Y}_i s$).[4] Note that this represents another way of thinking about OLS -- it minimizes the variance of the predicted values of *Y* from the actual values of *Y*.

C. In order to obtain the equation for the sample regression function, we have to derive estimated values for β_1 and β_2 with $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively (or b_1 and b_2), which indicate the least squares estimates. Therefore, the sample regression function can be written as:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad \textbf{(3.18)}$$

Defining the *residual* as $(Y_i - \hat{Y}_i) = \hat{u}_i$, which represents the vertical distance between the regression line predicted value for Y_i and the observed value of Y_i in the sample, this equation can also be written as:

$$(Y_i - \hat{Y}_i) = \hat{u}_i \quad \text{or} \quad \hat{Y}_i = Y_i - \hat{u}_i$$

Substituting the latter equation for $\hat{Y}_i$ in equation 3.18, we get:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \qquad \textbf{(3.19)}$$

D. The idea behind OLS is to minimize $\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2$, which can be done with calculus. This procedure, which is shown on pages 82 to 83 for those interested, yields two simultaneous equations. These are termed the *normal equations* for a line (or the least squares normal equations), and are:

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad \textbf{(3.20)}$$

$$\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \qquad \textbf{(3.21)}$$

Solving equations 3.20 and 3.21 for $\hat{\beta}_1$ and $\hat{\beta}_2$, we get:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} \qquad \textbf{(3.22)}$$

---

[4] The Lind, Marchal, and Wathen text uses the symbol Y' instead of $\hat{Y}$ to denote the predicted value of *Y* for a particular *X*.

$$\hat{\beta}_2 = \frac{n(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{n(\sum X_i^2) - (\sum X_i)^2} \quad \textbf{(3.23)}$$

By some convoluted algebra, equation 3.23 reduces to:

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad \textbf{(3.24)}$$

Equation 3.24 can be rewritten in deviation form as:

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \textbf{(3.25)}$$

$$\text{where } x_i = (X_i - \bar{X}) \text{ and } y_i = (Y_i - \bar{Y})$$

E. Equation 3.22 and 3.25 (or alternatively, equation 3.24) give the least squares estimators for β_1 and β_2, respectively.

1. Equations 3.24 and 3.25 also provide some revealing information about the nature of $\hat{\beta}_2$. By multiplying the numerator and denominator by 1/*n*, the numerator is shown to be the covariance of *X* and *Y*, and the denominator is the variance of *X*.

2. The value of $\hat{\beta}_2$ is interpreted as *the estimated change in the dependent variable Y given a one unit change in the independent variable X*.

3. Note that when data are in deviation form, $\hat{\beta}_1 = 0$ (i.e., the SRF passes through the origin). The SRF also passes through the point $(\bar{X}, \bar{Y})$ which is another property of least squares.

F. Variance of the OLS Estimators

1. The variances and standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ can respectively be shown to be:

$$Var\ \hat{\beta}_1 = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2 \qquad Var\ \hat{\beta}_2 = \frac{\sigma^2}{\sum x_i^2} \quad \textbf{(3.26)}$$

$$se_{\hat{\beta}_1} = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2} \sigma^2} \qquad se_{\hat{\beta}_2} = \sqrt{\frac{\sigma^2}{\sum x_i^2}} \quad \textbf{(3.27)}$$

2. But since σ^2, the variance of the random error term (*u*), is a characteristic of the population and is generally unobservable, we estimate σ^2 with $\hat{\sigma}^2$ as:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \qquad \textbf{(3.28)}$$

The $(n-2)$ in the denominator corrects for the appropriate number of degrees of freedom so as to provide an unbiased estimate of σ^2.

G. Gauss - Markov Theorem

1. It can be demonstrated that $\hat{\beta}_1$ and $\hat{\beta}_2$ are the best linear unbiased estimators of β_1 and β_2. "Best" implies most efficient (the variance of the sampling distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$ will be the minimum for any estimators of β_1 and β_2).

2. The Gauss-Markov Theorem states that, *given the assumptions we have made about the two variable linear regression model, the OLS estimators for β_1 and β_2 will be the best linear unbiased estimators (BLUE).*

H. Derivation of OLS Estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ (Optional)

In order for $\sum \hat{u}_i^2$ to be a minimum, a necessary condition (from calculus) is that the partial derivatives of this sum ($\sum \hat{u}_i^2$) with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ must be zero. Thus, given our definition of $\hat{u}_i$:

