<table>
<thead>
<tr>
<th>الموضوع</th>
<th>اليوم</th>
<th>الوقت</th>
</tr>
</thead>
<tbody>
<tr>
<td>مبادئ ومبادئ الإحصاء</td>
<td>السبت</td>
<td>10:15</td>
</tr>
<tr>
<td>استراحة</td>
<td></td>
<td>10:30</td>
</tr>
<tr>
<td>أنواع العينات، مهارات الاستدلال الإحصائي</td>
<td></td>
<td>12:00</td>
</tr>
<tr>
<td>بعض التوزيعات الاحتمالية، التقدير</td>
<td>الأحد</td>
<td>10:15</td>
</tr>
<tr>
<td>استراحة</td>
<td></td>
<td>10:30</td>
</tr>
<tr>
<td>اختيارات الفروض</td>
<td></td>
<td>12:00</td>
</tr>
<tr>
<td>اختيارات مربع كاي</td>
<td></td>
<td>10:15</td>
</tr>
<tr>
<td>استراحة</td>
<td></td>
<td>10:30</td>
</tr>
<tr>
<td>الارتباط والانحدار البسيط</td>
<td></td>
<td>12:00</td>
</tr>
<tr>
<td>تحليل التباين في اتجاه واحد، في اتجاهين</td>
<td></td>
<td>10:15</td>
</tr>
<tr>
<td>استراحة</td>
<td></td>
<td>10:30</td>
</tr>
<tr>
<td>تطبيقات</td>
<td></td>
<td>12:00</td>
</tr>
<tr>
<td>الاختيارات اللامعنية</td>
<td></td>
<td>10:15</td>
</tr>
<tr>
<td>استراحة</td>
<td></td>
<td>10:30</td>
</tr>
<tr>
<td>إعداد وكتابة التقارير الإحصائية</td>
<td></td>
<td>12:00</td>
</tr>
</tbody>
</table>
Contents:

1. Introduction to Statistics
2. Describing Univariate Data
3. Describing Bivariate Data
4. Introduction to Probability (elementary)
5. Normal Distribution
6. Sampling Distributions
7. Point Estimation
8. Confidence Intervals
9. The Logic of Hypothesis Testing
10. Testing Hypotheses with Standard Errors
11. Power
12. Introduction to Between-Subjects ANOVA
13. Factorial Between-Subjects ANOVA
14. Within-Subjects/Repeated Measures ANOVA
15. Prediction
16. Chi Square
17. Distribution-Free Tests
18. Measuring Effect Size
CHAPTER 1

Introduction to Statistics
Descriptive Statistics

One important use of descriptive statistics is to summarize a collection of data in a clear and understandable way. For example, assume a psychologist gave a personality test measuring shyness to all 2500 students attending a small college. How might these measurements be summarized? There are two basic methods: numerical and graphical. Using the numerical approach one might compute statistics such as the mean and standard deviation. These statistics convey information about the average degree of shyness and the degree to which people differ in shyness. Using the graphical approach one might create a stem and leaf display and a box plot. These plots contain detailed information about the distribution of shyness scores.

Graphical methods are better suited than numerical methods for identifying patterns in the data. Numerical approaches are more precise and objective.

Since the numerical and graphical approaches compliment each other, it is wise to use both.

Inferential Statistics

Inferential statistics are used to draw inferences about a population from a sample. Consider an experiment in which 10 subjects who performed a task after 24 hours of sleep deprivation scored 12 points lower than 10 subjects who performed after a normal night's sleep. Is the difference real or could it be due to chance? How much larger could the real difference be than the 12 points found in the sample? These are the types of questions answered by inferential statistics.

There are two main methods used in inferential statistics: estimation and hypothesis testing. In estimation, the sample is used to estimate a parameter and a confidence interval about the estimate is constructed.

In the most common use of hypothesis testing, a "straw man" null hypothesis is put forward and it is determined whether the data are strong enough to reject it. For the sleep deprivation study, the null hypothesis would be that sleep deprivation has no effect on performance.

Variable

A variable is any measured characteristic or attribute that differs for different subjects. For example, if the weight of 30 subjects were measured, then weight would be a variable.

Quantitative and Qualitative

Variables can be quantitative or qualitative. (Qualitative variables are sometimes called "categorical variables.") Quantitative variables are measured on an ordinal, interval, or ratio scale; qualitative variables are measured on a nominal scale. If five-year old subjects were asked to name their favorite color, then the variable would be qualitative. If the time it took them to respond were measured, then the variable would be quantitative.

Independent and Dependent

When an experiment is conducted, some variables are manipulated by the experimenter and others are measured from the subjects. The former variables are called "independent variables" or "factors" whereas the latter are called "dependent variables" or "dependent measures."

For example, consider a hypothetical experiment on the effect of drinking alcohol on reaction time: Subjects drank either water, one beer, three beers, or six beers and then had their reaction times to the onset of a stimulus measured. The independent variable would be the number of beers drunk (0, 1, 3, or 6) and the dependent variable would be reaction time.

Continuous and Discrete
Some variables (such as reaction time) are measured on a continuous scale. There is an infinite number of possible values these variables can take on. Other variables can only take on a limited number of values. For example, if a dependent variable were a subject's rating on a five-point scale where only the values 1, 2, 3, 4, and 5 were allowed, then only five possible values could occur. Such variables are called "discrete" variables.

**Parameter**

A parameter is a numerical quantity measuring some aspect of a population of scores. For example, the mean is a measure of central tendency.

Greek letters are used to designate parameters. At the bottom of this page are shown several parameters of great importance in statistical analyses and the Greek symbol that represents each one. Parameters are rarely known and are usually estimated by statistics computed in samples. To the right of each Greek symbol is the symbol for the associated statistic used to estimate it from a sample.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Parameter</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>μ</td>
<td>M</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>σ</td>
<td>s</td>
</tr>
<tr>
<td>Proportion</td>
<td>π</td>
<td>p</td>
</tr>
<tr>
<td>Correlation</td>
<td>ρ</td>
<td>r</td>
</tr>
</tbody>
</table>

**Statistics**

The word "statistics" is used in several different senses. In the broadest sense, "statistics" refers to a range of techniques and procedures for analyzing data, interpreting data, displaying data, and making decisions based on data. This is what courses in "statistics" generally cover.

In a second usage, a "statistic" is defined as a numerical quantity (such as the mean) calculated in a sample. Such statistics are used to estimate parameters.

The term "statistics" sometimes refers to calculated quantities regardless of whether or not they are from a sample. For example, one might ask about a baseball player's statistics and be referring to his or her batting average, runs batted in, number of home runs, etc. Or, "government statistics" can refer to any numerical indexes calculated by a governmental agency.

although the different meanings of "statistics" has the potential for confusion, a careful consideration of the context in which the word is used should make its intended meaning clear.

**Summation Notation**
The Greek letter Σ (a capital sigma) is used to designate summation. For example, suppose an experimenter measured the performance of four subjects on a memory task. Subject 1’s score will be referred to as $X_1$, Subject 2’s as $X_2$, and so on. The scores are shown below:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$X_1$</td>
</tr>
<tr>
<td>2</td>
<td>$X_2$</td>
</tr>
<tr>
<td>3</td>
<td>$X_3$</td>
</tr>
<tr>
<td>4</td>
<td>$X_4$</td>
</tr>
</tbody>
</table>

The way to use the summation sign to indicate the sum of all four $X$'s is:

$$
\sum_{i=1}^{4} X_i
$$

This notation is read as follows:
Sum the values of $X$ from $X_1$ through $X_4$.

The index $i$ (shown just under the Σ sign) indicates which values of $X$ are to be summed. The index $i$ takes on values beginning with the value to the right of the "=" sign (1 in this case) and continues sequentially until it reaches the value above the Σ sign (4 in this case). Therefore $i$ takes on the values 1, 2, 3, and 4 and the values of $X_1$, $X_2$, $X_3$, and $X_4$ are summed ($7 + 6 + 5 + 8 = 26$).

In order to make formulas more general, variables can be used with the summation notation. For example,

$$
\sum_{i=1}^{N} X_i
$$

means to sum up values of $X$ from 1 to $N$ where $N$ can be any number but usually indicates the sample size.

Often an abbreviated form of the summation notation is used. For example, $\Sigma X$ means to sum all the values of $X$. When only a subset of the values of $X$ are to be summed then the full version is required. Thus, the sum of all elements of $X$ except the first and the last (the N'th) would be indicated as:

$$
\sum_{i=2}^{N-1} X_i
$$

which would be read as the sum of $X$ with $i$ going from 2 to N-1. Some formulas require that each number be squared before the numbers are summed. This is indicated by:
and is equal to $7^2 + 6^2 + 5^2 + 8^2 = 174$.

The abbreviated version is simply: $\Sigma X^2$. It is very important to note that it makes a big difference whether the numbers are squared first and then summed or summed first and then squared. The symbol $(\Sigma X)^2$ indicates that the numbers should be summed first and then squared. For the present example, this equals:

$$(7 + 6 + 5 + 8)^2 = 26^2 = 676.$$ This, of course, is quite different from 174.

Sometimes a formula requires that the sum of cross products be computed. For instance, if 3 subjects were each tested twice, they might each have a score on $X$ and on $Y$.

<table>
<thead>
<tr>
<th>Subject</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

The sum of cross products $(2 \times 3) + (1 \times 6) + (4 \times 5) = 32$ can be represented in summation notation simply as: $\Sigma XY$

**Basic Theorems:**

The following data will be used to illustrate the theorems:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

$\Sigma (X + Y) = \Sigma X + \Sigma Y$

$\Sigma (X + Y) = 11 + 5 + 5 = 21$

$\Sigma X = 3 + 2 + 4 = 9$

$\Sigma Y = 8 + 3 + 1 = 12$

$\Sigma X + \Sigma Y = 9 + 12 = 21$

$\Sigma aX = a\Sigma X$ (a is a constant)

For an example, let $a = 2$.

$\Sigma aX = (2)(3) + (2)(2) + (2)(4) = 18$
\[ \Sigma X = (2)(9) = 18 \]

\[ \Sigma(X-M)^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \]

(N is the number of numbers, 3 in this case, and M is the mean which is also equal to 3 in this case.

\[ \Sigma(X-M)^2 = (3-3)^2 + (2-3)^2 + (4-3)^2 = 2 \]

\[ \Sigma X^2 = 3^2 + 2^2 + 4^2 = 29 \]

\[ \frac{(\Sigma X)^2}{N} = 9^2/3 = 27 \]

\[ \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 29 - 27 = 2 \]

**Measurement Scales**

Measurement is the assignment of numbers to objects or events in a systematic fashion. Four levels of measurement scales are commonly distinguished: nominal, ordinal, interval, and ratio.

There is a relationship between the level of measurement and the appropriateness of various statistical procedures. For example, it would be silly to compute the mean of nominal measurements. However, the appropriateness of statistical analyses involving means for ordinal level data has been controversial. One position is that data must be measured on an interval or a ratio scale for the computation of means and other statistics to be valid. Therefore, if data are measured on an ordinal scale, the median but not the mean can serve as a measure of central tendency.

The arguments on both sides of this issue will be examined in the context of an hypothetical experiment designed to determine whether people prefer to work with color or with black and white computer displays. Twenty subjects viewed black and white displays and 20 subjects viewed color displays.

Displays were rated on a 7 point scale where a 1 was the lowest rating and a 7 was the highest rating. This rating scale is only an ordinal scale since there is no assurance that the difference between a rating of 1 and a rating of 2 represents the same degree of difference in preference as the difference between a rating of 5 and a rating of 6.

The mean rating of the color display was 5.5 and the mean rating of the black and white display was 3.9. The first question the experimenter would ask is how likely is it that this big a difference between means could have occurred just because of chance factors such as which subjects saw the black and white display and which subjects saw the color display. Standard methods of statistical inference can answer this question. Assume these methods led to the conclusion that the difference was not due to chance but represented a "real" difference in means. Does the fact that the rating scale was ordinal instead of interval have any implications for the validity of the statistical conclusion that the difference between means was not due to chance?

The answer is an unequivocal "NO." There is really no room for argument here. What can be questioned, however, is whether it is worth knowing that the mean rating of color displays is higher than the mean rating for B & W displays.

The argument that it is not worth knowing assumes that means of ordinal data are meaningless. Supporting the notion that means of ordinal data are meaningless is the fact that examples can be made up showing that a difference between means on an ordinal scale can be in the opposite direction of what they would have been if the "true" measurement scale had been used.

If means of ordinal data are meaningless, why should anyone care whether the difference between two meaningless quantities (the two means) is due to chance or not. Naturally enough, the answer lies in challenging the proposition that means of ordinal data are meaningless. There are two counter arguments to the example showing that using an ordinal scale can reverse the direction of the difference between means.
The first is philosophical and challenges the validity of the notion that there is some unseen "true" measurement scale that is only being approximated by the rating scale. The second counter argument accepts the notion of an underlying scale but considers the examples to be very contrived and unlikely to occur in real data. Measurement scales used in behavioral research are invariably somewhere between ordinal and interval scales. In the preference experiment, it may not be the case that the difference between the ratings one and two is exactly the same as the difference between five and six, but it is unlikely to be many times larger either. The scale is roughly interval and it is exceedingly unlikely that the means on this scale would favor color displays while the means on the "true" scale would favor the B & W displays.

There are some cases where one can validly argue that the use of an ordinal instead of a ratio scale seriously distorts the conclusions. Consider an experiment designed to determine whether 5-year old children are more distractible than 10-year old children.

<table>
<thead>
<tr>
<th></th>
<th>No Distraction</th>
<th>Distraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-yr</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>10-yr</td>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>

It looks as though the 10-year olds are more distractible since distraction cost them 4 points but only cost the 5-year olds 3 points. However, it might be that a change from 3 to 6 represents a larger difference than a change from 8 to 12. Consider that the performance of 5-year olds dropped 50% from distraction but the performance of 10-year olds dropped only 33%.

Which age group is "really" more distractible? Unfortunately, there is no clearly right or wrong answer. If proportional change is considered, then 5-year olds are more distractible; if the amount of change is considered then 10-year olds are more distractible. Keep in mind that statistical conclusions are not affected by the choice of measurement scale even though the all-important interpretation of these conclusions can be.

In this example, a statistical test could validly rule out chance as an explanation of the finding that 10-year olds lost more points from distraction than did 5-year olds. However, the statistical test will not reveal whether a greater drop necessarily means 10-year olds are more distractible. So, the conclusion that distraction costs 10-year olds more points than it costs 5-year olds is valid. The interpretation depends on measurement issues.

In summary, statistical analyses provide conclusions about the numbers entered into them. Relating these conclusions to the substantive research issues depends on the measurement operations.
CHAPTER 2

Describing Univariate Data
Mean

Arithmetic Mean

The arithmetic mean is what is commonly called the average: When the word "mean" is used without a modifier, it can be assumed that it refers to the arithmetic mean. The mean is the sum of all the scores divided by the number of scores. The formula in summation notation is:

$$\mu = \frac{\sum X}{N}$$

where $\mu$ is the population mean and $N$ is the number of scores.

If the scores are from a sample, then the symbol $M$ refers to the mean and $N$ refers to the sample size. The formula for $M$ is the same as the formula for $\mu$.

$$M = \frac{\sum X}{N}$$

The mean is a good measure of central tendency for roughly symmetric distributions but can be misleading in skewed distributions since it can be greatly influenced by scores in the tail. Therefore, other statistics such as the median may be more informative for distributions such as reaction time or family income that are frequently very skewed.

Click here for an interactive demonstration of properties of the mean and median.

The sum of squared deviations of scores from their mean is lower than their squared deviations from any other number.

For normal distributions, the mean is the most efficient and therefore the least subject to sample fluctuations of all measures of central tendency.

The formal definition of the arithmetic mean is $\mu = E[X]$ where $\mu$ is the population mean of the variable $X$ and $E[X]$ is the expected value of $X$.

The geometric mean is the nth root of the product of the scores. Thus, the geometric mean of the scores: 1, 2, 3, and 10 is the fourth root of $1 \times 2 \times 3 \times 10$ which is the fourth root of 60 which equals 2.78. The formula can be written as:

Geometric mean $= \left( \prod X \right)^{\frac{1}{N}}$

where $\prod X$ means to take the product of all the values of $X$. The geometric mean can also be computed by:

1. taking the logarithm of each number
2. computing the arithmetic mean of the logarithms
3. raising the base used to take the logarithms to the arithmetic mean.

The next page shows an example of this method using natural logarithms.
The base of natural logarithms is 2.718. The expression: \( \text{EXP}[1.024] \) means that 2.718 is raised to the 1.024th power. \( \ln(X) \) is the natural log of \( X \).

Naturally, you get the same result using logs base 10 as shown below.

If any one of the scores is zero then the geometric mean is zero. The geometric mean does not make sense if any scores are less than zero.

The geometric mean is less affected by extreme values than is the arithmetic mean and is useful as a measure of central tendency for some positively skewed distributions.

The geometric mean is an appropriate measure to use for averaging rates. For example, consider a stock portfolio
that began with a value of $1,000 and had annual returns of 13%, 22%, 12%, -5%, and -13%. The table below shows the value after each of the five years.

<table>
<thead>
<tr>
<th>Year</th>
<th>Return</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13%</td>
<td>1,130</td>
</tr>
<tr>
<td>2</td>
<td>22%</td>
<td>1,379</td>
</tr>
<tr>
<td>3</td>
<td>12%</td>
<td>1,544</td>
</tr>
<tr>
<td>4</td>
<td>-5%</td>
<td>1,467</td>
</tr>
<tr>
<td>5</td>
<td>-13%</td>
<td>1,276</td>
</tr>
</tbody>
</table>

The question is how to compute annual rate of return? The answer is to compute the geometric mean of the returns. Instead of using the percents, each return is represented as a multiplier indicating how much higher the value is after the year. This multiplier is 1.13 for a 13% return and 0.95 for a 5% loss. The multipliers for this example are 1.13, 1.22, 1.12, 0.95, and 0.87. The geometric mean of these multipliers is 1.05. Therefore, the average annual rate of return is 5%. The following table shows how a portfolio gaining 5% a year would end up with the same value ($1,276) as the one shown above.

<table>
<thead>
<tr>
<th>Year</th>
<th>Return</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5%</td>
<td>1,050</td>
</tr>
<tr>
<td>2</td>
<td>5%</td>
<td>1,103</td>
</tr>
<tr>
<td>3</td>
<td>5%</td>
<td>1,158</td>
</tr>
<tr>
<td>4</td>
<td>5%</td>
<td>1,216</td>
</tr>
<tr>
<td>5</td>
<td>5%</td>
<td>1,276</td>
</tr>
</tbody>
</table>

**Harmonic Mean**

The harmonic mean is used to take the mean of sample sizes. If there are $k$ samples each of size $n$, then the harmonic mean is defined as:

$$ n_h = \frac{k}{\frac{1}{n_1} + \frac{1}{n_2} + \ldots + \frac{1}{n_k}} $$
For the numbers 1, 2, 3, and 10, the harmonic mean is:

$$n_h = \frac{4}{\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{10}} = 2.069.$$ 

= 2.069. This is less than the geometric mean of 2.78 and the arithmetic mean of 4.

**Median**

The median is the middle of a distribution: half the scores are above the median and half are below the median. The median is less sensitive to extreme scores than the mean and this makes it a better measure than the mean for highly skewed distributions. The median income is usually more informative than the mean income, for example.

The sum of the absolute deviations of each number from the median is lower than is the sum of absolute deviations from any other number. Click here for an example.

The mean, median, and mode are equal in symmetric distributions. The mean is typically higher than the median in positively skewed distributions and lower than the median in negatively skewed distributions, although this may not be the case in bimodal distributions. Click here for examples.

**Computation of Median**

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4.

When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is \((4+7)/2 = 5.5\).

The mode is the most frequently occurring score in a distribution and is used as a measure of central tendency. The advantage of the mode as a measure of central tendency is that its meaning is obvious. Further, it is the only measure of central tendency that can be used with nominal data.

The mode is greatly subject to sample fluctuations and is therefore not recommended to be used as the only measure of central tendency. A further disadvantage of the mode is that many distributions have more than one mode. These distributions are called "multi modal."

In a normal distribution, the mean, median, and mode are identical.

**Trimean**

The trimean is computed by adding the 25th percentile plus twice the 50th percentile plus the 75th percentile and dividing by four. What follows is an example of how to compute the trimean. The 25th, 50th, and 75th percentile of the dataset "Example 1" are 51, 55, and 63 respectively. Therefore, the trimean is computed as:

$$\frac{51 + (2)(55) + 63}{4} = 56.$$ 

The trimean is almost as resistant to extreme scores as the median and is less subject to sampling fluctuations than
the arithmetic mean in extremely skewed distributions. It is less efficient than the mean for normal distributions. The trimean is a good measure of central tendency and is probably not used as much as it should be.

**Trimmed Mean**

A trimmed mean is calculated by discarding a certain percentage of the lowest and the highest scores and then computing the mean of the remaining scores. For example, a mean trimmed 50% is computed by discarding the lower and higher 25% of the scores and taking the mean of the remaining scores.

The median is the mean trimmed 100% and the arithmetic mean is the mean trimmed 0%.

A trimmed mean is obviously less susceptible to the effects of extreme scores than is the arithmetic mean. It is therefore less susceptible to sampling fluctuation than the mean for extremely skewed distributions. It is less efficient than the mean for normal distributions.

Trimmed means are often used in Olympic scoring to minimize the effects of extreme ratings possibly caused by biased judges.

**Summary of Measures of Central Tendency**

Of the five measures of central tendency discussed, the mean is by far the most widely used. It takes every score into account, is the most efficient measure of central tendency for normal distributions and is mathematically tractable making it possible for statisticians to develop statistical procedures for drawing inferences about means. On the other hand, the mean is not appropriate for highly skewed distributions and is less efficient than other measures of central tendency when extreme scores are possible. The geometric mean is a viable alternative if all the scores are positive and the distribution has a positive skew.

The median is useful because its meaning is clear and it is more efficient than the mean in highly-skewed distributions. However, it ignores many scores and is generally less efficient than the mean, the trimean, and trimmed means.

The mode can be informative but should almost never be used as the only measure of central tendency since it is highly susceptible to sampling fluctuations.

Click here for an interactive demonstration of properties of the mean and median.

**Summary of Measures of Central Tendency**

The trimean and trimmed means are both examples of statistics developed to resist sampling fluctuations. It is highly recommended that at least one of these two be computed in addition to the mean.

**Range**

The range is the simplest measure of spread or dispersion: It is equal to the difference between the largest and the smallest values. The range can be a useful measure of spread because it is so easily understood. However, it is very sensitive to extreme scores since it is based on only two values. The range should almost never be used as the only measure of spread, but can be informative if used as a supplement to other measures of spread such as the standard deviation or semi-interquartile range.
Example:
The range of the numbers 1, 2, 4, 6, 12, 15, 19, 26 = 26 -1 = 25

**Semi-Interquartile Range**

The semi-interquartile range is a measure of spread or dispersion. It is computed as one half the difference between the 75th percentile [often called (Q3)] and the 25th percentile (Q1). The formula for semi-interquartile range is therefore: \( \frac{Q3 - Q1}{2} \).

Since half the scores in a distribution lie between Q3 and Q1, the semi-interquartile range is 1/2 the distance needed to cover 1/2 the scores. In a symmetric distribution, an interval stretching from one semi-interquartile range below the median to one semi-interquartile above the median will contain 1/2 of the scores. This will not be true for a skewed distribution, however.

The semi-interquartile range is little affected by extreme scores, so it is a good measure of spread for skewed distributions. However, it is more subject to sampling fluctuation in normal distributions than is the standard deviation and therefore not often used for data that are approximately normally distributed.

**Standard Deviation and Variance**

The variance and the closely-related standard deviation are measures of how spread out a distribution is. In other words, they are measures of variability.

The variance is computed as the average squared deviation of each number from its mean. For example, for the numbers 1, 2, and 3, the mean is 2 and the variance is:

\[
\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = 0.667
\]

The formula (in summation notation) for the variance in a population is

\[
\sigma^2 = \frac{\sum(X - \mu)^2}{N}
\]

where \( \mu \) is the mean and \( N \) is the number of scores.

When the variance is computed in a sample, the statistic

\[
S^2 = \frac{\sum(X - M)^2}{N}
\]

(where \( M \) is the mean of the sample) can be used. \( S^2 \) is a biased estimate of \( \sigma^2 \), however. By far the most common formula for computing variance in a sample is:

\[
S^2 = \frac{\sum(X - M)^2}{N - 1}
\]

which gives an unbiased estimate of \( \sigma^2 \). Since samples are usually used to estimate parameters, \( S^2 \) is the most
commonly used measure of variance. Calculating the variance is an important part of many statistical applications and analyses. It is the first step in calculating the standard deviation.

**Standard Deviation**

The standard deviation formula is very simple: it is the square root of the variance. It is the most commonly used measure of spread.

An important attribute of the standard deviation as a measure of spread is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score. In a normal distribution, about 68% of the scores are within one standard deviation of the mean and about 95% of the scores are within two standard deviations of the mean.

The standard deviation has proven to be an extremely useful measure of spread in part because it is mathematically tractable. Many formulas in inferential statistics use the standard deviation.

(See next page for applications to risk analysis and stock portfolio volatility.)

**Computing the Standard Deviation in SPSS**

The standard deviation is a measure of variability. In SPSS, you compute it by choosing Analyze/Descriptive Statistics/Descriptives...

You then specify the variables you want for which you want to compute the standard deviation:
Here is the result. Note that "Std. Deviation" is used to stand for "standard deviation."

<table>
<thead>
<tr>
<th>Y1</th>
<th>Y2</th>
<th>Y3</th>
<th>var</th>
<th>var</th>
<th>var</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.00</td>
<td>8.50</td>
<td>11.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6.00</td>
<td>9.50</td>
<td>12.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.00</td>
<td>7.00</td>
<td>10.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7.00</td>
<td>6.50</td>
<td>5.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.00</td>
<td>5.75</td>
<td>6.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here is the result. Note that "Std. Deviation" is used to stand for "standard deviation."

### Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>5</td>
<td>3.00</td>
<td>7.00</td>
<td>5.2000</td>
<td>1.48324</td>
</tr>
<tr>
<td>Y2</td>
<td>5</td>
<td>5.75</td>
<td>9.50</td>
<td>7.4500</td>
<td>1.52480</td>
</tr>
<tr>
<td>Y3</td>
<td>5</td>
<td>5.00</td>
<td>12.00</td>
<td>8.8000</td>
<td>3.11448</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notice that, by default, you get N, the minimum, the maximum, and the mean in addition to the standard deviation. You could have chosen more or fewer statistics by clicking the "option" button.

The syntax for computing the standard deviation is:

```
DESCRIPTIVES VARIABLES=Y1 Y2 Y3
/STATISTICS=MEAN STDDEV MIN MAX.
```

### Learn More About the Standard Deviation

Overview of the [standard deviation](http://davidmlane.com/hyperstat/index.html)

[Standard deviation](http://davidmlane.com/hyperstat/index.html) as a measure of risk
Standard Deviation and Variance

although less sensitive to extreme scores than the range, the standard deviation is more sensitive than the semi-interquartile range. Thus, the standard deviation should be supplemented by the semi-interquartile range when the possibility of extreme scores is present.

If variable Y is a linear transformation of X such that:

\[ Y = bX + A, \]

then the variance of Y is:

\[ \sigma_y^2 = b^2 \sigma_x^2 \]

where \( \sigma_x^2 \) is the variance of X.

The standard deviation of Y is \( b \sigma_x \) where \( \sigma_x \) is the standard deviation of X.

Standard Deviation as a Measure of Risk

The standard deviation is often used by investors to measure the risk of a stock or a stock portfolio. The basic idea is that the standard deviation is a measure of volatility: the more a stock’s returns vary from the stock’s average return, the more volatile the stock. Consider the following two stock portfolios and their respective returns (in per cent) over the last six months. Both portfolios end up increasing in value from $1,000 to $1,058. However, they clearly differ in volatility. Portfolio A’s monthly returns range from -1.5% to 3% whereas Portfolio B’s range from -9% to 12%. The standard deviation of the returns is a better measure of volatility than the range because it takes all the values into account. The standard deviation of the six returns for Portfolio A is 1.52; for Portfolio B it is 7.24.

<table>
<thead>
<tr>
<th>A</th>
<th>Value</th>
<th>Return (%)</th>
<th>Final Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>0.75</td>
<td>1,008</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>1.00</td>
<td>1,010</td>
<td></td>
</tr>
<tr>
<td>1,018</td>
<td>3.00</td>
<td>1,048</td>
<td></td>
</tr>
<tr>
<td>1,040</td>
<td>-1.50</td>
<td>1,032</td>
<td></td>
</tr>
<tr>
<td>1,032</td>
<td>0.50</td>
<td>1,038</td>
<td></td>
</tr>
<tr>
<td>1,088</td>
<td>2.00</td>
<td>1,058</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>Value</th>
<th>Return (%)</th>
<th>Final Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>1.50</td>
<td>1,015</td>
<td></td>
</tr>
<tr>
<td>1,015</td>
<td>5.00</td>
<td>1,066</td>
<td></td>
</tr>
<tr>
<td>1,046</td>
<td>12.00</td>
<td>1,194</td>
<td></td>
</tr>
<tr>
<td>1,194</td>
<td>-9.00</td>
<td>1,006</td>
<td></td>
</tr>
<tr>
<td>1,086</td>
<td>-4.00</td>
<td>1,043</td>
<td></td>
</tr>
<tr>
<td>1,043</td>
<td>1.50</td>
<td>1,058</td>
<td></td>
</tr>
</tbody>
</table>

How to compute the standard deviation in SPSS.

Summary of Measures of Spread (Variability)

The standard deviation is by far the most widely used measure of spread. It takes every score into account, has extremely useful properties when used with a normal distribution, and is tractable mathematically and, therefore, it appears in many formulas in inferential statistics. The standard deviation is not a good measure of spread in highly-skewed distributions and should be supplemented in those cases by the semi-interquartile range. The range is a useful statistic, but it cannot stand alone as a measure of spread since it takes into account only two
The semi-interquartile range is rarely used as a measure of spread, in part because it is not very mathematically tractable. However, it is influenced less by extreme scores than the standard deviation, is less subject to sampling fluctuations in highly-skewed distributions, and has a good intuitive meaning. It should be used to supplement the standard deviation in most cases.

**Skew**

A distribution is skewed if one of its tails is longer than the other. The first distribution shown has a positive skew. This means that it has a long tail in the positive direction. The distribution below it has a negative skew since it has a long tail in the negative direction. Finally, the third distribution is symmetric and has no skew. Distributions with positive skew are sometimes called "skewed to the right" whereas distributions with negative skew are called "skewed to the left."

Distributions with positive skew are more common than distributions with negative skews. One example is the distribution of income. Most people make under $40,000 a year, but some make quite a bit more with a small number making many millions of dollars per year. The positive tail therefore extends out quite a long way whereas the negative tail stops at zero.

For a more psychological example, a distribution with a positive skew typically results if the time it takes to make a response is measured. The longest response times are usually much longer than typical response times whereas the shortest response times are seldom much less than the typical response time. A histogram of the author's performance on a perceptual motor task in which the goal is to move the mouse to and click on a small target as quickly as possible is shown below. The X axis shows times in milliseconds.
Negatively skewed distributions do occur, however. Consider the following frequency polygon of test grades on a statistics test where most students did very well but a few did poorly. It has a large negative skew.

Skew can be calculated as:

\[
\text{skew} = \frac{\sum (X-\mu)^3}{N\sigma^3}
\]

where \(\mu\) is the mean and \(\sigma\) is the standard deviation.

The normal distribution has a skew of 0 since it is a symmetric distribution.

As a general rule, the mean is larger than the median in positively skewed distributions and less than the median in negatively skewed distributions. (Click here for an illustration.) There are counter examples. For example it is not uncommon for the median to be higher than the mean in a positively skewed bimodal distribution or with discrete distributions. See "Mean, Median, and Skew: Correcting a Textbook Rule" by Paul Hippel, for more details.

Click here for an interaction demonstration.

**Kurtosis**

Kurtosis is based on the size of a distribution's tails. Distributions with relatively large tails are called "leptokurtic"; those with small tails are called "platykurtic." A distribution with the same kurtosis as the normal distribution is called "mesokurtic."

The following formula can be used to calculate kurtosis:

\[
\text{kurtosis} = \frac{\sum (X-\mu)^4}{N\sigma^4} - 3
\]

where \(\sigma\) is the standard deviation. The kurtosis of a normal distribution is 0.

The following two distributions have the same variance, approximately the same skew, but differ markedly in kurtosis.
Frequency Polygon

A frequency polygon is a graphical display of a frequency table. The intervals are shown on the X-axis and the number of scores in each interval is represented by the height of a point located above the middle of the interval. The points are connected so that together with the X-axis they form a polygon.

A frequency table and a relative frequency polygon for response times in a study on weapons and aggression are shown below. The times are in hundredths of a second.

<table>
<thead>
<tr>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Count</th>
<th>Cumulative Count</th>
<th>Per Cent</th>
<th>Cumulative Per Cent</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>3.12</td>
<td>3.12</td>
</tr>
<tr>
<td>30</td>
<td>35</td>
<td>4</td>
<td>5</td>
<td>12.48</td>
<td>15.62</td>
</tr>
<tr>
<td>35</td>
<td>40</td>
<td>8</td>
<td>13</td>
<td>24.96</td>
<td>40.62</td>
</tr>
<tr>
<td>40</td>
<td>45</td>
<td>15</td>
<td>28</td>
<td>46.80</td>
<td>87.50</td>
</tr>
<tr>
<td>45</td>
<td>50</td>
<td>3</td>
<td>31</td>
<td>9.36</td>
<td>96.88</td>
</tr>
<tr>
<td>50</td>
<td>55</td>
<td>1</td>
<td>32</td>
<td>3.12</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Note: Values in each category are > the lower limit and ≤ to the upper limit.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. The figure below provides an example. The data come from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together. The figure shows that although there is some overlap in times, it generally took longer to move the mouse to the small target than to the large one.
Frequency polygons can be based on the actual frequencies or the relative frequencies. When based on relative frequencies, the percentage of scores instead of the number of scores in each category is plotted.

In a cumulative frequency polygon, the number of scores (or the percentage of scores) up to and including the category in question is plotted. A cumulative frequency polygon is shown below.

Histogram

A histogram is constructed from a frequency table. The intervals are shown on the X-axis and the number of scores in each interval is represented by the height of a rectangle located above the interval. A histogram of the response times from the dataset Target RT is shown below.

The shapes of histograms will vary depending on the choice of the size of the intervals. Click here to see an interactive histogram demonstrating the effect of changing the size of the intervals.

A bar graph is much like a histogram, differing in that the columns are separated from each other by a small distance. Bar graphs are commonly used for qualitative variables.

Stem and Leaf Plots

A stem and leaf display (also called a stem and leaf plot) is a graphical method of displaying data. It is particularly useful when the data are not too numerous. A stem and leaf plot is similar to a histogram except it portrays a little more precision. A stem and leaf plot of the tournament players from the dataset "chess" as well as the data themselves are shown below:
The largest value, 85.3, is approximated as:

10 × 8 + 5.

This is represented in the plot as a stem of 8 and a leaf of 5. It is shown as the 5 in the first line of the plot. Similarly, 80.3 is approximated as 10 × 8 + 0; it has a stem of 8 and a leaf of 0. It is shown as the "0" in the first line of the plot.

Depending on the data, each stem is displayed 1, 2, or 5 times. When a stem is displayed only once (as on the plot shown above), the leaves can take on the values from 0-9.

When a stem is displayed twice, (as in the example shown below) one stem is associated with the leaves 5-9 and the other stem is associated with the leaves 0-4.

Finally, when a stem is displayed five times, the first has the leaves 8-9, the second 6-7, the third 4-5, and so on.

If positive and negative numbers are both present, +0 and -0 are used as stems as they are in the plot. A stem of -0 and a leaf of 7 is a value of (-0 × 1) + (-0.1 × 7) = -0.7.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common column of stems. The figure below shows such a graph. It compares two distributions. The stems are in the middle, the leaves to the left are for one distribution, and the leaves to the right are for the other. For example, the second-to-last row shows that the distribution on the left contains the values 11, 12, and 13 whereas the distribution on the right contains two 12's and three 14's.
A box plot provides an excellent visual summary of many important aspects of a distribution. The box stretches from the lower hinge (defined as the 25th percentile) to the upper hinge (the 75th percentile) and therefore contains the middle half of the scores in the distribution.

The median is shown as a line across the box. Therefore 1/4 of the distribution is between this line and the top of the box and 1/4 of the distribution is between this line and the bottom of the box.

The "H-spread" is defined as the difference between the hinges and a "step" is defined as 1.5 times the H-spread.

Inner fences are 1 step beyond the hinges. Outer fences are 2 steps beyond the hinges.
There are two adjacent values: the largest value below the upper inner fence and the smallest value above the lower inner fence. For the data plotted in the figure, the minimum value is above the lower inner fence and is therefore the lower adjacent value. The maximum value is the inner fences so it is not the upper adjacent value.

As shown in the figure, a line is drawn from the upper hinge to the upper adjacent value and from the lower hinge to the lower adjacent value.
Every score between the inner and outer fences is indicated by an "o" whereas a score beyond the outer fences is indicated by a "*".

It is often useful to compare data from two or more groups by viewing box plots from the groups side by side. Plotted are data from Example 2a and Example 2b. The data from 2b are higher, more spread out, and have a positive skew. That the skew is positive can be determined by the fact that the mean is higher than the median and the upper whisker is longer than the lower whisker.

Some computer programs present their own variations on box plots. For example, SPSS does not include the mean. JMP distinguishes between "outlier" box plots which are the same as those described here and quantile box plots that show the 10th, 25th, 50th, 75th, and 90th percentiles.

For apply the exercises at following link:
http://davidmlane.com/hyperstat/desc_univ.html
CHAPTER 3

Describing Bivariate Data
Scatterplots

A scatterplot shows the scores on one variable plotted against scores on a second variable. Below is a plot showing the relationship between grip strength and arm strength for 147 people working at physically-demanding jobs. The data are from a case study in the Rice Virtual Lab in Statistics. The plot shows a very strong but certainly not a perfect relationship between these two variables.

![Scatterplot of Grip Strength vs Arm Strength](image)

Scatterplots should almost always be constructed when the relationship between two variables is of interest. Statistical summaries are no substitute for a full plot of the data.

Pearson's Correlation

The correlation between two variables reflects the degree to which the variables are related. The most common measure of correlation is the Pearson Product Moment Correlation (called Pearson's correlation for short). When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho (ρ). When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between variables. The scatterplot shown on this page depicts such a relationship. It is a positive relationship because high scores on the X-axis are associated with high scores on the Y-axis.

A correlation of -1 means that there is a perfect negative linear relationship between variables. The scatterplot shown below depicts a negative relationship. It is a negative relationship because high scores on the X-axis are associated with low scores on the Y-axis.

![Scatterplot with Correlation of -1](image)

A correlation of 0 means there is no linear relationship between the two variables. The second graph shows a
Pearson correlation of 0.

Correlations are rarely if ever 0, 1, or -1. Some real data showing a moderately high correlation are shown on the next page.

The scatterplot below shows arm strength as a function of grip strength for 147 people working in physically-demanding jobs (click here for details about the study). The plot reveals a strong positive relationship. The value of Pearson’s correlation is 0.63.

Other information about Pearson's correlation can be obtained by clicking one of the following links:

- Computational formula
- Sampling distribution
- Confidence interval
- Confidence interval on difference between r's

**Computing Pearson's Correlation Coefficient**

The formula for Pearson's correlation takes on many forms. A commonly used formula is shown below. The formula looks a bit complicated, but taken step by step as shown in the numerical example, it is really quite simple.

\[
 r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}}
\]
A simpler looking formula can be used if the numbers are converted into z scores:

\[ r = \frac{\sum z_x z_y}{N} \]

where \( z_x \) is the variable X converted into z scores and \( z_y \) is the variable Y converted into z scores.

**Computing Pearson's Correlation in SPSS**
The Pearson's correlation is a measure of the relationship between two variables. In SPSS, you compute it by choosing

Analyze/Correlate/Bivariate...

Then select the variable(s) you want to correlate and click the arrow button that points to the right.
This will move the variables into the "variables" column.

Finally, click "OK." The results are shown below:
Example Correlations

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
<th>Y3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>Pearson Correlation</td>
<td>.1000</td>
<td>.116</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td></td>
<td>.853</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Y2</td>
<td>Pearson Correlation</td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.853</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Y3</td>
<td>Pearson Correlation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>-.368</td>
<td>.879*</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.05 level (2-tailed).