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

$$= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \qquad \text{substituting eqn 3.18 for } \hat{Y}_i$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) \qquad \text{partial derivative w/r to } \hat{\beta}_1$$

$$-2\sum Y_i + 2\sum \hat{\beta}_1 + 2\sum \hat{\beta}_2 X_i = 0 \qquad \text{setting = to 0}$$

$$-2\sum Y_i = -2\sum \hat{\beta}_1 - 2\sum \hat{\beta}_2 X_i$$

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad \textbf{(3.29)}$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = -2 \sum X_i (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) \qquad \text{partial derivative w/r to } \hat{\beta}_2$$

$$-2\sum X_i Y_i + 2\sum \hat{\beta}_1 X_i + 2\sum \hat{\beta}_2 X_i^2 = 0 \qquad \text{setting = to 0}$$

$$-2\sum X_i Y_i = -2\sum \hat{\beta}_1 X_i - 2\sum \hat{\beta}_2 X_i^2$$

$$\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \qquad \textbf{(3.30)}$$

Equations 3.29 and 3.30 are the least squares normal equations. To derive the formulas for the estimators, we solve these simultaneously for $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$n\hat{\beta}_1 = \sum Y_i - \hat{\beta}_2 \sum X_i \qquad \text{from equation 3.29}$$

$$\hat{\beta}_1 = \frac{\sum Y_i}{n} - \hat{\beta}_2 \frac{\sum X_i}{n}$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} \qquad \text{which is the OLS estimator for } \beta_1$$

Substituting this quantity for $\hat{\beta}_1$ in equation 3.30 above gives:

$$\sum X_i Y_i = (\overline{Y} - \hat{\beta}_2 \overline{X}) \sum X_i + \hat{\beta}_2 \sum X_i^2$$

$$\sum X_i Y_i = \overline{Y} \sum X_i - \hat{\beta}_2 \overline{X} \sum X_i + \hat{\beta}_2 \sum X_i^2 \qquad \text{expanding the product}$$

$$\sum X_i Y_i - \overline{Y} \sum X_i = -\hat{\beta}_2 \overline{X} \sum X_i + \hat{\beta}_2 \sum X_i^2$$

$$= \hat{\beta}_2 (\sum X_i^2 - \overline{X} \sum X_i)$$

$$\hat{\beta}_2 = \frac{\sum X_i Y_i - \overline{Y} \sum X_i}{\sum X_i^2 - \overline{X} \sum X_i} \qquad \text{which is the OLS estimator of } \beta_2$$

Substituting $1/n \sum X_i$ and $1/n \sum Y_i$ for $\overline{X}$ and $\overline{Y}$, respectively:

$$= \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2}$$

Multiplying by *n*/*n* which gives eqn 3.23:

$$= \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

Note that the condition of setting the partial derivatives of $\hat{u}_i$ equal to zero implies that the sum of the residuals is zero.

BIVARIATE REGRESSION: STRENGTH OF RELATIONSHIP AND HYPOTHESIS TESTING

I. Measuring the Strength of Relationship: r^2

A. To measure the strength of the hypothesized statistical relationship between the dependent and independent variables of the regression equation, we use the square of the correlation coefficient "*r*". The value of r^2 can also be thought of as an indicator of "goodness of fit," or how well the sample regression line fits the sample data.

B. To see how this statistic is used, we decompose the variation of *Y* in the sample into two components.

1. We begin with the total variation of $Y = \sum (Y_i - \overline{Y})^2$, and add 0 to the right side of this expression in the form of $(\hat{Y}_i - \hat{Y}_i)$:

$$\begin{aligned}\sum (Y_i - \overline{Y})^2 &= \sum (Y_i + \hat{Y}_i - \hat{Y}_i - \overline{Y})^2 \\ &= \sum [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})]^2 \\ &= \sum (Y_i - \hat{Y}_i)^2 + 2\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) + \sum (\hat{Y}_i - \overline{Y})^2\end{aligned}$$

the middle term in this last expression can be shown to = 0, giving

$$\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \overline{Y})^2 \qquad \textbf{(3.31)}$$

Total Variation	=	Unexplained or Residual Variation	+	Explained Variation (due to regression)
Total Sum of Squares (TSS)	=	Residual Sum of Squares (RSS)	+	Explained Sum of Squares (ESS)

2. The RSS represents the "unexplained" variation, since it indicates the amount of error (or the residual) in the prediction of *Y*; i.e., it represents the difference between the actual value of *Y* and its predicted value.

3. The ESS represents the variation of the predicted values of *Y* around the mean of *Y*, and indicates the gain in predictive power achieved by using $\hat{Y}$ as a predictor of *Y* instead of $\overline{Y}$. Hence, the ESS is the amount of total variation in *Y* which is accounted for, or "explained by" the regression line or function.

4. Note that since TSS = RSS + ESS:

$$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$

5. r^2 is termed the *coefficient of determination*, and is defined as:

$$r^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \qquad \textbf{(3.32)}$$

6. r^2 can be interpreted as the proportion (or percentage) of the total variation in Y explained by (or associated with) the regression on X.