Restricted Range

Consider a study that investigated the correlation between arm strength and grip strength for 147 people working in physically-demanding jobs. Do you think the correlation was higher for all 147 workers tested, or for the workers who were above the median in grip strength? The upper portion of the figure below shows that the scatterplot for the entire sample of 147 workers. The lower portion of the figure shows the scatterplot for the 73 workers who scored highest on grip strength. The correlation is 0.63 for the sample of 147 but only 0.47 for the sample of 73. Notice that the scales of the X-axes are different.
Whenever a sample has a restricted range of scores, the correlation will be reduced. To take the most extreme example, consider what the correlation between high-school GPA and college GPA would be in a sample where every student had the same high-school GPA. The correlation would necessarily be 0.0.

Click [here](http://davidmlane.com/hyperstat/index.html) for an interactive demonstration of the effect of restriction of range and [here](http://davidmlane.com/hyperstat/index.html) for a related demonstration.

### Effect of Linear Transformations on Pearson's r

Linear transformations have no effect on Pearson's correlation coefficient. Thus, the correlation between height and weight is the same regardless of whether height is measured in inches, feet, centimeters or even miles. This is a very desirable property since, with the exception of ratio scales, choices among measurement scales that are linear transformations of each other are arbitrary. For instance, scores on the Scholastic Aptitude Test (SAT) range from 200-800. It was an arbitrary decision to set 200 to 800 as the range. The test would not be any different if 100 points were subtracted from each score and then each score were multiplied by 3. Scores on the SAT would then range from 300-2100. The Pearson's correlation between SAT and some other variable (such as college grade point average) would not be affected by this linear transformation.

### Spearman's rho

Spearman's rho is a measure of the linear relationship between two variables. It differs from Pearson's correlation only in that the computations are done after the numbers are converted to ranks. When converting to ranks, the smallest value on X becomes a rank of 1, etc. Consider the following X-Y pairs:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

Converting these to ranks would result in the following:
The first value of X (which was a 7) is converted into a 2 because 7 is the second lowest value of X. The X value of 5 is converted into a 1 since it is the lowest. Spearman's rho can be computed with the formula for Pearson's r using the ranked data. For this example, Spearman's rho = 0.60 Spearman's rho is an example of a "rank-randomization" test.

For apply the exercises at following link:
http://davidmlane.com/hyperstat/desc_biv.html
CHAPTER 4

Introduction to Probability
Simple probability

What is the probability that a card drawn at random from a deck of cards will be an ace? Since of the 52 cards in the deck, 4 are aces, the probability is 4/52. In general, the probability of an event is the number of favorable outcomes divided by the total number of possible outcomes. (This assumes the outcomes are all equally likely.) In this case there are four favorable outcomes: (1) the ace of spades, (2) the ace of hearts, (3) the ace of diamonds, and (4) the ace of clubs. Since each of the 52 cards in the deck represents a possible outcome, there are 52 possible outcomes.

The same principle can be applied to the problem of determining the probability of obtaining different totals from a pair of dice. As shown below, there are 36 possible outcomes when a pair of dice is thrown.

<table>
<thead>
<tr>
<th>Die 1</th>
<th>Die 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>

To calculate the probability that the sum of the two dice will equal 5, calculate the number of outcomes that sum to 5 and divide by the total number of outcomes (36). Since four of the outcomes have a total of 5 (1,4; 2,3; 3,2; 4,1), the probability of the two dice adding up to 5 is 4/36 = 1/9. In like manner, the probability of obtaining a sum of 12 is computed by dividing the number of favorable outcomes (there is only one) by the total number of outcomes (36). The probability is therefore 1/36.

Conditional Probability

A conditional probability is the probability of an event given that another event has occurred. For example, what is the probability that the total of two dice will be greater than 8 given that the first die is a 6? This can be computed by considering only outcomes for which the first die is a 6. Then, determine the proportion of these outcomes that total more than 8. All the possible outcomes for two dice are shown below:

<table>
<thead>
<tr>
<th>Die 1</th>
<th>Die 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>

There are 6 outcomes for which the first die is a 6, and of these, there are four that total more than 8 (6,3; 6,4; 6,5;
6,6). The probability of a total greater than 8 given that the first die is 6 is therefore 4/6 = 2/3.

More formally, this probability can be written as:

\[ p(\text{total} > 8 \mid \text{Die 1} = 6) = \frac{2}{3}. \]

In this equation, the expression to the left of the vertical bar represents the event and the expression to the right of the vertical bar represents the condition. Thus it would be read as "The probability that the total is greater than 8 given that Die 1 is 6 is 2/3." In more abstract form, \( p(A \mid B) \) is the probability of event A given that event B occurred.

**Probability of A and B**

**If A and B are Independent**

A and B are two events. If A and B are independent, then the probability that events A and B both occur is:

\[ p(A \text{ and } B) = p(A) \times p(B). \]

In other words, the probability of A and B both occurring is the product of the probability of A and the probability of B.

What is the probability that a fair coin will come up with heads twice in a row? Two events must occur: a head on the first toss and a head on the second toss. Since the probability of each event is 1/2, the probability of both events is: \( \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \).

Now consider a similar problem: Someone draws a card at random out of a deck, replaces it, and then draws another card at random. What is the probability that the first card is the ace of clubs and the second card is a club (any club). Since there is only one ace of clubs in the deck, the probability of the first event is 1/52. Since 13/52 = 1/4 of the deck is composed of clubs, the probability of the second event is 1/4. Therefore, the probability of both events is: \( \frac{1}{52} \times \frac{1}{4} = \frac{1}{208} \).

**If A and B are Not Independent**

If A and B are not independent, then the probability of A and B is:

\[ p(A \text{ and } B) = p(A) \times p(B \mid A) \]

where \( p(B \mid A) \) is the conditional probability of B given A.

If someone draws a card at random from a deck and then, without replacing the first card, draws a second card, what is the probability that both cards will be aces? Event A is that the first card is an ace. Since 4 of the 52 cards are aces, \( p(A) = \frac{4}{52} = \frac{1}{13} \). Given that the first card is an ace, what is the probability that the second card will be an ace as well? Of the 51 remaining cards, 3 are aces. Therefore, \( p(B \mid A) = \frac{3}{51} = \frac{1}{17} \) and the probability of A and B is:

\[ \frac{1}{13} \times \frac{1}{17} = \frac{1}{221}. \]

If events A and B are mutually exclusive, then the probability of A or B is simply:

\[ p(A \text{ or } B) = p(A) + p(B). \]
What is the probability of rolling a die and getting either a 1 or a 6? Since it is impossible to get both a 1 and a 6, these two events are mutually exclusive. Therefore,

$$p(1 \text{ or } 6) = p(1) + p(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

If the events A and B are not mutually exclusive, then

$$p(A \text{ or } B) = p(A) + p(B) - p(A \text{ and } B).$$

The logic behind this formula is that when $p(A)$ and $p(B)$ are added, the occasions on which A and B both occur are counted twice. To adjust for this, $p(A \text{ and } B)$ is subtracted. What is the probability that a card selected from a deck will be either an ace or a spade? The relevant probabilities are:

- $p(\text{ace}) = \frac{4}{52}$
- $p(\text{spade}) = \frac{13}{52}$

**Probability of A or B**

The only way in which an ace and a spade can both be drawn is to draw the ace of spades. There is only one ace of spades, so:

$$p(\text{ace and spade}) = \frac{1}{52}.$$

The probability of an ace or a spade can be computed as:

$$p(\text{ace}) + p(\text{spade}) - p(\text{ace and spade}) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}.$$

Consider the probability of rolling a die twice and getting a 6 on at least one of the rolls. The events are defined in the following way:

- Event A: 6 on the first roll: $p(A) = \frac{1}{6}$
- Event B: 6 on the second roll: $p(B) = \frac{1}{6}$

$$p(A \text{ and } B) = \frac{1}{6} \times \frac{1}{6}$$

$$p(A \text{ or } B) = \frac{1}{6} + \frac{1}{6} - \frac{1}{6} \times \frac{1}{6} = \frac{11}{36}$$

The same answer can be computed using the following admittedly convoluted approach: Getting a 6 on either roll is the same thing as not getting a number from 1 to 5 on both rolls. This is equal to: $1 - p(1 \text{ to } 5 \text{ on both rolls})$.

The probability of getting a number from 1 to 5 on the first roll is $\frac{5}{6}$. Likewise, the probability of getting a number from 1 to 5 on the second roll is also $\frac{5}{6}$ . Therefore, the probability of getting a number from 1 to 5 on both rolls is: $\frac{5}{6} \times \frac{5}{6} = \frac{25}{36}$. This means that the probability of not getting a 1 to 5 on both rolls (getting a 6 on at least one roll) is:

$$1 - \frac{25}{36} = \frac{11}{36}.$$

Despite the convoluted nature of this method, it has the advantage of being easy to generalize to three or more events. For example, the probability of rolling a die three times and getting a six on at least one of the three rolls is:

$$1 - \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = 0.421$$
In general, the probability that at least one of \( k \) independent events will occur is:

\[
1 - (1 - \alpha)^k
\]

where each of the events has probability \( \alpha \) of occurring

**Binomial distribution**

When a coin is flipped, the outcome is either a head or a tail; when a magician guesses the card selected from a deck, the magician can either be correct or incorrect; when a baby is born, the baby is either born in the month of March or is not. In each of these examples, an event has two *mutually exclusive* possible outcomes. For convenience, one of the outcomes can be labeled "success" and the other outcome "failure." If an event occurs \( N \) times (for example, a coin is flipped \( N \) times), then the binomial distribution can be used to determine the probability of obtaining exactly \( r \) successes in the \( N \) outcomes. The binomial probability for obtaining \( r \) successes in \( N \) trials is:

\[
P(r) = \frac{N!}{r!(N-r)!} \pi^r (1-\pi)^{N-r}
\]

where \( P(r) \) is the probability of exactly \( r \) successes, \( N \) is the number of events, and \( \pi \) is the probability of success on any one trial. This formula for the binomial distribution assumes that the events:

1. are dichotomous (fall into only two categories)
2. are *mutually exclusive*
3. are independent and
4. are randomly selected

Consider this simple application of the binomial distribution: What is the probability of obtaining exactly 3 heads if a fair coin is flipped 6 times?

For this problem, \( N = 6 \), \( r = 3 \), and \( \pi = 0.5 \). Therefore,

\[
P(3) = \frac{6!}{3!(6-3)!} (0.5)^3 (1 - 0.5)^{6-3} = \frac{6 \times 5 \times 4 \times 3 \times 2}{(3 \times 2)(3 \times 2)} \frac{(0.125)(0.125)}{(0.125)(0.125)} = 0.3125.
\]

Two binomial distributions are shown below. Notice that for \( \pi = 0.5 \), the distribution is symmetric whereas for \( \pi = 0.3 \), the distribution has a positive skew.
Often the cumulative form of the binomial distribution is used. To determine the probability of obtaining 3 or more successes with \( n = 6 \) and \( \pi = 0.3 \), you compute \( P(3) + P(4) + P(5) + P(6) \). This can also be written as:

\[
\sum_{i=3}^{5} F(r_i)
\]

and is equal to 0.1852 + 0.0595 + 0.0102 + 0.0007 = 0.2556. The binomial distribution can be approximated by a normal distribution (click here to see how). Click here for an interactive demonstration of the normal approximation to the binomial.

**Subjective probability**

For some purposes, probability is best thought of as subjective. Questions such as "What is the probability that Boston will defeat New York in an upcoming baseball game?" cannot be calculated by dividing the number of favorable outcomes by the number of possible outcomes. Rather, assigning probability 0.6 (say) to this event seems to reflect the speaker's personal opinion --- perhaps his or her willingness to bet according to certain odds. Such an approach to probability, however, seems to lose the objective content of the idea of chance; probability becomes mere opinion. Two people might attach different probabilities to the outcome, yet there would be no criterion for calling one "right" and the other "wrong." We cannot call one of the two people right simply because he or she assigned a higher probability to the outcome that actually occurred. After all, you would be right to attribute probability 1/6 to throwing a six with a fair die, and your friend who attributes 2/3 to this event would be wrong. And you are still right (and your friend is still wrong) even if the die ends up showing a six!

The following example illustrates the present approach to probabilities. Suppose you wish to know what the weather will be like next Saturday because you are planning a picnic. You turn on your radio, and the weather person says, “There is a 10% chance of rain.” You decide to have the picnic outdoors and, lo and behold, it rains. You are furious with the weather person. But was he or she wrong? No, they did not say it would not rain, only that rain was unlikely. The weather person would have been flatly wrong only if they said that the probability is 0 and it subsequently rained. However, if you kept track of the weather predictions over a long periods of time and found that it rained on 50% of the days that the weather person said the probability was 0.10, you could say his or her probability assessments are wrong.

So when is it sensible to say that the probability of rain is 0.10? According to a frequency interpretation, it means that it will rain 10% of the days on which rain is forecast with this probability.

**For apply the exercises at following link:**

http://davidmlane.com/hyperstat/probability.html
CHAPTER 5

Normal Distribution
What is a Normal Distribution?

Normal distributions are a family of distributions that have the same general shape. They are symmetric with scores more concentrated in the middle than in the tails. Normal distributions are sometimes described as bell shaped.

Examples of normal distributions are shown below. Notice that they differ in how spread out they are. The area under each curve is the same. The height of a normal distribution can be specified mathematically in terms of two parameters: the mean (μ) and the standard deviation (σ).

Mathematical Formula for Height of a Normal Distribution

The height (ordinate) of a normal curve is defined as:

\[
\frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}
\]

where μ is the mean and σ is the standard deviation, π is the constant 3.14159, and e is the base of natural logarithms and is equal to 2.718282.

x can take on any value from -infinity to +infinity.

f(x) is very close to 0 if x is more than three standard deviations from the mean (less than -3 or greater than +3).

Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. Normal distributions can be transformed to standard normal distributions by the formula:

\[
z = \frac{X - \mu}{\sigma}
\]

where X is a score from the original normal distribution, μ is the mean of the original normal distribution, and σ is
the standard deviation of original normal distribution. The standard normal distribution is sometimes called the z
distribution. A z score always reflects the number of standard deviations above or below the mean a particular score
is. For instance, if a person scored a 70 on a test with a mean of 50 and a standard deviation of 10, then they scored
2 standard deviations above the mean. Converting the test scores to z scores, an X of 70 would be:

\[ z = \frac{70 - 50}{10} = 2 \]

So, a z score of 2 means the original score was 2 standard deviations above the mean. Note that the z distribution
will only be a normal distribution if the original distribution (X) is normal.

Applying the formula

\[ z = \frac{X - \mu}{\sigma} \]

will always produce a transformed distribution with a mean of zero and a standard deviation of one. However, the
shape of the distribution will not be affected by the transformation. If X is not normal then the transformed
distribution will not be normal either. One important use of the standard normal distribution is for converting
between scores from a normal distribution and percentile ranks.

Areas under portions of the standard normal distribution are shown below. About 0.68 (0.34 + 0.34) of the
distribution is between -1 and 1 while about 0.96 of the distribution is between -2 and 2.

What's So Important about the Normal Distribution?

One reason the normal distribution is important is that many psychological and educational variables are distributed
approximately normally. Measures of reading ability, introversion, job satisfaction, and memory are among the many
psychological variables approximately normally distributed. although the distributions are only approximately
normal, they are usually quite close.

A second reason the normal distribution is so important is that it is easy for mathematical statisticians to work with.
This means that many kinds of statistical tests can be derived for normal distributions. Almost all statistical tests
discussed in this text assume normal distributions. Fortunately, these tests work very well even if the distribution is
only approximately normally distributed. Some tests work well even with very wide deviations from normality.

Finally, if the mean and standard deviation of a normal distribution are known, it is easy to convert back and forth
from raw scores to percentiles.

Converting to Percentiles and Back
If the **mean** and **standard deviation** of a **normal distribution** are known, it is relatively easy to figure out the **percentile rank** of a person obtaining a specific score. To be more concrete, assume a test in Introductory Psychology is normally distributed with a mean of 80 and a standard deviation of 5. What is the percentile rank of a person who received a score of 70 on the test?

Mathematical statisticians have developed ways of determining the proportion of a distribution that is below a given number of standard deviations from the mean. They have shown that only 2.3% of the population will be less than or equal to a score two standard deviations below the mean. (Click here to see why 70 is two standard deviations below the mean.) In terms of the Introductory Psychology test example, this means that a person scoring 70 would be in the 2.3rd percentile.

This graph shows the distribution of scores on the test. The shaded area is 2.3% of the total area. The proportion of the area below 70 is equal to the proportion of the scores below 70.

What about a person scoring 75 on the test? The proportion of the area below 75 is the same as the proportion of scores below 75.

A score of 75 is one standard deviation below the mean because the mean is 80 and the standard deviation is 5. Mathematical statisticians have determined that 15.9% of the scores in a normal distribution are lower than a score one standard deviation below the mean. Therefore, the proportion of the scores below 75 is 0.159 and a person scoring 75 would have a percentile rank score of 15.9.

The table below page gives the proportion of the scores below various values of z; z is computed with the formula:

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

where z is the number of standard deviations (σ) X is above the mean (μ).

Naturally, it is more convenient to use a **z calculator**.
When \( z \) is negative it means that \( X \) is below the mean. Thus, a \( z \) of \(-2\) means that \( X \) is \(-2\) standard deviations above the mean which is the same thing as being \(+2\) standard deviations below the mean.

To take another example, what is the percentile rank of a person receiving a score of 90 on the test?

The graph shows that most people scored below 90. Since 90 is 2 standard deviations above the mean

\[
z = \frac{(90 - 80)}{5} = 2
\]

it can be determined from the table that a \( z \) score of 2 is equivalent to the 97.7th percentile: The proportion of people scoring below 90 is thus .977.

What score on the Introductory Psychology test would it have taken to be in the 75th percentile? (Remember the test has a mean of 80 and a standard deviation of 5.) The answer is computed by reversing the steps in the previous problems.

First, determine how many standard deviations above the mean one would have to be to be in the 75th percentile. This can be found by using a \( z \) table and finding the \( z \) associated with 0.75. The value of \( z \) is 0.674. Thus, one must be .674 standard deviations above the mean to be in the 75th percentile.

Second, the standard deviation is 5, one must be:

\[
(5)(.674) = 3.37
\]

points above the mean. Since the mean is 80, a score of \( 80 + 3.37 = 83.37 \) is necessary. Rounding off, a score of 83 is needed to be in the 75th percentile. Since
a little algebra demonstrates that $X = \mu + z \sigma$. For the present example,

$$X = 80 + (.674)(5) = 83.37$$ as just shown in the figure.

**Area Under a Portion of the Normal Curve**

If a test is normally distributed with a mean of 60 and a standard deviation of 10, what proportion of the scores is above 85? This problem is very similar to figuring out the percentile rank of a person scoring 85. The first step is to figure out the proportion of scores less than or equal to 85. This is done by figuring out how many standard deviations above the mean 85 is. Since 85 is 85 - 60 = 25 points above the mean and since the standard deviation is 10, a score of 85 is $25/10 = 2.5$ standard deviations above the mean. Or, in terms of the formula,

$$z = \frac{X - \mu}{\sigma} = \frac{85-60}{10} = 2.5$$

A $z$ table can be used to calculate that 0.9938 of the scores are less than or equal to a score 2.5 standard deviations above the mean. It follows that only 1 - 0.9938 = 0.0062 of the scores are above a score 2.5 standard deviations above the mean. Therefore, only 0.0062 of the scores are above 85.

Suppose you wanted to know the proportion of students receiving scores between 70 and 80. The approach is to figure out the proportion of students scoring below 80 and the proportion below 70. The difference between the two proportions is the proportion scoring between 70 and 80. First, the calculation of the proportion below 80. Since 80 is 20 points above the mean and the standard deviation is 10, 80 is 2 standard deviations above the mean.
A z table can be used to determine that .9772 of the scores are below a score 2 standard deviations above the mean.

To calculate the proportion below 70,

\[ z = \frac{X - \mu}{\sigma} = \frac{70 - 60}{10} = 1 \]

A z table can be used to determine that the proportion of scores less than 1 standard deviation above the mean is 0.8413. So, if 0.1587 of the scores are above 70 and 0.0228 are above 80, then the proportion between 70 and 80 can be computed by subtracting 0.0228 from 0.1587:

\[ 0.1587 - 0.0228 = 0.1359 \]

Assume a test is normally distributed with a mean of 100 and a standard deviation of 15. What proportion of the scores would be between 85 and 105? The solution to this problem is similar to the solution to the last one. The first step is to calculate the proportion of scores below 85. Next, calculate the proportion of scores below 105. Finally, subtract the first result from the second to find the proportion scoring between 85 and 105.