7. Alternative formulas for r^2 (or r) include:

$$r^2 = \hat{\beta}_2^2 \frac{\sum x_i^2}{\sum y_i^2} \qquad \textbf{(3.33)}$$

$$r^2 = \hat{\beta}_2^2 \left(\frac{s_x^2}{s_y^2} \right) \qquad \text{where } s^2 \text{ is the sample variance}$$

$$r = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n\sum X_i^2 - (\sum X_i)^2][n\sum Y_i^2 - (\sum Y_i)^2]}} \qquad \textbf{(3.34)}$$

II. Hypothesis Testing

A. To carry out tests of hypotheses concerning either the existence of a statistically significant relationship between X and Y, or specific values of the regression parameters, it is necessary to make an additional assumption about u. We must assume, for the purposes of computing distributions of test statistics, that u is normally distributed: $u \sim N$

B. There are two types of hypotheses which are generally tested with the two variable linear model.

1. The first concerns the existence of a linear relationship between X and Y, and can be considered a test of the statistical significance of the regression itself. This hypothesis can be tested either with an F-statistic from an analysis of variance framework or with a t-statistic.

2. The second hypothesis concerns explicitly hypothesized values for β_1 and β_2, and can only be tested using the t-statistic.

C. Tests for the Existence of a Linear Relationship between X and Y

1. The lack of a linear relationship between X and Y would imply that both r and β_2 should be equal to zero. H_0: $\rho = 0$ or H_0: $\beta_2 = 0$ can be treated in an ANOVA (analysis of variance) framework in which an F-statistic is used to compare estimates of the variance of Y based on the explained sum of squares and the residual sum of squares divided by their respective degrees of freedom.

a. The residual sum of squares has $(n - 2)$ degrees of freedom associated with it, for

we must use up two degrees of freedom to compute $\hat{\beta}_1$ and $\hat{\beta}_2$, from which we can then compute the residuals.

b. The explained sum of squares ESS = $\sum(\hat{Y}_i - \overline{Y})^2$ has one d.f. associated with it. Since $\hat{Y}_i$ has two d.f. and $\overline{Y}$ has one d.f., the difference is one d.f. Given that *u* is normally distributed, then

$$\frac{ESS/1}{RSS/(n-2)} \sim F_{1,n-2} \qquad \textbf{(3.35)}$$

or

$$\frac{\sum(\hat{Y}_i - \overline{Y})^2/1}{\sum(Y_i - \hat{Y}_i)^2/(n-2)} \sim F_{1,n-2}$$

or

$$\frac{r^2/1}{(1-r^2)/(n-2)} = \frac{r^2(n-2)}{1-r^2} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum \hat{u}_i^2/(n-2)} \sim F_{1,n-2}$$

The table of the *F*-distribution is used in the usual manner to establish a critical region at a specified level of significance. (Note - use α, not $\alpha/2$.)

D. Tests of Hypotheses Concerning Specific Values of β_1 or β_2

1. Given the assumption that *u* is normally distributed, the following statistic has a *t*-distribution with $n - 2$ degrees of freedom (we lose two degrees of freedom to derive the values of $\hat{\beta}_1$ and $\hat{\beta}_2$):

$$\frac{\hat{\beta} - \beta}{se_{\hat{\beta}}} \sim t_{n-2} \qquad \textbf{(3.36)}$$

2. To test a hypothesis concerning a specified value of β_2, we have:

$$\frac{\hat{\beta}_2 - \beta_2}{se_{\hat{\beta}_2}} = \frac{\hat{\beta}_2 - \beta_2}{\left[\frac{\sum \hat{u}_i^2 (n-2)}{\sum x_i^2}\right]^{\frac{1}{2}}} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\sqrt{\sum \hat{u}_i^2/(n-2)}}$$

$$= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\sqrt{\sum (y_i^2 - \hat{\beta}_2^2 \sum x_i^2)/(n-2)}} \qquad \textbf{(3.37)}$$

$se_{\hat{\beta}_2}$ can also be shown to equal

$$\text{se}_{\hat{\beta}_2} = \sqrt{\frac{\sum (Y_i - \hat{Y})^2/(n-2)}{\sum (X_i - \overline{X})^2}}$$

or

$$se_{\hat{\beta}_2} = \sqrt{\frac{[n\sum Y_i^2 - (\sum Y_i)^2] - \hat{\beta}_2[n\sum X_i Y_i - \sum X_i \sum Y_i]}{[n\sum X_i^2 - (\sum X_i)^2][n-2]}} \qquad \textbf{(3.38)}$$

which is sometimes useful for carrying out computations. Note that the existence of a linear relationship between *X* and *Y* can be tested with the above *t*-statistic by specifying H_0: $\beta_2 = 0$.

3. In the case of the two variable linear model, it can be shown that:

$$t_{n-2} = \sqrt{F_{1,\,n-2}}$$

so that the two tests for the existence of a linear relationship between *X* and *Y* are equivalent. *But this is true only for the two variable linear model*.

4. To test a hypothesis concerning a specified value of β_1, we have:

$$\frac{\hat{\beta}_1 - \beta_1}{se_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sum X_i^2 \sum \hat{u}_i^2}{n(\sum x_i^2)(n-2)}}}$$

Tests of hypotheses concerning β_1 are much less frequent then tests concerning β_2.

E. Interval Estimation

It is also possible to derive confidence intervals (interval estimates) for β_1 and β_2 in the usual fashion using the *t*-distribution; i.e.:

$$-t_{\alpha/2} \le \frac{\hat{\beta} - \beta}{se_{\hat{\beta}}} \le t_{\alpha/2} \quad \text{or} \quad \hat{\beta} - t_{\alpha/2} se_{\hat{\beta}} \le \beta \le \hat{\beta} + t_{\alpha/2} se_{\hat{\beta}} \qquad \textbf{(3.39)}$$

where α = the desired level of significance.

F. Standard Error of Estimate OR Standard Error of the Regression - This is the average error in predicting *Y*.

$$se_{\hat{Y}} = \sqrt{\frac{(1-r^2)[\sum Y^2 - \frac{(\sum Y)^2}{n}]}{n-2}}$$

III. V506 BIVARIATE REGRESSION EXAMPLE

Observation	Number of Drinks	Blood-Alcohol Level
1	17	160
2	20	140
3	7	120
4	14	160
5	6	110
6	30	215
7	20	210
8	28	175
9	4	100
10	12	150

$\sum X$ = **158** $\sum Y$ = **1540** $\sum XY$ = **27010** $\sum X^2$ = **3214**
$\sum Y^2$ = **250750** $\overline{X}$ = **15.8** $\overline{Y}$ = **154**

The Regression Equation (equations 3.23, 3.22, and 3.19):

$$\hat{\beta}_2 = \frac{10(27010) - (158)(1540)}{10(3214) - 24964} = 3.732$$

$$\hat{\beta}_1 = 154 - (3.732)(15.80) = 95.034$$

Indicates that an increase of one drink raises blood-alcohol level by 3.73. It also indicates that a person who has had zero drinks will have a b-a level of 95.04. However, 0 drinks was not included in our sample so extrapolations of this nature are unreliable.

Coefficient of Determination (equation 3.34):

$$r = \frac{(10)(27010) - (158)(1540)}{\sqrt{[(10)(3214) - (158)^2][(10)(250750) - (1540)^2]}} = 0.85775$$

$$r^2 = 0.7354$$

This *r* value indicates the existence of a strong positive relationship between *X* and *Y*, and implies that 73.54% of the variation in blood-alcohol level is associated with (or "explained by") the number of drinks a person has had, and that the model fits the data quite well.

***F* Test** using the formula on page 87 of V506 notes:

H_0: $\rho = 0$ Critical Value: $F_{.05(1, 8)} = 5.32$
H_1: $\rho \neq 0$

$$F = \frac{(.7354)(10-2)}{1-.7354} = 22.23$$

Since 22.23 > 5.32, reject H_0 at the .05 level of significance. The *F* test examines the significance or fit of the entire regression.

Standard Error of the Slope Coefficient (equation 3.38):

The standard error of $\hat{\beta}_2$ measures the average amount of variation in the estimate of β_2.

$$se_{\hat{\beta}_2} = \sqrt{\frac{(10)(250750) - (1540)^2 - 3.73[(10)(27010) - (158)(1540)]}{[(10)(3214) - (158)^2][8]}} = 0.792$$

***t* test** (equation 3.36):

H_0: $\beta_2 = 0$ Critical Value: $t_{.025(8)} = 2.306$
H_1: $\beta_2 \neq 0$

$$t = \frac{3.732 - 0}{.792} = 4.72$$

Therefore reject H_0. The *t* test examines the significance of a single parameter. In this case, we conclude that $\hat{\beta}_2$ is statistically significant at the .05 level.

95% Confidence Interval for β_2 (equation 3.39):

$$3.73 \pm (2.306)(.792)$$
$$1.91 \leq \beta_2 \leq 5.55$$

This interval will contain the actual value of β_2 95% of the time.

Standard Error of the Regression on page 88 of V506 notes:

$$se_{\hat{Y}} = \sqrt{\frac{(1-r^2)[\sum Y^2 - \frac{(\sum Y)^2}{n}]}{n-2}} = \sqrt{\frac{(1-.7354)[250750 - \frac{(1540)^2}{10}]}{102}} = 21.2$$

This is the average amount by which the predicted values of blood-alcohol level derived from the

regression equation differ from the observed or actual levels. Thus, it is the "average error" in predicting the dependent variable.

Note that this value by itself is difficult to interpret, so it is helpful to compare it to the mean of Y (it is in the units of measurement used for the dependent variable). In fact, dividing the standard error of measurement by the mean of Y yields a Coefficient of Variation. This is generated by SAS and is indicated by the abbreviation "Coeff Var". In this case, Coeff Var = 0.138, which indicates that the average (standard) error in predicting Y is 13.8% of its mean.