Begin by calculating the proportion below 85. You can calculate that 85 is one standard deviation below the mean:

\[ z = \frac{X - \mu}{\sigma} = \frac{85 - 100}{15} = -1 \]

Using a z table with the value of -1 for z, the area below -1 (or 85 in terms of the raw scores) is 0.1587.
Doing the same thing for 105,

\[
z = \frac{X - \mu}{\sigma} = \frac{105 - 100}{15} = 0.333
\]

A *z table* shows that the proportion scoring below 0.333 (105 in raw scores) is .6306. The difference is .6306 - .1587 = .4719. So, .472 of the scores are between 85 and 105.

For apply the exercises at following link:
http://davidmlane.com/hyperstat/normal_distribution.html
CHAPTER 7

Point estimation
Overview of Point Estimation

When a parameter is being estimated, the estimate can be either a single number or it can be a range of scores. When the estimate is a single number, the estimate is called a "point estimate"; when the estimate is a range of scores, the estimate is called an interval estimate. Confidence intervals are used for interval estimates.

As an example of a point estimate, assume you wanted to estimate the mean time it takes 12-year-olds to run 100 yards. The mean running time of a random sample of 12-year-olds would be an estimate of the mean running time for all 12-year-olds. Thus, the sample mean, $M$, would be a point estimate of the population mean, $\mu$.

Often point estimates are used as parts of other statistical calculations. For example, a point estimate of the standard deviation is used in the calculation of a confidence interval for $\mu$. Point estimates of parameters are often used in the formulas for significance testing.

Point estimates are not usually as informative as confidence intervals. Their importance lies in the fact that many statistical formulas are based on them.

Characteristics of Estimators

Statistics are used to estimate parameters. Three important attributes of statistics as estimators are covered in this text: unbiasedness, consistency, and relative efficiency.

Most statistics you will see in this text are unbiased estimates of the parameter they estimate. For example, the sample mean, $M$, is an unbiased estimate of the population mean, $\mu$.

All statistics covered will be consistent estimators. It is hard to imagine a reasonably-chosen statistic that is not consistent.

When more than one statistic can be used to estimate a parameter, one will naturally be more efficient than the other(s). In general the relative efficiency of two statistics differs depending on the shape of the distribution of the numbers in the population. Statistics that minimize the sum of squared deviations such as the mean are generally the most efficient estimators for normal distributions but may not be for highly skewed distributions.

Estimating Variance

The formula for the variance computed in the population, $\sigma^2$, is different from the formula for an unbiased estimate of variance, $s^2$, computed in a sample. The two formulas are shown below:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

$$s^2 = \frac{\sum(X-M)^2}{(N-1)}$$

The unexpected difference between the two formulas is that the denominator is $N$ for $\sigma^2$ and is $N-1$ for $s^2$. That there should be a difference in formulas is very counterintuitive. To understand the reason that $N-1$ rather than $N$ is needed in the denominator of the formula for $s^2$, consider the problem of estimating $\sigma^2$ when the population mean, $\mu$, is already known.

Assume that you knew that the mean amount of practice it takes student pilots to master a particular maneuver is 12 hours. If you sampled one pilot and found he or she took 14 hours to master the maneuver, what would be your estimate of $\sigma^2$? The answer lies in considering the definition of variance: It is the average squared deviation of individual scores from $\mu$. 
With only one score, you have one squared deviation of a score from $\mu$. In this example, the one squared deviation is: $(X - \mu)^2 = (14-12)^2 = 4$.

This single squared deviation from the mean is the best estimate of the average squared deviation and is an unbiased estimate of $\sigma^2$. Since it is based on only one score, the estimate is not a very good estimate although it is still unbiased. It follows that if $\mu$ is known and $N$ scores are sampled from the population, then an unbiased estimate of $\sigma^2$ could be computed with the following formula:

$$\Sigma(X - \mu)^2/N.$$ 

Now it is time to consider what happens when $\mu$ is not known and $M$ is used as an estimate of $\mu$. Which value is going to be larger for a sample of $N$ values of $X$:

$$\Sigma(X - M)^2/N \text{ or } \Sigma(X - \mu)^2/N?$$

Since $M$ is the mean of the $N$ values of $X$ and since the sum of squared deviations of a set of numbers from their own mean is smaller than the sum of squared deviations from any other number, the quantity $\Sigma(X - M)^2/N$ will always be smaller than $\Sigma(X - \mu)^2/N$.

The argument goes that since $\Sigma(X - \mu)^2/N$ is an unbiased estimate of $\sigma^2$ and since $\Sigma(X - M)^2/N$ is always smaller than $\Sigma(X - \mu)^2/N$, then $\Sigma(X - M)^2/N$ must be biased and will have a tendency to underestimate $\sigma^2$. It turns out that dividing by $N-1$ rather than by $N$ increases the estimate just enough to eliminate the bias exactly.

Another way to think about why you divide by $N-1$ rather than by $N$ has to do with the concept of degrees of freedom. When $\mu$ is known, each value of $X$ provides an independent estimate of $\sigma^2$: Each value of $(X - \mu)^2$ is an independent estimate of $\sigma^2$. The estimate of $\sigma^2$ based on $N$ X's is simply the average of these $N$ independent estimates. Since the estimate of $\sigma^2$ is the average of these $N$ estimates, it can be written as:

$$\frac{\sum (X - \mu)^2}{df}$$

where there are $N$ degrees of freedom and therefore $df = N$. When $\mu$ is not known and has to be estimated with $M$, the $N$ values of $(X-M)^2$ are not independent because if you know the value of $M$ and the value of $N-1$ of the X's, then you can compute the value of the $N'th$ X exactly.

The number of degrees of freedom an estimate is based upon is equal to the number of independent scores that went into the estimate minus the number of parameters estimated en route to the estimation of the parameter of interest. In this case, there are $N$ independent scores and one parameter ($\mu$) is estimated en route to the estimation of the parameter of interest, $\sigma^2$. Therefore the estimate has $N-1$ degrees of freedom. The formula for $s^2$ can then be written as:

$$s^2 = \frac{\sum(X - M)^2}{df}$$
where df = N-1. Naturally, the greater the degrees of freedom the closer the estimate is likely to be to \( \sigma^2 \).

For apply the exercises at following link:
http://davidmlane.com/hyperstat/point_estimation.html
Overview of Confidence Intervals

Before a simple research question such as "What is the mean number of digits that can be remembered?" can be answered, it is necessary to specify the population of people to which it is addressed. The researcher could be
interested in, for example, adults over the age of 18, all people regardless of age, or students attending high school.

For the present example, assume the researcher is interested in students attending high school.

Once the population is specified, the next step is to take a random sample from it. In this example, let’s say that a sample of 10 students were drawn and each student’s memory tested. The way to estimate the mean of all high school students would be to compute the mean of the 10 students in the sample. Indeed, the sample mean is an unbiased estimate of μ, the population mean. But it will certainly not be a perfect estimate. By chance it is bound to be at least either a little bit too high or a little bit too low (or, perhaps, much too high or much too low).

For the estimate of μ to be of value, one must have some idea of how precise it is. That is, how close to μ is the estimate likely to be?

An excellent way to specify the precision is to construct a confidence interval. If the number of digits remembered for the 10 students were: 4, 4, 5, 5, 5, 6, 6, 7, 8, 9 then the estimated value of μ would be 5.9 and the 95% confidence interval would range from 4.71 to 7.09. (Click here to see how to compute the interval.)

The wider the interval, the more confident you are that it contains the parameter. The 99% confidence interval is therefore wider than the 95% confidence interval and extends from 4.19 to 7.61.

Below are shown some examples of possible confidence intervals. Although the parameter μ₁ - μ₂ represents the difference between two means, it is still valid to think of it as one parameter; π₁ - π₂ can also be thought of as one parameter.

<table>
<thead>
<tr>
<th>Lower Limit</th>
<th>Parameter</th>
<th>Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>π</td>
<td>0.7</td>
</tr>
<tr>
<td>-3.2</td>
<td>μ</td>
<td>4.5</td>
</tr>
<tr>
<td>3.5</td>
<td>μ₁ - μ₂</td>
<td>7.9</td>
</tr>
<tr>
<td>0.4</td>
<td>π</td>
<td>0.8</td>
</tr>
<tr>
<td>0.3</td>
<td>π₁ - π₂</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Confidence Interval for μ, Standard Deviation Known

This section explains how to compute a confidence interval for the mean of a normally-distributed variable for which the population standard deviation is known. In practice, the population standard deviation is rarely known. However, learning how to compute a confidence interval when the standard deviation is known is an excellent introduction to how to compute a confidence interval when the standard deviation has to be estimated.

Three values are used to construct a confidence interval for μ: the sample mean (M), the value of z (which depends on the level of confidence), and the standard error of the mean (σₓM). The confidence interval has M for its center and extends a distance equal to the product of z and σₓM in both directions. Therefore, the formula for a confidence interval is:
STATISTICS

http://davidmlane.com/hyperstat/index.html

\[ M - z \sigma_M \leq \mu \leq M + z \sigma_M. \]

Assume that the standard deviation of SAT verbal scores in a school system is known to be 100. A researcher wishes to estimate the mean SAT score and compute a 95% confidence interval from a random sample of 10 scores.

The 10 scores are: 320, 380, 400, 420, 500, 520, 600, 660, 720, and 780. Therefore, \( M \) = 530, \( N = 10 \), and \[ \sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{100}{\sqrt{10}} = 31.62. \]

The value of \( z \) for the 95% confidence interval is the number of standard deviations one must go from the mean (in both directions) to contain 0.95 of the scores.

It turns out that one must go 1.96 standard deviations from the mean in both directions to contain 0.95 of the scores. The value of 1.96 was found using a \( z \) table. Since each tail is to contain 0.025 of the scores, you find the value of \( z \) for which 1-0.025 = 0.975 of the scores are below. This value is 1.96.

All the components of the confidence interval are now known:
\( M = 530, \sigma_M = 31.62, z = 1.96. \)

Lower limit = \( 530 - (1.96)(31.62) = 468.02 \)

Upper limit = \( 530 + (1.96)(31.62) = 591.98 \)

Therefore, \( 468.02 \leq \mu \leq 591.98 \). Naturally, if a larger sample size had been used, the range of scores would have been smaller.

The computation of the 99% confidence interval is exactly the same except that 2.58 rather than 1.96 is used for \( z \). The 99% confidence interval is: \( 448.54 \leq \mu \leq 611.46 \). As it must be, the 99% confidence interval is even wider than the 95% confidence interval.

Summary of Computations

1. Compute \( M = \Sigma X/N. \)
2. Compute \( \sigma_M = \frac{\sigma}{\sqrt{N}} \)
3. Find \( z \) (1.96 for 95% interval; 2.58 for 99% interval)
4. Lower limit = \( M - z \sigma_M \)
5. Upper limit = \( M + z \sigma_M \)
Assumptions:

1. Normal distribution
2. \( \sigma \) is known
3. Scores are sampled randomly and are independent

Confidence Interval for \( \mu \), Standard Deviation Estimated

It is very rare for a researcher wishing to estimate the mean of a population to already know its standard deviation. Therefore, the construction of a confidence interval almost always involves the estimation of both \( \mu \) and \( \sigma \).

When \( \sigma \) is known, the formula:

\[ M - z\sigma_M \leq \mu \leq M + z\sigma_M \]

is used for a confidence interval. When \( \sigma \) is not known,

\[ S_M = \frac{S}{\sqrt{N}} \quad (N \text{ is the sample size)} \]

is used as an estimate of \( \sigma_M \). Whenever the standard deviation is estimated, the \( t \) rather than the normal \( (z) \) distribution should be used. The values of \( t \) are larger than the values of \( z \) so confidence intervals when \( \sigma \) is estimated are wider than confidence intervals when \( \sigma \) is known.

The formula for a confidence interval for \( \mu \) when \( \sigma \) is estimated is:

\[ M - t \, s_M \leq \mu \leq M + t \, s_M \]

where \( M \) is the sample mean, \( s_M \) is an estimate of \( \sigma_M \), and \( t \) depends on the degrees of freedom and the level of confidence.

The value of \( t \) can be determined from a \( t \) table. The degrees of freedom for \( t \) is equal to the degrees of freedom for the estimate of \( \sigma_M \) which is equal to \( N-1 \).

Suppose a researcher were interested in estimating the mean reading speed (number of words per minute) of high-school graduates and computing the 95% confidence interval. A sample of 6 graduates was taken and the reading speeds were: 200, 240, 300, 410, 450, and 600. For these data,

\[ M = 366.6667 \]
\[ s_M = 60.9736 \]
\[ df = 6-1 = 5 \]
\[ t = 2.571 \]
Therefore, the lower limit is: \( M - (t)(s_m) = 209.904 \) and the upper limit is: \( M + (t)(s_m) = 523.430 \), and the 95% confidence interval is:

\[ 209.904 \leq \mu \leq 523.430 \]

Thus, the researcher can conclude based on the rounded off 95% confidence interval that the mean reading speed of high-school graduates is between 210 and 523.

**Summary of Computations**

1. Compute \( M = \Sigma X/N \)
2. Compute \( s (\text{click for formula}) \)
3. \( S_M = \frac{s}{\sqrt{N}} \)
4. Compute \( df = N-1 \)
5. Find \( t \) for these \( df \) using a \( t \) table
6. Lower limit = \( M - t \ s_m \)
7. Upper limit = \( M + t \ s_m \)
8. Lower limit \( \leq \mu \leq \) Upper limit

**Assumptions:**

1. Normal distribution
2. Scores are sampled randomly and are independent

**General Formula for Confidence Intervals (1 of 2)**

If a statistic is normally distributed and the standard error of the statistic is known, then a confidence interval for that statistic can be computed as follows:

\[ \text{statistic} \pm (z) (\sigma_{\text{stat}}) \]

where \( \sigma_{\text{stat}} \) is the standard error of the statistic. For instance, the confidence interval for the mean is:

\[ M \pm z \sigma_M \]

where \( M \) is the sample mean and \( \sigma_M \) is the standard error of the mean. If the standard error has to be estimated then the formula uses \( t \):

\[ \text{statistic} \pm (t) (s_{\text{stat}}) \]

where \( s_{\text{stat}} \) is the estimated standard error of the statistic. For example, the confidence interval for the mean when \( \sigma \) is estimated is

\[ M \pm t \ s_m \]
The general formula applies to a whole host of other statistics such as the difference between means and Pearson's correlation. The one exception to this rule you will encounter is that when confidence intervals on proportions (or differences between proportions) are computed, z rather than t is used even though the standard error is estimated.

Confidence Interval on Difference between Means, Independent Groups, Standard Deviation Known

Following the general formula for a confidence interval, the formula for a confidence interval on the difference between means is:

\[ M_d \pm (z) (\sigma_{M_d}) \]

where

\[ M_d = M_1 - M_2 \] (the difference between means) is the statistic and

\[ \sigma_{M_d} \] (the standard error of the difference between means) is the standard error.

\[ z \] depends on the level of confidence desired (1.96 for the 95% interval and 2.58 for the 99% interval).

For a specific example, consider an hypothetical experiment on reaction time (the time it takes to push a button in response to a signal). One group is given a simple reaction time task (a light comes on and they push the button.) A second group is given a choice reaction time task (an "A" or a "B" is presented and the subjects push one button for "A" and a different button for "B"). Assume that it is known that the standard deviation (\( \sigma \)) for the simple reaction time task is 75 and the standard deviation for the choice reaction time task is 100. Five subjects are tested with the simple reaction time task and six subjects with the choice reaction time task.

Their scores (in msec) are shown below:

Simple RT: 205, 230, 280, 320, 390
Choice RT: 300, 310, 380, 410, 502, 570

\[ M_{\text{Simple}} = 285 \] and \[ M_{\text{Choice}} = 412. \]

The problem is to estimate the difference between simple and choice reaction time by computing the 95% confidence interval on the difference. The elements required to compute the interval are: the \( z \) for the 95% confidence interval, the mean difference (\( M_d \)), and the standard error of the mean difference (\( \sigma_{M_d} \)).

\[ z = 1.96 \]

\[ M_d = 285 - 412 = -127 \]

\[ \sigma_{M_d} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{75^2}{5} + \frac{100^2}{6}} = 52.836. \]

The 95% confidence interval is: \(-127 \pm (1.96)(52.836) = -127 \pm 103.559 \)
here $\mu_1$ is the mean for simple RT and $\mu_2$ is the mean for choice RT.

Summary of Computations

1. Compute $M_d = M_1 - M_2$

$$\sigma_{M_d} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

2. Compute

3. Find $z$ (1.96 for 95% interval; 2.58 for 99% interval)

4. Lower limit = $M_d - z \sigma_{M_d}$

5. Upper limit = $M_d + z \sigma_{M_d}$

6. Lower limit $\leq \mu_1 - \mu_2 \leq$ Upper limit

Assumptions:

1. The populations are each normally distributed.

2. $\sigma$ is known

3. Scores are sampled randomly and are independent

Confidence Interval on Difference Between Means, Independent Groups, Standard Deviation Estimated

Following the general formula for a confidence interval, the formula for a confidence interval on the difference between means ($M_1 - M_2$) is:

$$M_d \pm (t)(S_{M_d})$$

where $M_d = M_1 - M_2$ is the statistic and $S_{M_d}$ is an estimate of $\sigma_{M_d}$ (the standard error of the difference between means). $t$ depends on the level of confidence desired and on the degrees of freedom. The estimated standard error, $S_{M_d}$, is computed assuming that the variances in the two populations are equal. If the two sample sizes are equal ($n_1 = n_2$) then the population variance $\sigma^2$ (it is the same in both populations) is estimated by using the following formula:

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

where $MSE$ (which stands for mean square error) is an estimate of $\sigma^2$. Once $MSE$ is calculated, $S_{M_d}$ can be computed...
A concrete example should make the procedure for computing the confidence interval clearer. Assume that an experimenter were interested in computing the 99% confidence interval on the difference between the memory spans of seven- and nine-year old children. Four children at each age level are tested and their memory spans are shown below:

<table>
<thead>
<tr>
<th>Age 7</th>
<th>Age 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

The first step is to compute the means of each group:

\[ M_{\text{Age 7}} = 3.75 \text{ and } M_{\text{Age 9}} = 6.00. \]

Therefore, \( M_d = 3.75 - 6.00 = -2.25. \)

To obtain a value of \( t \), one must first compute the degrees of freedom (df). The degrees of freedom is equal to the degrees of freedom for MSE (MSE is used to estimate \( \sigma^2 \)). Since MSE is made up of two estimates of \( \sigma^2 \) (one for each sample), the df for MSE is the sum of the df for these two estimates. Therefore, the df for MSE is \((n - 1) + (n - 1) = 3 + 3 = 6.\)

A \textit{t table} shows that the value of \( t \) for a 99% confidence interval for 6 df is 3.707. The only remaining term is \( S_{M_d}. \)

The first step is to compute \( S^2_1 \) and \( S^2_2 \):

\[ S^2_1 = 0.917 \text{ and } S^2_2 = 0.667. \]

\[ \text{MSE} = (0.917 + 0.667)/2 = .792. \]

From the formula:

\[ S_{M_d} = \sqrt{\frac{2\text{MSE}}{n}} \]
All the terms needed to construct the confidence interval have now been computed. The lower limit (LL) of the interval is:

\[ LL = M_d - s_{M_d t} \]

\[ = -2.25 - (3.707)(0.629) \]

\[ = -4.58. \]

UL = -2.25 + (3.707)(0.629) = 0.09.

Therefore -4.58 \( \leq \mu_1 - \mu_2 \leq 0.09. \)

The calculations are only slightly more complicated when the sample sizes are different \( (n_1 \text{ does not equal } n_2). \) The first difference in the calculations is that MSE is computed differently. If the two values of \( s^2 \) were simply averaged as they are in the case of equal sample sizes, then the estimate based on the smaller sample size would count as much as the estimate based on the larger sample size. Instead the formula for MSE is:

\[ \text{MSE} = \frac{\text{SSE}}{df} \]

where df is the degrees of freedom and SSE is the sum of squares error and is defined as:

\[ \text{SSE} = \text{SSE}_1 + \text{SSE}_2 \]

\[ \text{SSE}_1 = \sum(X - M_1)^2 \] where the X's are from the first group (sample) and \( M_1 \) is the mean of the first group.

Similarly, \( \text{SSE}_2 = \sum(X - M_2)^2 \)

where the X's are from the second group and \( M_2 \) is the mean of the second group. The formula:

\[ s_{M_d} = \sqrt{\frac{2\text{MSE}}{n}} \]

cannot be used without modification since there is not one value of \( n \) but two: \( (n_1 \text{ and } n_2). \)

The solution is to use the harmonic mean of the two sample sizes for \( n. \) The harmonic mean \( (n_h) \) of \( n_1 \) and \( n_2 \) is:

\[ n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}} \]

Therefore the formula for the estimated standard error of the difference between means is:
For the example of the confidence interval on the difference between memory spans of seven-year olds and nine-year olds, assume that one more seven year old was tested and the resulting memory span score was 3.

For these data, $M_1 = 3.6$, $M_2 = 6.0$, and $M_d = -2.4$

$SSE_1 = (3-3.6)^2 + (3-3.6)^2 + (3 - 3.6)^2 + (4 - 3.6)^2 + (5 - 3.6)^2 = 3.2.$

$SSE_2 = (5-6)^2 + (6-6)^2 + (6-6)^2 + (7-6)^2 = 2.0.$

Therefore $SSE = 3.2 + 2.0 = 5.2$. The df are equal to the sum of the df for $S_1$ and $S_2$ which is $4 + 3 = 7$.

Since $MSE = SSE/df$, $MSE = 5.2/7 = 0.743$.

The harmonic mean of the n’s is:

$$n_h = \frac{2}{\frac{1}{5} + \frac{1}{4}} = \frac{2}{0.2 + 0.25} = 4.4444.$$  

Finally, a t table can be used to find that the value of t (with 7 df) to be used for the 99% confidence interval is 3.499.

The confidence interval is therefore:

$$LL = -2.4 - (3.499)(0.578) = -4.42$$

$$UL = -2.4 + (3.499)(0.578) = -0.38$$

$$-4.42 \leq \mu_1 - \mu_2 \leq -0.38$$

Summary of Computations

1. Compute $M_d = M_1 - M_2$
2. Compute \( \text{SSE}_1 = \sum (X - M_1)^2 \) for Group 1 and \( \text{SSE}_2 = \sum (X - M_2)^2 \) for Group 2

3. Compute \( \text{SSE} = \text{SSE}_1 + \text{SSE}_2 \)

4. Compute \( \text{df} = N - 2 \) where \( N = n_1 + n_2 \)

5. Compute \( \text{MSE} = \frac{\text{SSE}}{\text{df}} \)

6. Find \( t \) from a \( t \) table.

7. Compute

\[
 n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}}
\]

(If the sample sizes are equal then \( n_h = n_1 = n_2 \)).

8. Compute:

\[
 s_{M_d} = \sqrt{\frac{2 \text{MSE}}{n_h}}
\]

9. Lower limit = \( M_d - t \cdot s_{M_d} \)

10. Upper limit = \( M_d + t \cdot s_{M_d} \)

11. Lower limit \( \leq \mu_1 - \mu_2 \leq \) Upper limit

Assumptions:

1. The populations each are normally distributed.

2. Homogeneity of variance

3. Scores are sampled randomly and independently from 2 different populations

Confidence Interval on Linear Combination of Means, Independent Groups

Following the general formula for a confidence interval, the formula for a confidence interval on a linear combination of means is:

\[
 L \pm t \cdot s_i
\]

where \( L \) is the linear combination of the sample means, \( t \) depends on the level of confidence desired and the degrees of freedom, and \( S_i \) is an estimate of \( \sigma_i \), the standard error of a linear combination of means.

The formula for \( S_i \) is:
which is the same as the formula for σ except that MSE (an estimate of σ²) is used in place of σ². The formula for MSE here is very similar to the formula for MSE used in the calculation of a confidence interval on the difference between two means. The formula is:

\[ \text{MSE} = \frac{\sum s_i^2}{k} \]

where \( s_i^2 \) is the sample variance of the ith group and k is the number of groups. This formula assumes homogeneity of variance and that the k sample sizes are equal.

As an example, consider a hypothetical experiment in which aspirin, Tylenol, and a placebo are tested to see how much pain relief each provides. The amount of pain relief is rated on a five-point scale. Four subjects are tested in each group. The data are shown below:

<table>
<thead>
<tr>
<th>Group</th>
<th>Aspirin</th>
<th>Tylenol</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

The problem is to calculate the 95% confidence interval on the difference between the average of the aspirin and Tylenol groups and the placebo. That parameter being estimated is:

\[ \frac{\mu_{\text{aspirin}} + \mu_{\text{Tylenol}}}{2} - \mu_{\text{placebo}} \]

where \( \mu_{\text{aspirin}} \) is the population mean for aspirin, \( \mu_{\text{Tylenol}} \) is the population mean for Tylenol, and \( \mu_{\text{placebo}} \) is the population mean for the placebo.

The first step is to find the coefficients (a's) so that

\[ \sum a_i \mu_i = \frac{\mu_{\text{aspirin}} + \mu_{\text{Tylenol}}}{2} - \mu_{\text{placebo}} \]

where aspirin is Condition 1, Tylenol is Condition 2, and placebo is Condition 3. The solution is: \( a_1 = 0.5, a_2 = 0.5 \) and \( a_3 = -1 \). The next step is to compute the 3 sample means.

They are:

\[ M_{\text{aspirin}} = 4 \]
\[ M_{\text{Tylenol}} = 3, \text{ and} \]
\[ M_{\text{placebo}} = 2. \]

L can now be computed as: \( L = (.5)(4) + (.5)(3) + (-1)(2) = 1.5 \).

The estimated value of
In order to compute MSE, it is necessary to compute \( s^2 \) for each group. The values are: 1.334, 1.334, and 0.666. Therefore,

\[
\text{MSE} = \frac{(1.334 + 1.334 + 0.666)}{3} = 1.111.
\]

\[\Sigma a^2 = [.5^2 + .5^2 + (-1)^2] = 1.5\]

Since there are four subjects in each group, \( n = 4 \)

The component parts of \( s \) are now computed:

\[
s_L = \sqrt{\frac{(1.5)(1.111)}{4}} = 0.6455.
\]

Finally, to compute \( t \) it is necessary to know the degrees of freedom. Since three values of \( s^2 \) went into MSE and each \( s^2 \) was based on \( n - 1 = 3 \) df, the degrees of freedom for MSE and for \( t \) are \( 3 \times 3 = 9 \). The value of \( t \) for the 95% confidence interval with 9 df can be found from a \textit{t} table: It is 2.262. The lower limit (LL) of the confidence interval is:

\[
\text{LL} = 1.5 - (2.262)(.6455) = 0.04.
\]

The upper limit (UL) is:

\[
\text{UL} = 1.5 + (2.262)(.6455) = 2.96.
\]

The 95% confidence interval is therefore:

\[
0.04 \leq \frac{\mu_1 + \mu_2}{2} - \mu_3 \leq 2.96.
\]

The results of this experiment show that, based on the 95% confidence interval, the average effect of aspirin and Tylenol is somewhere between 0.04 and 2.96 units more effective than a placebo.

**Summary of Computations**

1. Compute the sample mean (M) for each group.
2. Compute the sample variance (\( s^2 \)) for each of the \( k \) groups.
3. Find the coefficients (\( a's \)) so that \( \Sigma a\mu_i \) is the parameter to be estimated.
4. Compute \( L = a_1M_1 + a_2M_2 + \ldots + a_kM_k \)
5. Compute MSE = \( \Sigma s^2 / k \)
6. Compute

\[ s_1 = \sqrt{\frac{\sum a_i^2}{n}} \]

7. Compute \( df = k(n-1) \) where \( k \) is the number of groups and \( n \) is the number of subjects in each group.

8. Find \( t \) for the \( df \) and level of confidence desired using a \( t \) table

9. Lower limit = \( L - t \cdot s_L \)

10. Upper limit = \( L + t \cdot s_L \)

11. Lower limit \( \leq \sum a_i \mu_i \leq \) Upper limit

Assumptions:

1. All populations are normally distributed.

2. All population variances are equal (homogeneity of variance)

3. Scores are sampled randomly and independently from \( k \) different populations.

4. The sample sizes are equal.

Confidence Interval on Pearson's Correlation

Since the sampling distribution of Pearson's \( r \) is not normally distributed, Pearson's \( r \) is converted to Fisher's \( z' \) and the confidence interval is computed using Fisher's \( z' \). The values of Fisher's \( z' \) in the confidence interval are then converted back to Pearson's \( r \)'s. For example, assume a researcher wished to construct a 99% confidence interval on the correlation between SAT scores and grades in the first year in college at a large state university. The researcher obtained data from 100 students chosen at random and found that the sample value of Pearson's \( r \) was 0.60. The first step in computing the confidence interval is to convert 0.60 to a value of \( z' \) using the \( r \) to \( z' \) table. The value is: \( z' = 0.69 \).

The sampling distribution of \( z' \) is known to be approximately normal with a standard error of

\[ \sigma_{z'} = \frac{1}{\sqrt{N - 3}} \]

where \( N \) is the number of pairs of scores.

From the general formula for a confidence interval, the formula for a confidence interval on \( z' \) is:

\[ z' \pm z \cdot \sigma_{z'} \]

which equals
A z table can be used to find that for the 99% confidence interval you should use a z of 2.58. For N = 100, the standard error of z' is 0.102.

The confidence interval for z' can be computed as:

\[
\text{Lower limit} = 0.69 - (2.58)(0.102) = 0.43 \\
\text{Upper limit} = 0.69 + (2.58)(0.102) = 0.95
\]

Using the r to z' table to convert the values of 0.43 and 0.95 back to Pearson r's, it turns out that the confidence interval for the population value of Pearson's correlation, \( \rho \), is:

\[0.41 \leq \rho \leq 0.74\]

Therefore, the correlation between SAT scores and college grades is highly likely to be somewhere between 0.41 and 0.74.

**Summary of Computations**

1. Compute the sample r.
2. Use the r to z' table to convert the value of r to z'.
3. Use a z table to find the value of z for the level of confidence desired.

\[
\sigma_z = \frac{1}{\sqrt{N - 3}}
\]

4. Compute:
   5. Lower limit = \( z' - (z)(\sigma_z) \)
   6. Upper limit = \( z' + (z)(\sigma_z) \)
   7. Use procedure the r to z' table to convert the lower and upper limits from z' to r

**Assumptions**

1. Each subject (observation) is sampled independently from each other subject.
2. Subjects are sampled randomly.

**Confidence Interval, Difference between Independent Correlations**

The procedure for computing a confidence interval on the difference between two independent correlations is similar to the procedure for computing a confidence interval on one correlation. The first step is to convert both values of r to \( z' \). Then a confidence interval is constructed based on the general formula for a confidence interval where the statistic is \( z'_1 - z'_2 \), and the standard error of the statistic is:
where $N_1$ is the number of pairs of scores in $r_1$ and $N_2$ is the number of pairs of scores in $r_2$. Once the confidence interval is computed, the upper and lower limits are converted back from $z'$ to $r$. As an example, assume a researcher were interested in whether the correlation between verbal and quantitative SAT scores (VSAT and QSAT) is different for females than it is for males. Samples of 80 females and 75 males are tested and the correlation for females is 0.55 and the correlation for males is 0.42.

The problem is to construct the 95% confidence on the difference between correlations. The formula is:

$$z_{1'} - z_{2'} \pm z \sigma_{z_{1'}z_{2'}}$$

The first step is to use the $r$ to $z'$ table to convert the two $r$'s. The $r$ of 0.55 corresponds to a $z'$ of 0.62 and the $r$ of 0.42 corresponds to a $z'$ of 0.45. Therefore,

$$z_{1'} - z_{2'} = 0.62 - 0.45 = 0.17.$$  

The next step is to find $z$. From a $z$ table, it can be determined that the $z$ for 95% confidence intervals is 1.96. Finally,

$$\sigma_{z_{1'}z_{2'}} = \frac{1}{\sqrt{N_1 - 3}} + \frac{1}{\sqrt{N_2 - 3}} = \sqrt{\frac{1}{77} + \frac{1}{72}} = 0.164.$$

Therefore,

Lower limit = $0.17 - (1.96)(0.164) = -0.15$
Upper limit = $0.17 + (1.96)(0.164) = 0.49$

Converting the lower and upper limits back to $r$, results in $r$'s of -0.15 and 0.45 respectively. Therefore, the confidence interval is:

$-0.15 \leq \rho_1 - \rho_2 \leq 0.45$

where $\rho_1$ is the population correlation for females and $\rho_2$ is the population correlation for males.

The experiment shows that there is still uncertainty about the difference in correlations. It could be that the correlation for females is as much as 0.49 higher. However, it could also be 0.15 lower.

**Summary of Computations**

1. Compute the sample $r$'s.
2. Use an $r$ to $z'$ table to convert the values of $r$ to $z'$. 

**STATISTICS**

http://davidmlane.com/hyperstat/index.html
3. Compute $z_1' - z_2'$.

4. Use a to $z$ table to find the value of $z$ for the level of confidence desired.

$$
\sigma_{z_1' - z_2'} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}
$$

5. Compute:

6. Lower limit = $z_1' - z_2' - (z)(\sigma_{z_1' - z_2'})$

7. Upper limit = $z_1' - z_2' + (z)(\sigma_{z_1' - z_2'})$

8. Use an $r$ to $z'$ table to convert the lower and upper limits to $r$'s.

Assumptions:

1. Each subject (observation) is sampled independently from each other subject.

2. Subjects are sampled randomly.

3. The two correlations are from independent samples (different groups of subjects). If the same subjects were used for both correlations then the assumption of independence would be violated.

Confidence Interval for a Proportion

Applying the general formula for a confidence interval, the confidence interval for a proportion, $\pi$, is:

$$p \pm z \sigma_p$$

where $p$ is the proportion in the sample, $z$ depends on the level of confidence desired, and $\sigma_p$, the standard error of a proportion, is equal to:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}}$$

where $\pi$ is the proportion in the population and $N$ is the sample size. Since $\pi$ is not known, $p$ is used to estimate it. Therefore the estimated value of $\sigma_p$ is:

$$\sqrt{\frac{p(1-p)}{N}}$$

As an example, consider a researcher wishing to estimate the proportion of X-ray machines that malfunction and produce excess radiation. A random sample of 40 machines is taken and 12 of the machines malfunction. The problem is to compute the 95% confidence interval on $\pi$, the proportion that malfunction in the population.

The value of $p$ is $12/40 = 0.30$. The estimated value of $\sigma_p$ is $\sqrt{\frac{(0.3)(0.7)}{40}} = 0.072$. 
A z table can be used to determine that the z for a 95% confidence interval is 1.96. The limits of the confidence interval are therefore:

Lower limit = .30 - (1.96)(0.072) = .16
Upper limit = .30 + (1.96)(0.072) = .44.

The confidence interval is: 0.16 ≤ π ≤ .44.

Correction for Continuity
Since the sampling distribution of a proportion is not a continuous distribution, a slightly more accurate answer can be arrived at by applying the correction for continuity. This is done simply by subtracting 0.5/N from the lower limit and adding 0.5/N to the upper limit. For the present example, 0.5/N = 0.5/40 = 0.01. Therefore the corrected interval is: 0.15 ≤ π ≤ 0.45.

Summary of Computations

1. Compute p
2. Estimate σp by \( \sqrt{\frac{p(1-p)}{N}} \)
3. Find z for the level of confidence desired with a z table.
4. Lower limit = p - (z)(Estimated σp) - 0.5/N
5. Upper limit = p + (z)(Estimated σp) + 0.5/N
6. Lower limit ≤ π ≤ Upper limit

Assumptions

1. Observations are sampled randomly and independently.
2. The adequacy of the normal approximation depends on the sample size (N) and π. Although there are no hard and fast rules, the following is a guide to needed sample size:

   If π is between 0.4 and 0.6 then an N of 10 is adequate. If π is as low as 0.2 or as high as 0.8 then N should be at least 25. For π as low as 0.1 or as high as 0.9, N should be at least 30.

   A more conservative rule of thumb that is often recommended is that Nπ and N(1 - π) should both be at least 10.

Click here for an interactive demonstration of the normal approximation to the binomial to explore the validity of these rules of thumb.

Confidence Interval on the Difference Between Proportions

The confidence interval on the difference between proportions is based on the same general formula as are other confidence intervals. A confidence interval on the difference between two proportions is computed in the following
situation: There are two populations and the members of each population can be classified as falling into one of two categories. For example, the categories might be such things as whether or not one has a high-school degree or whether or not one has ever been arrested.

Consider a researcher interested in whether people who majored in psychology are more or less likely than physics majors to solve a problem that involves a certain type of statistical reasoning. The researcher is interested in estimating the difference in the proportions of people in the two populations that can solve the problem and in computing a 99% confidence interval on the difference. Random samples of 100 psychology majors and 110 physics majors are taken and each person is given a chance to solve the problem. Of the 100 psychology majors, 65 solve the problem; of the 110 physics majors only 45 solve it.

Therefore the proportions who solve the problem are 0.65 for the psychology majors ($p_1$) and 0.41 for the physics majors ($p_2$). The goal of the experiment is to estimate the difference between the proportion of psychology majors in the population that can solve the problem ($\pi_1$) and the proportion of physics majors in the population that can solve it ($\pi_2$). The statistic $p_1 - p_2$ is used as an estimate of $\pi_1 - \pi_2$. In this experiment, the estimated value of $\pi_1 - \pi_2$ is $0.65 - 0.41 = 0.24$.

To compute the confidence interval, the standard error of $p_1 - p_2$ is needed. The standard error is:

$$\sigma_{p_1-p_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

The estimated standard error is:

$$s_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.65(1-0.65)}{100} + \frac{0.41(1-0.41)}{110}} = 0.067$$

Finally, the value of $z$ for the 99% confidence interval is computed using a $z$ table; it is 2.58. The lower limit of the confidence interval is simply:

$$p_1 - p_2 - (z)(\text{estimated } \sigma_{p_1-p_2})$$

$$= 0.65 - 0.41 - (2.58)(0.067) = 0.07$$

The upper limit is:

$$= 0.65 - 0.41 + (2.58)(0.067) = 0.41$$

The 99% confidence interval is therefore:

$$0.07 \leq \pi_1 - \pi_2 \leq 0.41.$$ 

This indicates that the proportion of psychology majors that can solve the problem is from 0.07 to 0.41 higher than the proportion of physics majors that can solve it.
Summary of Computations

1. Compute $p_1 - p_2$.

2. Find $z$ for confidence interval using a $z$ table.

3. $\sigma_{p_1 - p_2}$ with the formula:

$$s_{p_1 - p_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} = \sqrt{\frac{0.65(1 - 0.65)}{100} + \frac{0.41(1 - 0.41)}{110}}$$

4. lower limit = $p_1 - p_2 - (z) (\text{estimated} \sigma_{p_1 - p_2})$

5. upper limit = $p_1 - p_2 + (z) (\text{estimated} \sigma_{p_1 - p_2})$

6. Confidence interval: lower limit $\leq \pi_1 - \pi_2 \leq$ upper limit.

Assumptions

1. The two proportions are independent.

2. The adequacy of the normal approximation depends on the sample size ($N$) and $\pi$. Although there are no hard and fast rules, the following is a guide to needed sample size:

   If $\pi$ is between 0.4 and 0.6 then an $N$ of 10 is adequate. If $\pi$ is as low as 0.2 or as high as 0.8 then $N$ should be at least 25. For $\pi$ as low as 0.1 or as high as 0.9, $N$ should be at least 30.

   A more conservative rule of thumb that is often recommended is that $N\pi$ and $N(1 - \pi)$ should both be at least 10.

Click here for an interactive demonstration of the normal approximation to the binomial to explore the validity of these rules of thumb.

For apply the exercises at following link:
http://davidmlane.com/hyperstat/confidence_intervals.html
CHAPTER 9

Logic of hypothesis testing

Ruling out Chance as an Explanation

When an independent variable appears to have an effect, it is very important to be able to state with confidence that the effect was really due to the variable and not just due to chance. For instance, consider a hypothetical
experiment on a new antidepressant drug. Ten people suffering from depression were sampled and treated with the new drug (the experimental group); an additional 10 people were sampled from the same population and were treated only with a placebo (the control group). After 12 weeks, the level of depression in all subjects was measured and it was found that the mean level of depression (on a 10-point scale with higher numbers indicating more depression) was 4 for the experimental group and 6 for the control group. The most basic question that can be asked here is: "How can one be sure that the drug treatment rather than chance occurrences were responsible for the difference between the groups?" It could be, that by chance, the people who were randomly assigned to the treatment group were initially somewhat less depressed than those randomly assigned to the control group.