MULTIVARIATE REGRESSION: ASSUMPTIONS, ESTIMATION, AND THE GAUSS-MARKOV THEOREM

I. Assumptions of the General Linear Model (GLM)

A. The General Linear Model is simply an extension of the Two Variable Linear Model to the case of $k - 1$ independent (X) variables. The population regression function is written as:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{k,i} + u_i \qquad \textbf{(3.40)}$$

B. It is necessary to add an assumption which restricts the independent variables to *linear independence*. This assumption implies that each independent X variable is unique, and cannot be formed from any *linear combination* of the other X variables in the model.

1. If one or more of the X's are linearly dependent, we have a problem called *pure* or *perfect multicollinearity*. For example, in the following equation, X_2 is perfectly collinear with (or linearly dependent on) X_3 and X_4:

$$X_2 = aX_3 + bX_4$$

(where a and b are constants), whereas in the next equation, X_2 is not linearly dependent on these variables, since this combination is multiplicative rather than linear:

$$X_2 = X_3 X_4$$

The presence of perfect multicollinearity results in an inability to derive OLS estimators for the GLM parameters. Intuitively, this assumption implies that if all of the effects of a variable are already included in a model by the presence of other independent variables, its participation in the regression equation is superfluous.

2. Linear independence does not imply statistical independence. Two variables can be linearly independent and still be related statistically. However, statistical independence does imply linear independence. If two variables are statistically independent they will also be linearly independent.

3. This is a very important concept for regression analysis, and means that *the X variables cannot be perfect linear combinations of each other*. But they can be related in a statistical sense. The concept that gives rise to the term *independent variable* is linear independence, not statistical independence.

C. The General Linear Model consists of equation 3.40 (which implies that the regression model is correctly specified) and the following assumptions:

1. E (u_i) = 0
2. Cov ($u_i u_j$) = 0 for $i \neq j$
3. Var (u_i) = σ^2
4. X's are fixed [or Cov ($X_{ji} u_i$) = 0 for j = 2,..., k)
5. X's are linearly independent (note also that k must be less than or equal to n)
6. Correct specification

D. Although the GLM specifies $k - 1$ independent variables, theory concerning the GLM can generally be developed using just a three variable model. This approach will be used whenever feasible. Thus, the general population equation we will be working with is:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \textbf{(3.41)}$$

II. Interpretation of the Regression Function and Parameters

A. Assuming a three variable regression model, a linear relationship between Y, X_2, and X_3 in the population can be represented by a regression plane.

B. β_1 is still the intercept term, representing the point on the Y axis through which the regression plane passes. β_2 represents the slope of the line formed by the intersection of the regression plane and the plane of the Y and X_2 axes. β_3 represents the slope of the line formed by the intersection of the regression plane and the plane formed by the Y and X_3 axes. Since β_2 and β_3 pertain to lines which are obtained by holding the effects of all but one independent variable constant, they are termed *partial regression coefficients*.

C. This interpretation of the β's can be extended to k dimensions with a regression hyperplane. Unfortunately, human minds have difficulty visualizing this, implying a necessity for "faith"!

III. Estimation and the Gauss-Markov Theorem

A. In order to estimate the values of the β's in the population regression function using sample data, we can extend the concept of least squares. The procedure consists of minimizing the sum of the squared deviations of Y from the regression plane (i.e., minimize $\sum \hat{u}_i^2$).

B. The same minimization procedure from calculus used for the Two Variable Model is utilized for the GLM. Set first partial derivatives of the following function equal to zero.

$$\sum u_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})^2 \quad \textbf{(3.42)}$$

1. This gives the <u>three</u> normal equations for the regression plane:

$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3 \quad \textbf{(3.43)}$$

$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} \quad \textbf{(3.44)}$$

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 \quad \textbf{(3.45)}$$

2. Solving these three normal equations simultaneously gives the formulas for the OLS estimators of β_1, β_2, and β_3:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3 \quad \textbf{(3.46)}$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad \textbf{(3.47)}$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad \textbf{(3.48)}$$

3. The value of $\hat{\beta}_2$ is interpreted as the estimated change in the dependent variable *Y* given a one unit change in X_2 ***while holding the effects of X_3 constant***. The value of $\hat{\beta}_3$ is interpreted analogously.

C. The *Gauss-Markov Theorem* states that, given the assumptions of the General Linear Model, the OLS estimators of β_1, β_2, and β_3 will be best linear unbiased estimators (BLUE).