Or, it could be that, by chance, more pleasant things happened to the experimental group than to the control group over the 12 weeks of the experiment. The way this problem is approached statistically is to calculate how often one would get a difference as large or larger than the one obtained in the experiment if the experimental treatment really had no effect (and thus the differences were due to chance). If a difference as large or larger than the one obtained in the experiment could be expected to occur by chance relatively frequently, say, one out of every four times, then chance would remain a viable explanation of the effect. If such a difference would only occur by chance very rarely, then chance would not be a viable explanation.

Returning to the study on the effectiveness of the antidepressant, recall that the experimental group differed from the control group by 6 - 4 = 2 units on the depression scale. For the sake of argument, assume that if there were no true difference between means, the sampling distribution of the difference between means would be as shown on the next page.

The mean of the sampling distribution is 0 since the assumption is that there are no differences between population means and therefore the average difference between sample means would be zero. The graph of the sampling distribution shows that a difference between sample means of two or more is not an improbable occurrence even when there is no difference between population means. Twenty percent of the time the Drug group would be two or more points lower and 20% of the time the Control group would be two or more points higher. How should the results of the experiment be interpreted in light of this analysis? Science requires a conservative approach to decision making: a conclusion is accepted by the scientific community only if the data supporting it are strong enough to convince a skeptic. No skeptic would be convinced that the antidepressant drug rather than chance caused the difference.

The skeptic would argue:
"Since the differences could easily have been produced by chance factors, why should I believe in the effectiveness of this drug? Perhaps the drug is effective, perhaps not. I am not convinced." Now, for the sake of argument, assume that if there were no true difference between means, the sampling distribution of the difference between means would be as shown below instead of the one shown on the previous page.
Under this assumption, a difference of two or more on the depression scale would occur extremely infrequently, about once in a thousand times. The skeptic would be hard pressed to argue that the antidepressant had no effect. The skeptic's thinking might be as follows on the next page:

"Of course, things that happen extremely infrequently occasionally do happen. This might be one of those times. However, being skeptical does not mean being totally unwilling to accept new findings. The data are quite strong so I am forced to conclude that the drug is more effective than the placebo."

Naturally, there is a large degree of subjectivity in deciding how unlikely results must be (given that only chance is operating) before it should be concluded that chance is not responsible for the effect. Traditionally, researchers have used either the 5% or the 1% significance levels. When using the 5% significance level, one concludes that the experimental treatment has a real effect if chance alone would produce a difference as large or larger than the one obtained only 5% of the time or less. A good portion of this text is devoted to methods used for figuring out the probability of outcomes when chance alone is operating.

**Null Hypothesis**

The null hypothesis is an hypothesis about a population parameter. The purpose of hypothesis testing is to test the viability of the null hypothesis in the light of experimental data. Depending on the data, the null hypothesis either will or will not be rejected as a viable possibility.

Consider a researcher interested in whether the time to respond to a tone is affected by the consumption of alcohol. The null hypothesis is that \( \mu_1 - \mu_2 = 0 \) where \( \mu_1 \) is the mean time to respond after consuming alcohol and \( \mu_2 \) is the mean time to respond otherwise. Thus, the null hypothesis concerns the parameter \( \mu_1 - \mu_2 \) and the null hypothesis is that the parameter equals zero.

The null hypothesis is often the reverse of what the experimenter actually believes; it is put forward to allow the data to contradict it. In the experiment on the effect of alcohol, the experimenter probably expects alcohol to have a harmful effect. If the experimental data show a sufficiently large effect of alcohol, then the null hypothesis that alcohol has no effect can be rejected.

It should be stressed that researchers very frequently put forward a null hypothesis in the hope that they can discredit it. For a second example, consider an educational researcher who designed a new way to teach a particular concept in science, and wanted to test experimentally whether this new method worked better than the existing method. The researcher would design an experiment comparing the two methods. Since the null hypothesis would be that there is no difference between the two methods, the researcher would be hoping to reject the null hypothesis and conclude that the method he or she developed is the better of the two.

The symbol \( H_0 \) is used to indicate the null hypothesis. For the example just given, the null hypothesis would be designated by the following symbols:

\[
H_0: \mu_1 - \mu_2 = 0
\]
The null hypothesis is typically a hypothesis of no difference as in this example where it is the hypothesis of no difference between population means. That is why the word "null" in "null hypothesis" is used -- it is the hypothesis of no difference.

Despite the "null" in "null hypothesis," there are occasions when the parameter is not hypothesized to be 0. For instance, it is possible for the null hypothesis to be that the difference between population means is a particular value. Or, the null hypothesis could be that the mean SAT score in some population is 600. The null hypothesis would then be stated as: $H_0: \mu = 600$. Although the null hypotheses discussed so far have all involved the testing of hypotheses about one or more population means, null hypotheses can involve any parameter. An experiment investigating the correlation between job satisfaction and performance on the job would test the null hypothesis that the population correlation ($\rho$) is 0. Symbolically, $H_0: \rho = 0$.

Some possible null hypotheses are given below:

- $H_0: \mu = 0$
- $H_0: \mu = 10$
- $H_0: \pi = 0.5$
- $H_0: \pi_1 - \pi_2 = 0$
- $H_0: \mu_1 = \mu_2 = \mu_3$
- $H_0: \rho_1 - \rho_2 = 0$

When a one-tailed test is conducted, the null hypothesis includes the direction of the effect. A one-tailed test of the differences between means might test the null hypothesis that $\mu_1 - \mu_2 \geq 0$. If $M_1 - M_2$ were much less than 0 then the null hypothesis would be rejected in favor of the alternative hypothesis: $\mu_1 - \mu_2 < 0$.

**Steps in Hypothesis Testing**

The basic logic of hypothesis testing has been presented somewhat informally in the sections on "Ruling out chance as an explanation" and the "Null hypothesis." In this section the logic will be presented in more detail and more formally.

1. The first step in hypothesis testing is to specify the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$). If the research concerns whether one method of presenting pictorial stimuli leads to better recognition than another, the null hypothesis would most likely be that there is no difference between methods ($H_0: \mu_1 - \mu_2 = 0$). The alternative hypothesis would be $H_1: \mu_1 \neq \mu_2$. If the research concerned the correlation between grades and SAT scores, the null hypothesis would most likely be that there is no correlation ($H_0: \rho = 0$). The alternative hypothesis would be $H_1: \rho \neq 0$.

2. The next step is to select a significance level. Typically the 0.05 or the 0.01 level is used.

3. The third step is to calculate a statistic analogous to the parameter specified by the null hypothesis. If the null hypothesis were defined by the parameter $\mu_1 - \mu_2$, then the statistic $M_1 - M_2$ would be computed.

4. The fourth step is to calculate the probability value (often called the p value). The p value is the probability of obtaining a statistic as different or more different from the parameter specified in the null hypothesis as the statistic computed from the data. The calculations are made assuming that the null hypothesis is true. (click here for a concrete example)
5. The probability value computed in Step 4 is compared with the significance level chosen in Step 2. If the probability is less than or equal to the significance level, then the null hypothesis is rejected; if the probability is greater than the significance level then the null hypothesis is not rejected. When the null hypothesis is rejected, the outcome is said to be "statistically significant" when the null hypothesis is not rejected then the outcome is said be "not statistically significant."

6. If the outcome is statistically significant, then the null hypothesis is rejected in favor of the alternative hypothesis. If the rejected null hypothesis were that $μ_1 - μ_2 = 0$, then the alternative hypothesis would be that $μ_1 ≠ μ_2$. If $M_1$ were greater than $M_2$ then the researcher would naturally conclude that $μ_1 ≥ μ_2$. (Click here to see why you can conclude more than $μ_1 ≠ μ_2$)

7. The final step is to describe the result and the statistical conclusion in an understandable way. Be sure to present the descriptive statistics as well as whether the effect was significant or not. For example, a significant difference between a group that received a drug and a control group might be described as follow:

Subjects in the drug group scored significantly higher (M = 23) than did subjects in the control group (M = 17), t(18) = 2.4, p = 0.027.

The statement that "t(18) = 2.4" has to do with how the probability value (p) was calculated. A small minority of researchers might object to two aspects of this wording. First, some believe that the significance level rather than the probability level should be reported. The argument for reporting the probability value is presented in another section. Second, since the alternative hypothesis was stated as $μ_1 ≠ μ_2$, some might argue that it can only be concluded that the population means differ and not that the population mean for the drug group is higher than the population mean for the control group.

This argument is misguided. Intuitively, there are strong reasons for inferring that the direction of the difference in the population is the same as the difference in the sample. There is also a more formal argument. A non significant effect might be described as follows:

Although subjects in the drug group scored higher (M = 23) than did subjects in the control group, (M = 20), the difference between means was not significant, t(18) = 1.4, p = 0.179.

It would not have been correct to say that there was no difference between the performance of the two groups. There was a difference. It is just that the difference was not large enough to rule out chance as an explanation of the difference. It would also have been incorrect to imply that there is no difference in the population. Be sure not to accept the null hypothesis.

At this point you may wish to see a concrete example of using these seven steps in hypothesis testing. If so, jump to the section on "Tests of $μ$, $σ$ known."

Why the Null Hypothesis is Not Accepted

A null hypothesis is not accepted just because it is not rejected. Data not sufficient to show convincingly that a difference between means is not zero do not prove that the difference is zero. Such data may even suggest that the null hypothesis is false but not be strong enough to make a convincing case that the null hypothesis is false. For example, if the probability value were 0.15, then one would not be ready to present one’s case that the null hypothesis is false to the (properly) skeptical scientific community. More convincing data would be needed to do that. However, there would be no basis to conclude that the null hypothesis is true. It may or may not be true, there just is not strong enough evidence to reject it. Not even in cases where there is no evidence that the null hypothesis is false is it valid to conclude the null hypothesis is true. If the null hypothesis is that $μ_1 - μ_2$ is zero then the hypothesis is that the difference is exactly zero. No experiment can distinguish between the case of no difference
between means and an extremely small difference between means. If data are consistent with the null hypothesis, they are also consistent with other similar hypotheses.

Thus, if the data do not provide a basis for rejecting the null hypothesis that \( \mu_1 - \mu_2 = 0 \) then they almost certainly will not provide a basis for rejecting the hypothesis that \( \mu_1 - \mu_2 = 0.001 \). The data are consistent with both hypotheses. When the null hypothesis is not rejected then it is legitimate to conclude that the data are consistent with the null hypothesis. It is not legitimate to conclude that the data support the acceptance of the null hypothesis since the data are consistent with other hypotheses as well. In some respects, rejecting the null hypothesis is comparable to a jury finding a defendant guilty. In both cases, the evidence is convincing beyond a reasonable doubt. Failing to reject the null hypothesis is comparable to a finding of not guilty. The defendant is not declared innocent. There is just not enough evidence to be convincing beyond a reasonable doubt. In the judicial system, a decision has to be made and the defendant is set free. In science, no decision has to be made immediately. More experiments are conducted.

One experiment might provide data sufficient to reject the null hypothesis, although no experiment can demonstrate that the null hypothesis is true. Where does this leave the researcher who wishes to argue that a variable does not have an effect? If the null hypothesis cannot be accepted, even in principle, then what type of statistical evidence can be used to support the hypothesis that a variable does not have an effect. The answer lies in relaxing the claim a little and arguing not that a variable has no effect whatsoever but that it has, at most, a negligible effect. This can be done by constructing a confidence interval around the parameter value.

Consider a researcher interested in the possible effectiveness of a new psychotherapeutic drug. The researcher conducted an experiment comparing a drug-treatment group to a control group and found no significant difference between them. Although the experimenter cannot claim the drug has no effect, he or she can estimate the size of the effect using a confidence interval. If \( \mu_1 \) were the population mean for the drug group and \( \mu_2 \) were the population mean for the control group, then the confidence interval would be on the parameter \( \mu_1 - \mu_2 \).

Assume the experiment measured "well being" on a 50 point scale (with higher scores representing more well being) that has a standard deviation of 10. Further assume the 99% confidence interval computed from the experimental data was:

\[-0.5 \leq \mu_1 - \mu_2 \leq 1\]

This says that one can be confident that the mean "true" drug treatment effect is somewhere between -0.5 and 1. If it were -0.5 then the drug would, on average, be slightly detrimental; if it were 1 then the drug would, on average, be slightly beneficial. But, how much benefit is an average improvement of 1? Naturally that is a question that involves characteristics of the measurement scale. But, since 1 is only 0.10 standard deviations, it can be presumed to be a small effect. The overlap between two distributions whose means differ by 0.10 standard deviations is shown below. Although the blue distribution is

![Graph showing confidence intervals](https://example.com/confidence-interval-graph.png)

slightly to the right of the red distribution, the overlap is almost complete.
So, the finding that the maximum difference that can be expected (based on a 99% confidence interval) is itself a very small difference would allow the experimenter to conclude that the drug is not effective. The claim would not be that it is totally ineffective, but, at most, its effectiveness is very limited.

The Precise Meaning of the Probability Value

There is often confusion about the precise meaning of the probability computed in a significance test. As stated in Step 4 of the steps in hypothesis testing, the null hypothesis ($H_0$) is assumed to be true. The difference between the statistic computed in the sample and the parameter specified by $H_0$ is computed and the probability of obtaining a difference this large or larger is calculated. This probability value is the probability of obtaining data as extreme or more extreme than the current data (assuming $H_0$ is true). It is not the probability of the null hypothesis itself. Thus, if the probability value is 0.005, this does not mean that the probability that the null hypothesis is true is .005. It means that the probability of obtaining data as different or more different from the null hypothesis as those obtained in the experiment is 0.005.

The inferential step to conclude that the null hypothesis is false goes as follows: The data (or data more extreme) are very unlikely given that the null hypothesis is true. This means that: (1) a very unlikely event occurred or (2) the null hypothesis is false. The inference usually made is that the null hypothesis is false.

To illustrate that the probability is not the probability of the hypothesis, consider a test of a person who claims to be able to predict whether a coin will come up heads or tails. One should take a rather skeptical attitude toward this claim and require strong evidence to believe in its validity. The null hypothesis is that the person can predict correctly half the time ($H_0$: $\pi = 0.5$). In the test, a coin is flipped 20 times and the person is correct 11 times. If the person has no special ability ($H_0$ is true), then the probability of being correct 11 or more times out of 20 is 0.41. Would someone who was originally skeptical now believe that there is only a 0.41 chance that the null hypothesis is true? They almost certainly would not since they probably originally thought $H_0$ had a very high probability of being true (perhaps as high as 0.9999). There is no logical reason for them to decrease their belief in the validity of the null hypothesis since the outcome was perfectly consistent with the null hypothesis.

The proper interpretation of the test is as follows: A person made a rather extraordinary claim and should be able to provide strong evidence in support of the claim if the claim is to be believed. The test provided data consistent with the null hypothesis that the person has no special ability since a person with no special ability would be able to predict as well or better than 40% of the time. Therefore, there is no compelling reason to believe the extraordinary claim. However, the test does not prove the person cannot predict better than chance; it simply fails to provide evidence that he or she can. The probability that the null hypothesis is true is not determined by the statistical analysis conducted as part of hypothesis testing. Rather, the probability computed is the probability of obtaining data as different or more different from the null hypothesis (given that the null hypothesis is true) as the data actually obtained.

At What Level is $H_0$ Really Rejected?

According to one view of hypothesis testing, the significance level should be specified before any statistical calculations are performed. Then, when the probability ($p$) is computed from a significance test, it is compared with the significance level. The null hypothesis is rejected if $p$ is at or below the significance level; it is not rejected if $p$ is above the significance level. The degree to which $p$ ends up being above or below the significance level does not matter. The null hypothesis either is or is not rejected at the previously stated significance level. Thus, if an experimenter originally stated that he or she was using the 0.05 significance level and $p$ was subsequently calculated to be 0.042, then the person would reject the null hypothesis at the 0.05 level. If $p$ had been 0.0001 instead of 0.042 then the null hypothesis would still be rejected at the 0.05 level. The experimenter would not have any basis to be more confident that the null hypothesis was false with a $p$ of 0.0001 than with a $p$ of 0.041. Similarly, if the $p$ had been 0.051 then the experimenter would fail to reject the null hypothesis.
He or she would have no more basis to doubt the validity of the null hypothesis than if \( p \) had been 0.482. The conclusion would be that the null hypothesis could not be rejected at the 0.05 level. In short, this approach is to specify the significance level in advance and use \( p \) only to determine whether or not the null hypothesis can be rejected at the stated significance level.

Many statisticians and researchers find this approach to hypothesis testing not only too rigid, but basically illogical. Who in their right mind would not have more confidence that the null hypothesis is false with a \( p \) of 0.0001 then with a \( p \) of 0.042? The less likely the obtained results (or more extreme results) under the null hypothesis, the more confident one should be that the null hypothesis is false. The null hypothesis should not be rejected once and for all. The possibility that it was falsely rejected is always present, and, all else being equal, the lower the \( p \) value, the lower this possibility.

According to this view, research reports should not contain the values of \( p \), only whether or not the values were significant (at or below the significance level). It is much more reasonable to report the \( p \) values. That way each reader can make up his or her mind about just how convinced they are that the null hypothesis is false. For more discussion see Wilkenson et al. (1999).

**Statistical and Practical Significance**

It is important not to confuse the confidence with which the null hypothesis can be rejected with size of the effect. To make this point concrete, consider a researcher assigned the task of determining whether the video display used by travel agents for booking airline reservations should be in color or in black and white. Market research had shown that travel agencies were primarily concerned with the speed with which reservations can be made. Therefore, the question was whether color displays allow travel agents to book reservations faster. Market research had also shown that in order to justify the higher price of color displays, they must be faster by an average of at least 10 seconds per transaction. Fifty subjects were tested with color displays and 50 subjects were tested with black and white displays. Subjects were slightly faster at making reservations on a color display (\( M = 504.7 \) seconds) than on a black and white display (\( M = 508.2 \) seconds). although the difference is small, it was statistically significant at the .05 significance level. Box plots of the data are shown below.

The 95% confidence interval on the difference between means is:

\[-7.0 \leq \mu_{\text{color}} - \mu_{\text{black & white}} \leq -0.1\]

which means that the experimenter can be confident that the color display is between 7.0 seconds and 0.1 seconds faster. Clearly, the difference is not big enough to justify the more expensive color displays. Even the upper limit of the 95% confidence interval (seven seconds) is below the minimum needed to justify the cost (10 seconds). Therefore, the experimenter could feel confident in his or her recommendation that the black and white displays should be used. The fact that the color displays were significantly faster does not mean that they were much faster. It just means that the experimenter can reject the null hypothesis that there is no difference between the displays.
The experimenter presented this conclusion to management but management did not accept it. The color displays were so dazzling that despite the statistical analysis, they could not believe that color did not improve performance by at least 10 seconds. The experimenter decided to do the experiment again, this time using 100 subjects for each type of display. The results of the second experiment were very similar to the first. Subjects were slightly faster at making reservations on a color display (M = 504.7 seconds) than on a black and white display (M = 508.1 seconds). This time the difference was significant at the 0.01 level rather than the 0.05 level found in the first experiment. Despite the fact that the size of the difference between means was no larger, the difference was "more significant" due to the larger sample size used. If the population difference is zero, then a sample difference of 3.4 or larger with a sample size of 100 is less likely than a sample difference of 3.5 or larger with a sample size of 50.

The 95% confidence interval on the difference between means is:

\[-5.8 \leq \mu_{\text{color}} - \mu_{\text{black & white}} \leq -0.9\]

and the 99% interval is:

\[-6.6 \leq \mu_{\text{color}} - \mu_{\text{black & white}} \leq -0.1\]

Therefore, despite the finding of a "more significant" difference between means, the experimenter can be even more certain that the color displays are only slightly better than the black and white displays. The second experiment shows conclusively that the difference is less than 10 seconds.

This example was used to illustrate the following points: (1) an effect that is statistically significant is not necessarily large enough to be of practical significance and (2) the smaller of two effects can be "more significant" than the larger. Be careful how you interpret findings reported in the media. If you read that a particular diet lowered cholesterol significantly, this does not necessarily mean that the diet lowered cholesterol enough to be of any health value. It means that the effect on cholesterol in the population is greater than zero.

**Type I and II errors**

There are two kinds of errors that can be made in significance testing: (1) a true null hypothesis can be incorrectly rejected and (2) a false null hypothesis can fail to be rejected. The former error is called a Type I error and the latter error is called a Type II error. These two types of errors are defined in the table.

<table>
<thead>
<tr>
<th>Statistical Decision</th>
<th>True State of the Null Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$ True</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
</tr>
<tr>
<td>Do not Reject $H_0$</td>
<td>Correct</td>
</tr>
</tbody>
</table>

The probability of a Type I error is designated by the Greek letter alpha ($\alpha$) and is called the Type I error rate; the probability of a Type II error (the Type II error rate) is designated by the Greek letter beta ($\beta$). A Type II error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost. It is not an error in the sense that an incorrect conclusion was drawn since no conclusion is drawn when the null hypothesis is not rejected.
A Type I error, on the other hand, is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true. Therefore, Type I errors are generally considered more serious than Type II errors. The probability of a Type I error ($\alpha$) is called the **significance level** and is set by the experimenter. There is a tradeoff between Type I and Type II errors. The more an experimenter protects himself or herself against Type I errors by choosing a low level, the greater the chance of a Type II error. Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected. However, it increases the chance that a false null hypothesis will not be rejected, thus lowering **power**. The Type I error rate is almost always set at .05 or at .01, the latter being more conservative since it requires stronger evidence to reject the null hypothesis at the .01 level than at the .05 level.

### One- and Two-Tailed Tests

In the section on "Steps in hypothesis testing" the fourth step involves calculating the probability that a statistic would differ as much or more from parameter specified in the null hypothesis as does the statistic obtained in the experiment. This statement implies that a difference in either direction would be counted. That is, if the null hypothesis were:

$$H_0: \mu_1 - \mu_2 = 0$$

and the value of the statistic $M_1 - M_2$ were +5, then the probability of $M_1 - M_2$ differing from zero by five or more (in either direction) would be computed. In other words, **probability value** would be the probability that either $M_1 - M_2 \geq 5$ or $M_1 - M_2 \leq -5$.

Assume that the figure shown below is the sampling distribution of $M_1 - M_2$.

![Sampling Distribution](image)

The figure shows that the probability of a value of +5 or more is 0.036 and that the probability of a value of -5 or less is .036. Therefore the probability of a value either greater than or equal to +5 or less than or equal to -5 is 0.036 + 0.036 = 0.072.

A probability computed considering differences in both directions is called a "two-tailed" probability. The name makes sense since both tails of the sampling distribution are considered.

There are situations in which an experimenter is concerned only with differences in one direction. For example, an experimenter may be concerned with whether or not $\mu_1 - \mu_2$ is greater than zero. However, if $\mu_1 - \mu_2$ is not greater than zero, the experimenter may not care whether it equals zero or is less than zero. For instance, if a new drug treatment is developed, the main issue is whether or not it is better than a placebo. If the treatment is not better than a placebo, then it will not be used. It does not really matter whether or not it is worse than the placebo.

When only one direction is of concern to an experimenter, then a "one-tailed" test can be performed. If an experimenter were only concerned with whether or not $\mu_1 - \mu_2$ is greater than zero, then the one-tailed test would involve calculating the probability of obtaining a statistic as great or greater than the one obtained in the experiment.

In the example, the one-tailed probability would be the probability of obtaining a value of $M_1 - M_2$ greater than or equal to five given that the difference between population means is zero.
The shaded area in the figure is greater than five. The figure shows that the one-tailed probability is 0.036.

It is easier to reject the null hypothesis with a one-tailed than with a two-tailed test as long as the effect is in the specified direction. Therefore, one-tailed tests have lower Type II error rates and more power than do two-tailed tests. In this example, the one-tailed probability (0.036) is below the conventional significance level of 0.05 whereas the two-tailed probability (0.072) is not. Probability values for one-tailed tests are one half the value for two-tailed tests as long as the effect is in the specified direction.

One-tailed and two-tailed tests have the same Type I error rate. One-tailed tests are sometimes used when the experimenter predicts the direction of the effect in advance. This use of one-tailed tests is questionable because the experimenter can only reject the null hypothesis if the effect is in the predicted direction. If the effect is in the other direction, then the null hypothesis cannot be rejected no matter how strong the effect is. A skeptic might question whether the experimenter would really fail to reject the null hypothesis if the effect were strong enough in the wrong direction. Frequently the most interesting aspect of an effect is that it runs counter to expectations. Therefore, an experimenter who committed himself or herself to ignoring effects in one direction may be forced to choose between ignoring a potentially important finding and using the techniques of statistical inference dishonestly. One-tailed tests are not used frequently. Unless otherwise indicated, a test should be assumed to be two-tailed.

Confidence Intervals & Hypothesis Testing

There is an extremely close relationship between confidence intervals and hypothesis testing. When a 95% confidence interval is constructed, all values in the interval are considered plausible values for the parameter being estimated. Values outside the interval are rejected as relatively implausible. If the value of the parameter specified by the null hypothesis is contained in the 95% interval then the null hypothesis cannot be rejected at the 0.05 level. If the value specified by the null hypothesis is not in the interval then the null hypothesis can be rejected at the 0.05 level. If a 99% confidence interval is constructed, then values outside the interval are rejected at the 0.01 level.

Imagine a researcher wishing to test the null hypothesis that the mean time to respond to an auditory signal is the same as the mean time to respond to a visual signal. The null hypothesis therefore is:

\[ \mu_{\text{visual}} - \mu_{\text{auditory}} = 0. \]

Ten subjects were tested in the visual condition and their scores (in milliseconds) were: 355, 421, 299, 460, 600, 580, 474, 511, 550, and 586.

Ten subjects were tested in the auditory condition and their scores were: 275, 320, 278, 360, 430, 520, 464, 311, 529, and 326.

The 95% confidence interval on the difference between means is:

\[ 9 \leq \mu_{\text{visual}} - \mu_{\text{auditory}} \leq 196. \]
Therefore only values in the interval between 9 and 196 are retained as plausible values for the difference between population means. Since zero, the value specified by the null hypothesis, is not in the interval, the null hypothesis of no difference between auditory and visual presentation can be rejected at the 0.05 level. The probability value for this example is 0.034. Any time the parameter specified by a null hypothesis is not contained in the 95% confidence interval estimating that parameter, the null hypothesis can be rejected at the 0.05 level or less. Similarly, if the 99% interval does not contain the parameter then the null hypothesis can be rejected at the 0.01 level. The null hypothesis is not rejected if the parameter value specified by the null hypothesis is in the interval since the null hypothesis would still be plausible.

However, since the null hypothesis would be only one of an infinite number of values in the confidence interval, accepting the null hypothesis is not justified.

There are many arguments against accepting the null hypothesis when it is not rejected. The null hypothesis is usually a hypothesis of no difference. Thus null hypotheses such as:

\[ \mu_1 - \mu_2 = 0 \]
\[ \pi_1 - \pi_2 = 0 \]

in which the hypothesized value is zero are most common. When the hypothesized value is zero then there is a simple relationship between hypothesis testing and confidence intervals:

If the interval contains zero then the null hypothesis cannot be rejected at the stated level of confidence. If the interval does not contain zero then the null hypothesis can be rejected.

This is just a special case of the general rule stating that the null hypothesis can be rejected if the interval does not contain the hypothesized value of the parameter and cannot be rejected if the interval contains the hypothesized value.

Since zero is contained in the interval, the null hypothesis that \( \mu_1 - \mu_2 = 0 \) cannot be rejected at the 0.05 level since zero is one of the plausible values of \( \mu_1 - \mu_2 \). The interval contains both positive and negative numbers and therefore \( \mu_1 \) may be either larger or smaller than \( \mu_2 \). None of the three possible relationships between \( \mu_1 \) and \( \mu_2 \):

\[ \mu_1 - \mu_2 = 0, \]
\[ \mu_1 - \mu_2 > 0, \text{ and} \]
\[ \mu_1 - \mu_2 < 0 \]

can be ruled out. The data are very inconclusive. Whenever a significance test fails to reject the null hypothesis, the direction of the effect (if there is one) is unknown.

Now, consider the 95% confidence interval:

\[ 6 \leq \mu_1 - \mu_2 \leq 15. \]

Since zero is not in the interval, the null hypothesis that \( \mu_1 - \mu_2 = 0 \) can be rejected at the 0.05 level. Moreover, since all the values in the interval are positive, the direction of the effect can be inferred: \( \mu_1 > \mu_2 \).

Whenever a significance test rejects the null hypothesis that a parameter is zero, the confidence interval on that
parameter will not contain zero. Therefore either all the values in the interval will be positive or all the values in the interval will be negative. In either case, the direction of the effect is known.

**Following a Non-Significant Finding**

An experimenter wishes to test the hypothesis that sleep deprivation increases reaction time. An experiment is conducted comparing the reaction times of 10 people who have missed a night's sleep with 10 control subjects. Although the sleep-deprived subjects react more slowly, the difference is not significant, $p = 0.10$, two tailed. Should the experimenter be more or less certain that sleep deprivation increases reaction time than he or she was before the experiment was conducted? The naive approach is to argue that there was no significant difference between the sleep-deprived and the control group so the experimenter should now be less confident that sleep deprivation increases reaction time. This argument implicitly assumes that the null hypothesis should be accepted when it is not rejected. A more straightforward and more correct approach is to consider that the experimenter expected the sleep-deprived group to have slower reaction time, and they did. The experimenter's prediction was correct. It is just that the difference was not large enough to rule out chance as an explanation.

The experimenter's belief that sleep deprivation increases reaction time should be strengthened. Nonetheless, the data are not strong enough to convince a skeptic, so no attempt should be made to publish the results. Instead, the experimenter repeated the experiment. Once again, the sleep-deprived group had slower reaction times. Based on the results of the first experiment, the experimenter conducted a one-tailed test in the second experiment. However, it was not significant, $p = 0.08$, one tailed.

The naive interpretation of the two experiments is that the experimenter tried twice to find a significant result and failed both times. With each failure, the strength of the experimenter's case that sleep deprivation increases reaction time is weakened.

The correct interpretation is that in two out of two experiments the sleep-deprived subjects had the slower reaction times. The experimenter's case is strengthened by each experiment. Moreover, there are methods for combining the probability values across experiments. For these two experiments, the combined probability is 0.047.

Taken together, these two experiments provide relatively good evidence ($p<0.05$) that sleep deprivation increases reaction time. Naturally, the more times an outcome is replicated, the more believable the outcome. Assume the experimenter did the experiment six times and the sleep-deprived group was slower each time. If the probability values were: 0.10, 0.08, 0.12, 0.07, 0.19, and 0.13, the combined probability would be 0.009. Therefore, six non-significant probabilities combine to produce one highly significant probability. Compare this with the naive view that would state that the experimenter's hypothesis is almost certainly incorrect since not one of the six experiments found a significant difference between sleep-deprived and control subjects. Such a view implicitly accepts the null hypothesis, a serious error.

In summary, a non-significant result means that the data are inconclusive. Collecting additional data may be all that is needed to reject the null hypothesis. If the null hypothesis is true, then additional data will make clear that the effect is at most small. The additional data can never prove that the effect is nonexistent.

**For apply the exercises at following link:**

[http://davidmlane.com/hyperstat/logic_hypothesis.html](http://davidmlane.com/hyperstat/logic_hypothesis.html)
CHAPTER 10

Hypothesis Testing with Standard Error
General Formula for Testing Hypotheses

The formula shown below is used for testing hypotheses about a parameter.

\[ z = \frac{\text{statistic} - \text{hypothesized value}}{\text{standard error of the statistic}} \]

The "statistic" is an estimate of the parameter in question. The "hypothesized value" is the value of the parameter specified in the null hypothesis. The standard error of the statistic is assumed to be known and the sampling distribution of the statistic is assumed to be normal.

Consider an experiment designed to test the null hypothesis that \( \mu = 10 \). The test would be conducted with the following formula:

\[ z = \frac{M - 10}{\sigma_M} \]

where \( M \) (the statistic) is the sample mean, 10 is the hypothesized value of \( \mu \), and \( \sigma_M \) is standard error of the mean. Once \( z \) is determined, the probability value can be found using a z table. For example, if \( M = 15 \) and \( \sigma_M = 2 \), then \( z \) would be \( (15-10)/2 = 2.5 \). The two-tailed probability value would be 0.0124. The one-tailed probability value would be 0.0124/2 = 0.0062.

The \( t \) (Student's t) distribution is used if the standard error has to be estimated from the data. The formula then becomes:

\[ t = \frac{\text{statistic} - \text{hypothesized value}}{\text{estimated standard error of the statistic}} \]

(The one exception to this rule you will encounter is that in tests of differences between proportions, \( z \) is used even when the standard error is estimated.) For this example, the formula is:

\[ t = \frac{M - 10}{s_M} \]

where \( s_M \) is the estimated standard error of the mean. If \( M = 15 \) and \( s_M = 2 \), then \( t = 2.5 \). If \( N \) were 12 then the degrees of freedom would be 11. A t table can be used to calculate that the two-tailed probability value for a \( t \) of 2.5 with 11 df is 0.0295. The one-tailed probability is 0.0148.

Why the Formula Works

The formula is basically a method for calculating the area under the normal curve. If the mean and standard deviation of the sampling distribution of a statistic are known, it is easy to calculate the probability of a statistic being greater than or less than a specific value. That is what is done in hypothesis testing.

Tests of \( \mu \), Standard Deviation Known

This section explains how to compute a significance test for the mean of a normally-distributed variable for which the population standard deviation (\( \sigma \)) is known. In practice, the standard deviation is rarely known.
However, learning how to compute a significance test when the standard deviation is known is an excellent introduction to how to compute a significance test in the more realistic situation in which the standard deviation has to be estimated.

1. The first step in hypothesis testing is to specify the null hypothesis and the alternate hypothesis. In testing hypotheses about µ, the null hypothesis is a hypothesized value of µ. Suppose the mean score of all 10-year old children on an anxiety scale were 7. If a researcher were interested in whether 10-year old children with alcoholic parents had a different mean score on the anxiety scale, then the null and alternative hypotheses would be:

   \[ H_0: \mu_{\text{alcoholic}} = 7 \]

   \[ H_1: \mu_{\text{alcoholic}} \neq 7 \]

2. The second step is to choose a significance level. Assume the 0.05 level is chosen.

3. The third step is to compute the mean. Assume \( M = 8.1 \).

The fourth step is to compute \( p \), the probability (or probability value) of obtaining a difference between \( M \) and the hypothesized value of \( \mu \) (7.0) as large or larger than the difference obtained in the experiment. Applying the general formula to this problem,

\[
z = \frac{M - \mu}{\sigma_M} = \frac{8.1 - 7.0}{\sigma_M}
\]

The sample size (N) and the population standard deviation (\( \sigma \)) are needed to calculate \( \sigma_M \). Assume that \( N = 16 \) and \( \sigma = 2.0 \). Then,

\[
\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{2}{\sqrt{16}} = 0.50
\]

and

\[
\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{2}{\sqrt{16}} = 0.50
\]

A \( z \) table can be used to compute the probability value, \( p = 0.028 \). A graph of the sampling distribution of the mean is shown below. The area 8.1 - 7.0 = 1.1 or more units from the mean (in both directions) is shaded. The shaded area is 0.028 of the total area.
In step five, the probability computed in Step 4 is compared to the significance level stated in Step 2. Since the probability value (0.028) is less than the significance level (0.05) the effect is statistically significant.

Since the effect is significant, the null hypothesis is rejected. It is concluded that the mean anxiety score of 10-year-old children with alcoholic parents is higher than the population mean.

The results might be described in a report as follows:

The mean score of children of alcoholic parents (M = 8.1) was significantly higher than the population mean (μ = 7.0), z = 2.2, p = 0.028.

Summary of Computational Steps

1. Specify the null hypothesis and an alternative hypothesis.

2. Compute \( M = \frac{\Sigma X}{N} \).

3. Compute \( \sigma_M = \frac{\sigma}{\sqrt{N}} \).

4. Compute \( z = \frac{M - \mu}{\sigma_M} \), where \( M \) is the sample mean and \( \mu \) is the hypothesized value of the population.

5. Use a z table to determine p from z

ASSUMPTIONS

1. Normal distribution

2. Scores are independent

3. \( \sigma \) is known.

Tests of \( \mu \), Standard Deviation Estimated

Rarely does a researcher wishing to test a hypothesis about a population’s mean already know the population’s standard deviation (\( \sigma \)). Therefore, testing hypotheses about means almost always involves the estimation \( \sigma \). When \( \sigma \) is known, the formula:

\[
\sigma_M = \frac{\sigma}{\sqrt{N}}
\]

is used for a confidence interval. When \( \sigma \) is not known,
is used as an estimate of $\sigma_M$ (s is an estimate of the standard deviation and N is the sample size). When $\sigma_M$ is estimated by $s_M$, the significance test uses the \textit{t distribution} instead of the normal distribution.

Suppose a researcher wished to test whether the mean score of fifth graders on a test of reading achievement in his or her city differed from the national mean of 76. The researcher randomly sampled the scores of 20 students. The scores are shown below.

72  69  98  87
78  76  78  66
85  97  84  86
88  76  79  82
82  91  69  74

The first step in hypothesis testing is to specify the \textit{null hypothesis} and an \textit{alternative hypothesis}. When testing hypotheses about $\mu$, the null hypothesis is an hypothesized value of $\mu$. In this example, the null hypothesis is $\mu = 76$. The alternative hypothesis is: $\mu \neq 76$.

4. The second step is to choose a \textit{significance level}. Assume the .05 level is chosen.

5. The third step is to compute the \textit{mean}. For this example, $M = 80.85$.

6. The fourth step is to compute $p$, the probability (or \textit{probability value}) of obtaining a difference between $M$ and the hypothesized value of $\mu$ (76) as large or larger than the difference obtained in the experiment. Applying the \textit{general formula} to this problem,

\[
t = \frac{M - \mu}{s_M} = \frac{80.85 - 76}{1.984} = 2.44.
\]

The estimated standard error of the mean ($s_M$) was computed using the formula:

\[
s_M = \frac{s}{\sqrt{N}}
\]

= 8.87/4.47 = 1.984 where $s$ is the estimated \textit{standard deviation} and $N$ is the \textit{sample size}.

The probability value for $t$ can be determined using a \textit{t table}. The degrees of freedom for $t$ is equal to the degrees of freedom for the estimate of $\sigma_M$ which is $N - 1 = 20 - 1 = 19$. A \textit{t table} can be used to calculate that the two-tailed probability value of a $t$ of 2.44 with 19 df is 0.025.

7. In the fifth step, the probability computed in Step 4 is compared to the significance level stated in Step 2. Since the probability value (0.025) is less than the significance level (0.05) the effect is statistically significant.

8. Since the effect is significant, the null hypothesis is rejected. It is concluded that the mean reading achievement score of children in the city in question is higher than the population mean.

9. A report of this experimental result might be as follows:

The mean reading-achievement score of fifth grade children in the sample ($M = 80.85$) was significantly higher than the mean reading-achievement score nationally ($\mu = 76$), $t(19) = 2.44$, $p = 0.025$. 

\[
http://davidmlane.com/hyperstat/index.html
\]
The expression "t(19) = 2.44" means that a t test with 19 degrees of freedom was equal to 2.44. The probability value is given by "p = 0.025." Since it was not mentioned whether the test was one- or two-tailed, a two-tailed test is assumed.

**Summary of Computations**

1. Specify the null hypothesis and an alternative hypothesis.
2. Compute \( M = \frac{\sum X}{N} \).
3. Compute \( s_m = \frac{s}{\sqrt{N}} \).
4. Compute \( s_m = \frac{s}{\sqrt{N}} \).
5. Compute \( t = \frac{M - \mu}{s_m} \) where \( \mu \) is the hypothesized value of the population mean.
6. Compute \( df = N - 1 \).
7. Use a t table to compute p from t and df.

**Assumptions**

1. Normal distribution
2. Scores are independent

**Tests of Differences between Means, Independent Groups, Standard Deviation Estimated**

This section explains how to test the difference between group means for significance. The formulas are slightly simpler when the sample sizes are equal. These formulas are given first.

**Equal Sample Sizes**

An experiment was conducted comparing the memory of expert and novice chess players. The mean number of pieces correctly placed across several chess positions was computed for each subject. (These data are from the novices and tournament players in the dataset "Chess.") The question is whether the difference between the means of these two groups of subjects is statistically significant.

1. The first step is to specify the null hypothesis and an alternative hypothesis. For experiments testing differences between means, the null hypothesis is that the difference between means is some specified value. Usually the null hypothesis is that the difference is zero.

   For this example, the null and alternative hypotheses are:

   \[ H_0: \mu_1 - \mu_2 = 0 \]
   \[ H_1: \mu_1 - \mu_2 \neq 0 \]

2. The second step is to choose a significance level. Assume the .05 level is chosen.
3. The **third step** is to compute the difference between sample means ($M_d$). In this example, $M_I = 63.89$, $M_N = 46.79$ and $M_d = M_I - M_N = 17.10$.