D. *The variance (or standard error[2]) of* $\hat{\beta}_2$ *and* $\hat{\beta}_3$ is:

$$\text{Var}(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \sigma^2 \quad \textbf{(3.49)}$$

$$\text{Var}(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \sigma^2 \quad \textbf{(3.50)}$$

But since we generally do not know σ^2, we estimate it with:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k} \quad \textbf{(3.51)}$$

where:

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i} \quad \textbf{(3.52)}$$

MULTIVARIATE REGRESSION: STRENGTH OF RELATIONSHIP AND HYPOTHESIS TESTING

I. Strength of Relationship

A. R^2 - Multiple Coefficient of Determination or the Multiple Correlation Coefficient

1. Decomposition of the sample variation of *Y* into the RSS and ESS components is identical to that of the Two Variable Linear Model:

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \qquad \textbf{(3.53)}$$

$$\text{TSS} \quad = \quad \text{RSS} \quad + \quad \text{ESS}$$

2. The interpretation of R^2 is also identical to that of r^2 for the Two Variable Model, and

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \qquad \textbf{(3.54)}$$

But for the three variable model,

$$R^2 = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} \qquad \textbf{(3.55)}$$

R^2 also has a range of 0 to 1.

B. Although R^2 can be very useful in the context of evaluating the strength of relationship or "goodness of fit" of a multivariate regression, it has an annoying property which makes it difficult to use in comparing different regression models with the same dependent variable.

1. This property is that the value of R^2 is partly dependent on the number of independent variables in the equation, regardless of their explanatory power. Hence, the value of R^2 has the tendency to increase automatically as additional independent *X* variables are added to the model, thus erroneously indicating an increase in the explanatory power of the equation.

2. This problem can be overcome by adjusting for the number of independent variables in the calculation of R^2. If *k* equals the number of variables (or parameters) in the model, the *adjusted* R^2 (or $\bar{R}^2$) is computed as:

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)} = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)} \qquad \textbf{(3.56)}$$

$$= 1 - (1 - R^2)\frac{(n-1)}{(n-k)}$$

Note that if $R^2 = 0$, then the adjusted $\bar{R}^2 < 0$. If $k > 2$, then the value of the adjusted $\bar{R}^2 < R^2$, this difference growing larger as *k* increases.

3. It is also possible to derive partial correlation coefficients, analogous to partial regression coefficients, for relationships between two variables in which control is exercised for the effects of other variables.

II. Hypothesis Testing

A. Testing for the Significance of a Linear Relationship between the Dependent and Independent Variables (Testing the Significance of the Regression)

1. For purposes of hypothesis testing, it is necessary to again add the assumption of normally distributed errors to the General Linear Model: $u \sim N$. If the error term has a normal distribution, then it is possible to establish the distributions of the respective test statistics.

2. The lack of a linear relationship between Y and the X variables would imply H_0: $\beta_2 = \beta_3 = 0$. This hypothesis can be tested using an Analysis of Variance (ANOVA or AOV) framework and an F-statistic. The appropriate statistic with an F-distribution is:

$$\frac{\text{ESS}/(k-1)}{\text{RSS}/(n-k)} = \frac{(\text{ESS}/\text{TSS})/(k-1)}{(\text{RSS}/\text{TSS})/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,\ n-k}$$

Lind, Marchal, and Wathen (2002) use the following formula:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n-(k+1)]}$$

where SSR is the sum of the squares explained by the regression (or ESS) and SSE is the sum of squares error (or RSS).

Alternatively, for the 3 variable case:

$$\frac{(\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i})/2}{\sum \hat{u}_i^2/(n-3)} \sim F_{2,\ n-3}$$

If the null hypothesis is true, and the partial regression coefficients are equal to zero, then the numerator should not differ markedly from the denominator; i.e., they should both give the same estimate of σ^2. If a relationship does exist between Y and X_2 and X_3, the ESS in the numerator should significantly exceed the RSS in the denominator. Note that this is a test of a *joint hypothesis* concerning both β_2 and β_3, as is described below.

$$F_{k-1,\ n-k} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

3. As indicated above, the relationship between R^2 and the calculated F-statistic is:
   a. As the value of $\overline{R}^2$ increases, so does the value of the F-statistic. When R^2 is 0, F is also zero. When R^2 is 1, F becomes infinite.

   b. The F test, which is a measure of the overall significance of the estimated regression, is also a test of the significance of R^2.

B. Testing for the Significance of Individual Partial Regression Coefficients

1. Given that $u \sim N$, we can test hypotheses concerning specific values of the β's in the General Linear Model using the statistic:

$$\frac{\hat{\beta} - \beta}{se_{\hat{\beta}}} \sim t_{n-k}$$

2. If we specify H_0: $\beta = 0$, we are testing the significance of the specific β value, i.e., we are trying to determine whether the calculated value for β is significantly different from zero. For the three variable model and H_0: $\beta_2 = 0$, this test statistic is:

$$\frac{\hat{\beta}_2 - 0}{\sqrt{\sum x_{3i}^2 [(\sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}) / (n-3)] / [(\sum x_{2i}^2 \sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2]}} \sim t_{n-3}$$

This is obtained by substitution from equations (3.49) (3.51), and (3.52).

3. Note that a test of the hypothesis H_0: $\beta_2 = \beta_3 = 0$ is not the same as the individual tests of H_0: $\beta_2 = 0$ and H_0: $\beta_3 = 0$.

   a. The first case (H_0: $\beta_2 = \beta_3 = 0$) is a test of a joint hypothesis which states that both partial regression coefficients are simultaneously equal to zero, whereas the second case (H_0: $\beta_2 = 0$ and H_0: $\beta_3 = 0$) looks at each β independently.

   b. Since the overall significance of the regression relates to the joint hypothesis, individual tests of the parameters might indicate significant β's, whereas the joint hypothesis test might reveal lack of significance. However, the joint hypothesis test using the *F*-statistic is considerably more powerful than individual *t*-tests.

   c. The overall significance of the regression is first explored using the *F*-statistic before any interpretation is placed on the calculated values of the *t*-statistics.

C. Confidence Intervals: Confidence intervals can be derived for the β's of the General Linear Model in the usual fashion using the *t*-distribution:

$$-t_{\alpha/2} \leq \frac{\hat{\beta} - \beta}{se_{\hat{\beta}}} \leq t_{\alpha/2} \qquad \textbf{(3.57)}$$

$$\hat{\beta} - t_{\alpha/2} se_{\hat{\beta}} \leq \beta \leq \hat{\beta} + t_{\alpha/2} se_{\hat{\beta}} \qquad \textbf{(3.58)}$$

III. Multiple Regression Example

Suppose we wish to determine if there is a relationship between the number of deaths in large (400-500 beds) hospitals per month, hospital size, and equipment investment per bed. We sample 25 hospitals in the 400-500 bed size range and come up with the following results.

Deaths/Month	Beds	$1000/Bed	Deaths/Month	Beds	$1000/Bed
44	410	22.1	47	419	22.5
60	427	23.1	71	431	24.0
61	464	22.6	60	481	21.7
56	467	22.0	66	482	24.6
51	457	21.1	53	448	22.2
74	496	24.8	33	412	20.5
54	447	21.9	52	473	20.8
30	404	20.0	58	409	23.3
59	498	21.3	52	427	22.9
56	459	22.3	49	423	22.6
63	490	22.4	61	434	23.8
39	416	20.6	62	432	24.4
78	494	25.0			

$n = 25$

A. Derive the Regression Equation:

$$\sum Y = 1389 \qquad \sum X_2 = 11200 \qquad \sum X_3 = 562.5$$

$$\overline{Y} = 55.56 \qquad \overline{X}_2 = 448.00 \qquad \overline{X}_3 = 22.50$$

$$\sum y^2 = 3146.16 \qquad \sum x_2^2 = 22684 \qquad \sum x_3^2 = 46.38$$

$$\sum x_2 y = 5638 \qquad \sum x_3 y = 316.5 \qquad \sum x_2 x_3 = 270.8$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3$$

$$\hat{\beta}_2 = \frac{\sum x_2 y \sum x_3^2 - \sum x_2 x_3 \sum x_3 y}{\sum x_2^2 \sum x_3^2 - (\sum x_2 x_3)^2}$$

$$\hat{\beta}_3 = \frac{\sum x_3 y \sum x_2^2 - \sum x_2 x_3 \sum x_2 y}{\sum x_2^2 \sum x_3^2 - (\sum x_2 x_3)^2}$$

$$\hat{\beta}_2 = \frac{(5638)(46.38) - (270.8)(316.5)}{(22684)(46.38) - (270.8)^2} = 0.179598$$

$$\hat{\beta}_3 = \frac{(316.5)(22684) - (270.8)(5638)}{(22684)(46.38) - (270.8)^2} = 5.775436$$

$$\hat{\beta}_1 = 55.56 - .179598(448.00) - 5.775436(22.50) = -154.847214$$

$$\hat{Y}_i = -154.847 + 0.1796X_{2i} + 5.7754X_{3i}$$

B. Compute the R^2:

$$R^2 = \frac{\hat{\beta}_2 \sum x_2 y + \hat{\beta}_3 \sum x_3 y}{\sum y^2} \qquad \overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

$$R^2 = \frac{0.179598(5638) + 5.775436(316.5)}{3146.16} = .9028$$

$$R^2 = .9028 \qquad \overline{R}^2 = 1 - (1 - .9028)\frac{24}{22} = .8940$$

$\overline{R}^2$ = 0.8940 => ~ 89% of the variation in the dependent variable is explained by the independent variables

C. Test the Significance of the Regression (i.e., test H_0: $\beta_2 = \beta_3 = 0$):

$$\frac{R^2/2}{(1-R^2)/(n-3)} \sim F_{2,\,n-3} \qquad \text{let } \alpha = .01$$

$$\frac{.9028/2}{(1-.9028)/(25-3)} = \hat{F}$$

$$\hat{F} = 102.169 \; > \; F^*_{2,\,22} = 5.72$$

D. Test the Significance of the Individual Parameter Estimates for β_2 and β_3:

$$\frac{\hat{\beta}_2 - \beta_2}{se_{\hat{\beta}_2}} \sim t_{n-k} \qquad t^*_{22} = 2.819 \qquad \alpha = .01$$

$$se_{\hat{\beta}_2} = \sqrt{\frac{\sum x_{3i}^2 [\sum \hat{u}_i^2/(n-3)]}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}} \qquad \sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}$$

$$se_{\hat{\beta}_2} = \sqrt{\frac{(46.38)[(305.6611)/22)]}{(22684)(46.38)-(270.8)^2}} = 0.0257$$

$$\hat{t}_{\beta_2} = \frac{0.1796 - 0}{0.0257} = 6.988 \; > \; t^*_{22} = 2.819$$

$$\frac{\hat{\beta}_3 - \beta_3}{se_{\hat{\beta}_3}} \sim t_{n-k} \qquad t^*_{22} = 2.819 \qquad \alpha = .01$$

$$\hat{t}_{\beta_3} = \frac{5.7754 - 0}{0.5675} = 10.177 \; > \; t^*_{22} = 2.819$$

IV. Implicitly Linear Models and Variable Transformations

A. Nonlinear Models - Nonlinear regression models may be used in an OLS framework provided that the model is nonlinear in the variables but linear in the parameters (or can be transformed so it is linear in the parameters). Such a model is said to be *implicitly* or *intrinsically linear*. Nonlinear forms should only be utilized if theory so specifies, or theory does not specify a functional form, and there is reason to believe that a nonlinear form may be an appropriate specification. The major characteristic of such a model is that, with an appropriate data transformation, it can be converted into a linear model. For example:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3^2 + \beta_4 X_2 X_3 + u$$

can be transformed via:

$$Z_3 = X_3^2 \quad \text{and} \quad Z_4 = X_2 X_3$$

to the linear form:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 Z_3 + \beta_4 Z_4 + u$$

or:

$$Y = \beta_1 X_2^{\beta_2} X_3^{\beta_3} u$$

can be transformed via logarithmic transformations to:

$$\log Y = \log \beta_1 + \beta_2 \log X_2 + \beta_3 \log X_3 + \log u$$

Note that for such a transformation, *u* must have a "log normal" distribution.

B. A common type of nonlinear model is the semi-log function:

$$Y = \beta_1 + \beta_2 \log X + u$$

which can be transformed to the linear equation:

$$Y = \beta_1 + \beta_2 Z + u \quad \text{where } Z = \log X$$

Double-log functions of the following type are also fairly common:

$$\log Y = \beta_1 + \beta_2 \log X + u$$

But note that this form cannot be compared directly to the equation $Y = \beta_1 + \beta_2 X + u$, since the dependent variables have different forms.

SPEA V506
INTERVAL ESTIMATION AND HYPOTHESIS TESTING SUMMARY
(Excludes ANOVA and Regression)

PURPOSE	STATISTIC	DISTRIBUTION	COMMENTS[5]
Interval estimate or hypothesis test for μ when σ is known	z	Normal	
Interval estimate or hypothesis test for μ when σ is unknown	t	t ($n-1$ df) for $n \le 120$, normal for $n > 120$	
Interval estimate or hypothesis test for independent samples difference of means, σ is known	z	Normal	
Interval estimate or hypothesis test for independent samples difference of means, unknown σ but assume $\sigma_1^2 = \sigma_2^2$	pooled t	t ($n-1$ df) for $n \le 120$, normal for $n > 120$	Tradeoff variance and sample size differences
Interval estimate or hypothesis test for independent samples difference of means, unknown σ but cannot assume $\sigma_1^2 = \sigma_2^2$	t′	t ($n-1$ df) for $n \le 120$, normal for $n > 120$	Use when pooled t is inappropriate due to violations of assumptions
Interval estimate or hypothesis test for paired samples difference of means, unknown σ	$d \sim t$	t ($n-1$ df) for $n \le 120$, normal for $n > 120$	
Interval estimate or hypothesis test for population variance or standard deviation	χ^2	χ^2 ($n-1$ df)	Distribution is not symmetric. Use conversion formula for $n > 40$
Interval estimate or hypothesis test for independent samples difference between two variances	F	F (n_1-1, n_2-1 df)	Distribution is not symmetric. Place larger variance in numerator, allowing for right-tail values to always be used
Interval estimate or hypothesis test for a binomial proportion	z	Normal	
Hypothesis test for a proportion	P's X^2	χ^2 ($k-1$ df)	Only 2-sided H_1 possible. Do not divide α by 2
Hypothesis test for a contingency table	P's X^2	χ^2 [$(k_A-1)(k_B-1)$ df]	Only 2-sided H_1 possible. Do not divide α by 2

5. Unless otherwise noted, divide α by 2 for two-tailed tests or intervals. All statistics except z with $n > 30$ require assumption of normal populations.