4. The **fourth step** is to compute $p$, the probability (or **probability value**) of obtaining a difference between and the value specified by the null hypothesis (0) as large or larger than the difference obtained in the experiment. Applying the general **formula**,

$$t = \frac{M_d - (\mu_1 - \mu_2)}{s_{M_d}}$$

where $M_d$ is the difference between sample means, $\mu_1 - \mu_2$ is the difference between population means specified by the null hypothesis (usually zero), and $s_{M_d}$ is the estimated **standard error of the difference between means**.

The estimated standard error, $s_{M_d}$, is computed assuming that the variances in the two populations are **equal**. If the two sample sizes are equal ($n_1 = n_2$) then the population variance $\sigma^2$ (it is the same in both populations) is estimated by using the following formula:

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

where MSE (which stands for mean square error) is an estimate of $\sigma^2$. Once MSE is calculated, $s_{M_d}$ can be computed as follows:

$$s_{M_d} = \sqrt{\frac{2MSE}{n}}$$

where $n = n_1 = n_2$. This formula is derived from the formula for the standard error of the difference between means when the variance is **known**.

4. (continued) For the present example,

$$t = \frac{M_I - M_N - 0}{s_{M_d}}$$

$MSE = (81.54 + 244.03)/2 = 162.78$,

and

$$s_{M_d} = \frac{(2)(162.78)}{10} = 5.706$$

$$t = \frac{17.10 - 0}{5.706} = 2.99$$

The probability value for $t$ can be determined using a **t table**. The degrees of freedom for $t$ is equal to the degrees of freedom for MSE which is equal to $df = n_1 - 1 + n_2 -1 = 18$ or, $df = N - 2$ where $N = n_1 + n_2$ The probability is: $p = 0.008$. 
5. **In step 5**, the probability computed in Step 4 is compared to the significance level stated in Step 2. Since the probability value (0.008) is less than the significance level (0.05) the effect is significant.

6. Since the effect is significant, the null hypothesis is rejected. It is concluded that the mean memory score for experts is higher than the mean memory score for novices.

7. A report of this experimental result might be as follows:

   The mean number of pieces recalled by tournament players ($M_T = 63.89$) was significantly higher than the mean number of pieces recalled by novices ($M_N = 46.79$), $t(18) = 2.99$, $p = 0.008$.

   The expression "$t(18) = 2.99$" means that a t test with 18 degrees of freedom was equal to 2.99. The probability value is given by "$p = 0.008." Since it was not mentioned whether the test was one- or two-tailed, it is assumed the test was two tailed.

### Unequal Sample Sizes

The calculations in Step 4 are slightly more complex when $n_1 \neq n_2$. The first difference is that MSE is computed differently. If the two values of $s^2$ were simply averaged as they are for equal sample sizes, then the estimate based on the smaller sample size would count as much as the estimate based on the larger sample size. Instead the formula for MSE is:

$$MSE = \frac{SSE}{df}$$

where $df$ is the degrees of freedom ($n_1 - 1 + n_2 - 1$) and SSE is:

$$SSE = SSE_1 + SSE_2$$

$SSE_1 = \Sigma(X - M_1)^2$ where the $X$'s are from the first group (sample) and $M_1$ is the mean of the first group. Similarly, $SSE_2 = \Sigma(X - M_2)^2$ where the $X$'s are from the second group and $M_2$ is the mean of the second group.

The formula:

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

cannot be used without modification since there is not one value of $n$ but two: ($n_1$ and $n_2$). The solution is to use the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean) of the two sample sizes. The harmonic mean ($n_h$) of $n_1$ and $n_2$ is:

$$n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

The formula for the estimated standard error of the difference between means becomes:

$$s_{M_1} = \sqrt{\frac{2MSE}{n_h}}$$

The hypothetical data shown below are from an experimental group and a control group.
STATISTICS

http://davidmlane.com/hyperstat/index.html

<table>
<thead>
<tr>
<th>Experimental</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

\[ n_1 = 5, \quad n_2 = 4, \quad M_1 = 6, \quad M_2 = 5, \]

\[
SSE_1 = (3-6)^2 + (5-6)^2 + (7-6)^2 + (7-6)^2 + (8-6)^2 = 16,
\]

\[
SE_2 = (3-5)^2 + (4-5)^2 + (6-5)^2 + (7-5)^2 = 10
\]

\[
SSE = SSE_1 + SSE_2 = 16 + 10 = 26
\]

\[
df = n_1 - 1 + n_2 - 1 = 7
\]

\[
MSE = SSE/df = 26/7 = 3.71
\]

\[
S_{M_d} = \sqrt{\frac{2MSE}{n_h}} = \sqrt{\frac{(2)(3.71)}{4.44}} = 1.293
\]

\[
t = \frac{6 - 5}{1.293} = 0.77
\]

The p value associated with a t of 0.77 with 7 df is 0.47. Therefore, the difference between groups is not significant.

**Summary of Computations**

1. Specify the null hypothesis and an alternative hypothesis.
2. Compute \( M_d = M_1 - M_2 \)
3. Compute \( SSE_1 = \sum (X - M_1)^2 \) for Group 1 and \( SSE_2 = \sum (X - M_2)^2 \) for Group 2
4. Compute \( SSE = SSE_1 + SSE_2 \)
5. Compute \( df = N - 2 \) where \( N = n_1 + n_2 \)
6. Compute \( MSE = SSE/df \)
7. Compute:

\[
n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}}
\]

(If the sample sizes are equal then \( n_h = n_1 = n_2 \).)
8. Compute:
\[ s_{M_d} = \sqrt{\frac{2 \text{MSE}}{n}} \]

9. Compute:
\[ t = \frac{M_d - (\mu_1 - \mu_2)}{s_{M_d}} \]
where \( \mu_1 - \mu_2 \) is the difference between population means specified by the null hypothesis (and is usually 0).

10. Use a t table to compute p from t (step 9) and df (step 5).

Assumptions

1. The populations are normally distributed.
2. Variances in the two populations are equal.
3. Scores are independent (Each subject provides only one score)

Tests of Differences between Means, Dependent Means

When the same subjects are tested in two experimental conditions, scores in the two conditions are not independent because subjects who score well in one condition tend to score well in the other condition. This non-independence must be taken into account by the statistical analysis. The t test used when the scores are not independent is sometimes called a correlated t test and sometimes called a related-pairs t test. (Click here to see the advantage of testing the same subjects in both conditions.)

Consider an experimenter interested in whether the time it takes to respond to a visual signal is different from the time it takes to respond to an auditory signal. Ten subjects are tested with both the visual signal and with the auditory signal. (To avoid confounding with practice effects, half are in the auditory condition first and the other half are in the visual task first). The reaction times (in milliseconds) of the ten subjects in the two conditions are shown below.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Visual</th>
<th>Auditory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>420</td>
<td>380</td>
</tr>
<tr>
<td>2</td>
<td>235</td>
<td>230</td>
</tr>
<tr>
<td>3</td>
<td>280</td>
<td>300</td>
</tr>
<tr>
<td>4</td>
<td>360</td>
<td>260</td>
</tr>
<tr>
<td>5</td>
<td>305</td>
<td>295</td>
</tr>
<tr>
<td>6</td>
<td>215</td>
<td>190</td>
</tr>
<tr>
<td>7</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>8</td>
<td>460</td>
<td>410</td>
</tr>
<tr>
<td>9</td>
<td>345</td>
<td>330</td>
</tr>
<tr>
<td>10</td>
<td>375</td>
<td>380</td>
</tr>
</tbody>
</table>

The first step in testing the difference between means is to compute a difference score for each subject. The difference score is simply the difference between the two conditions. The difference scores are shown below:

<table>
<thead>
<tr>
<th>Sub</th>
<th>Visual</th>
<th>Auditory</th>
<th>Vis-Aud</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>420</td>
<td>380</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>235</td>
<td>230</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>280</td>
<td>300</td>
<td>-20</td>
</tr>
<tr>
<td>4</td>
<td>360</td>
<td>260</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>305</td>
<td>295</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>215</td>
<td>190</td>
<td>25</td>
</tr>
<tr>
<td>7</td>
<td>200</td>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>460</td>
<td>410</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>345</td>
<td>330</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>375</td>
<td>380</td>
<td>5</td>
</tr>
</tbody>
</table>
Notice that the difference between the visual mean (319.5) and the auditory mean (297.5) is the same as the mean of the difference scores (22). This will always be the case: The difference between the means will always equal the mean of the difference scores. The significance test of the difference between means consists of determining whether the mean of difference scores is significantly different from zero. If it is, then the difference between means is significant. The procedure for testing whether a mean is significantly different from zero is given in the section "Tests of μ, standard deviation estimated."

For this example,

\[ t = \frac{M - \mu}{s_M} = \frac{22 - 0}{10.88} = 2.02. \]

There are 10 - 1 = 9 degrees of freedom (Note that the sample size is the number of difference scores, not the total number of scores.) A t table can be used to find that the probability value is 0.074.

An alternate formula for \( s_M \) is:

\[ s_M = \sqrt{\frac{s_1^2 + s_2^2 - 2rs_1s_2}{n}} \]

where \( s_1 \) is the standard deviation of Condition 1, \( s_2 \) is the standard deviation of Condition 2, \( n \) is the number of subjects, and \( r \) is the correlation between scores in Conditions 1 and 2. For the example data, \( s_1 = 87.764, s_2 = 77.719, n = 10, \) and \( r = 0.9206 \) and therefore \( s_M = 10.88. \)

An example of a report of this result is shown below.
The mean time to respond to a visual stimulus (\( M = 319.5 \)) was longer than the mean time to respond to an auditory stimulus (\( M = 297.5 \)). However, this difference was not statistically significant, \( t(9) = 2.02, p = 0.074. \)

**Summary of Computations**

1. Compute a difference score for each subject.
2. Test whether the mean difference score is significantly different from zero (click here for a summary of the computational steps for this test)

Assumptions

1. Each subject is sampled independently from each other subject.

2. The difference scores are normally distributed. (If both raw scores are normally distributed then the difference score will be normally distributed.) The two raw scores from each subject do not have to be independent of each other.

Tests of Linear Combinations of Means, Independent Groups

In a hypothetical experiment, aspirin, Tylenol, and a placebo were tested to see how much pain relief each provides. Pain relief was rated on a five-point scale. Four subjects were tested in each group and their data are shown below:

<table>
<thead>
<tr>
<th></th>
<th>Aspirin</th>
<th>Tylenol</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Subject 2</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Subject 3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Subject 4</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Before conducting the experiment, the researcher planned to make two comparisons: (1) a comparison of the average of the aspirin and Tylenol groups with the placebo control and (2) a comparison between the aspirin and Tylenol groups. The comparisons can be framed in terms of linear combinations of means: The first comparison is:

\[
\frac{\mu_{\text{aspirin}} + \mu_{\text{Tylenol}}}{2} - \mu_{\text{placebo}}
\]

In terms of coefficients:

\[
(0.5)(\mu_{\text{aspirin}}) + (0.5)(\mu_{\text{Tylenol}}) + (-1)(\mu_{\text{placebo}})
\]

The steps for testing a linear contrast follow:

1. The first step is to specify the null hypothesis and an alternative hypothesis. For experiments testing linear combinations of means, the null hypothesis is: \( \Sigma a_i \mu_i \) is equal to some specified value, usually zero. In this experiment, \( a_1 = 0.5, a_2 = 0.5, \) and \( a_3 = -1. \) The null hypothesis is \( \Sigma a_i \mu_i = 0 \) which can be written as: \( a_1 \mu_1 + a_2 \mu_2 + a_3 \mu_3 = 0. \) For this experiment, the null hypothesis is:

\[
(0.5)(\mu_{\text{aspirin}}) + (0.5)(\mu_{\text{Tylenol}}) + (-1)(\mu_{\text{placebo}}) = 0.
\]

The alternative hypothesis is:

\( a_1 \mu_1 + a_2 \mu_2 + a_3 \mu_3 \neq 0. \)

2. The second step is to choose a significance level. Assume the .05 level is chosen.

3. The third step is to compute the value of the linear combination (L) based on the samples.

\[
L = a_1 M_1 + a_2 M_2 + \ldots + a_n M_n
\]
For these data,
\[ L = (0.5)(4) + (0.5)(3) + (-1)(2) = 1.5. \]

The fourth step is to calculate the probability of a value of \( L \) this different or more different from the value specified in the null hypothesis (zero). Applying the general formula, one obtains
\[ t = \frac{L}{s_L} \]
where \( L \) is the value of the linear combination in the sample and \( s_L \) is the standard error of \( L \). The formula for the standard error of \( L \) is given below.
\[
s_L = \sqrt{\frac{\sum a_i^2}{n} \text{MSE}}
\]

This formula is the same as the formula for \( s_L \) except that MSE (an estimate of \( \sigma_L^2 \)) is used in place of \( \sigma_L^2 \). The formula for MSE (and \( S_L^2 \)) is the same as the formula used in the calculation of a confidence interval on a linear combination of means.
\[
\text{MSE} = \frac{\sum s_i^2}{a}
\]
where \( s_i^2 \) is the sample variance of the ith group and "a" is the number of groups. This formula assumes homogeneity of variance and that the "a" sample sizes are equal. A related formula is described elsewhere that can be used with unequal sample sizes.

The degrees of freedom are equal to the sum of the degrees of freedom in the \( a = 3 \) groups. Therefore, \( df = a(n-1) = 3(4-1) = 9 \). A t table shows that the two-tailed probability value for a t of 2.33 with 9 df is 0.045

(continued) For this example,
\[ s_1^2 = 1.334, \quad s_2^2 = 1.334, \quad \text{and} \quad s_3^2 = 0.666 \] and, therefore,
\[
\text{MSE} = \frac{1.334 + 1.334 + 0.666}{3} = 1.111
\]
\[
\sum a_i^2 = 0.5^2 + 0.5^2 + (-1)^2 = 1.5.
\]
The sample size, \( n \), is 4 since there are four subjects in each group. Therefore,
\[
s_L = \sqrt{\frac{(1.5)(1.111)}{4}} = 0.645
\]
Plugging these values into the formula:
\[
t = \frac{L}{s_L}
\]
The probability computed in Step 4 is compared to the significance level stated in Step 2. Since the probability value (0.045) is less than the significance level (0.05), the effect is significant.

Since the effect is significant, the null hypothesis is rejected. It is concluded that the average relief provided by the two drugs is greater than the relief provided by the placebo.

An example of how the results of the experiment could be described in a report is shown below.

Both aspirin (M = 4) and Tylenol (M = 3) provided more pain relief than the placebo (M = 2). A planned comparison of the mean of the two experimental groups with the placebo control was significant, t(9) = 2.33, p = 0.045. The means of the two experimental groups were not significantly different from each other, t(9) = 1.34, p = 0.213.

It is important to state that the combination of means was planned in advance since a different procedure is used if the comparison is decided on after viewing the data.

The expression "t(9) = 2.33, p = 0.045" means that a t test with 9 degrees of freedom was conducted and that the probability value was 0.045. The test for the difference between the two experimental groups was conducted using the coefficients: a_1 = 1, a_2 = -1, a_3 = 0.

**Summary of Computations**

1. Specify the null and alternate hypotheses. This is done by choosing the coefficients (a's). The null hypothesis is \( \Sigma a_i \mu_i = 0 \). The alternate hypothesis is \( \Sigma a_i \mu_i \neq 0 \). (On rare occasions, the null hypothesis will be that \( \Sigma a_i \mu_i \) is some value other than zero.)

2. Compute \( L = \Sigma a_i M_i \).

3. Compute
   \[
   MSE = \frac{\sum s_i^2}{a}
   \]
   where "a" is the number of groups.

4. Compute
   \[
   s_L = \sqrt{\frac{\sum a_i^2}{n} \cdot MSE}
   \]
   where n is the number of subjects in each group (there must be equal n).
5. Compute

\[ t = \frac{L}{s_L} \]

6. Compute df = a(n-1) where "a" is the number of groups.

7. Use a t table to find the probability value of t.

Assumptions

1. All populations are normally distributed.
2. Variances in the "a" populations are equal.
3. Scores are independent (Each subject provides only one score).
4. There is an equal number of subjects in each group.
5. The comparison was planned in advance.

See also: "Confidence interval on linear combination of means, independent groups"

Tests of Pearson's Correlation

The sampling distribution of Pearson's r is normal only if the population correlation (ρ) equals zero; it is skewed if ρ is not equal to 0 (click here for illustration). Therefore, different formulas are used to test the null hypothesis that ρ = 0 and other null hypotheses.

Null Hypothesis: ρ = 0
A hypothetical experiment is conducted on the relationship between job satisfaction and job performance. A sample of 100 employees rate their own level of job satisfaction. This measure of job satisfaction is correlated with supervisors' ratings of performance. The question is whether there is a relationship between these two measures in the population.

1. The first step is to specify the null hypothesis and an alternative hypothesis. The null hypothesis is ρ = 0; the alternative hypothesis is ρ ≠ 0.
2. The second step is to choose a significance level. Assume the 0.05 level is chosen.
3. The third step is to compute the sample value of Pearson's correlation (click here for the formula). In this experiment, \( r = 0.27 \).

The fourth step is to compute p, the probability (or probability value) of obtaining a difference between and the value specified by the null hypothesis (zero) as large or larger than the difference obtained in the experiment. This can be calculated using the following formula:
The degrees of freedom are N-2 where N is the sample size.

For this example,
\[ t = \frac{0.27 \sqrt{100 - 2}}{\sqrt{1 - 0.27^2}} = 2.776 \]

The df are N - 2 = 100 - 2 = 98. A t table can be used to find that the two-tailed probability value for a t of 2.776 with 98 df is 0.007.

5. The probability computed in Step 4 is compared to the significance level stated in Step 2. Since the probability value (0.007) is less than the significance level (0.05), the correlation is significant.

6. Since the effect is significant, the null hypothesis is rejected. It is concluded that the correlation between job satisfaction and job performance is greater than zero.

7. A report of this finding might be as follows:

There was a small but significant relationship between job satisfaction and job performance, \( r = .27 \), \( t(98) = 2.78 \), \( p = .007 \).

The expression "\( t(98) = 2.78 \)" means that a t test with 98 degrees of freedom was equal to 2.78. The probability value is given by "\( p = .007 \)." Since it was not mentioned whether the test was one- or two-tailed, it is assumed the test was two tailed. The relationship between job satisfaction and job performance was described as "small." Refer to the section on "scatterplots of example values of r" to see what a relationship of that size looks like.

Other Null Hypotheses
For other null hypotheses, the significance test of \( r \) is done using Fisher's z' transformation. The formulas used in the fourth step of hypothesis testing are shown below:

\[ z = \frac{z_r - z_\rho}{\sigma_z} \]

where \( z_r \) is the value of \( r \) converted to \( z' \), \( z_\rho \) is the value of \( \rho \) converted to \( z' \), and \( \sigma_z \) is the standard error of \( z' \). This standard error can be computed as:

\[ \sigma_z = \frac{1}{\sqrt{N - 3}} \]

Once \( z \) is computed, a \( z \) table is used to calculate the \( p \) value

Suppose a theory of intelligence predicts that the correlation between intelligence and spatial ability is 0.50. A sample of 73 high school students was given both a test of intelligence and spatial ability. The correlation
between the tests was 0.677. Is the value of 0.677 significantly higher than 0.50? The null hypothesis therefore is \( \rho = 0.5 \).

The r to z' table can be used to convert both the r of 0.677 and the of 0.50 to Fisher's z'. The values of z' are 0.824 and 0.549 respectively. The standard error of z' is:

\[
\sigma_{z'} = \frac{1}{\sqrt{N - 3}} = 0.1195
\]

and therefore,

\[
z = (0.824 - 0.549)/0.1195 = 2.30.
\]

A z table shows that the two-tailed probability value for a z of 2.30 is 0.021. Therefore, the null hypothesis that the population correlation is 0.50 can be rejected.

**Summary of Computations**

1. Specify the null hypothesis and an alternative hypothesis.
2. Compute Pearson's r (click here for formula)
3. If the null hypothesis is \( \rho = 0 \) then compute t using the following formula:

\[
t = \frac{r \sqrt{N - 2}}{\sqrt{1 - r^2}}
\]

4. Use a t table to compute p for t and N-2 df.

4. If the null hypothesis is that is some specific value other than zero, then compute

\[
z = \frac{z_r - z_\rho}{\sigma_{z'}}
\]

and use a z table to find p.

**Assumptions**

1. The N pairs of scores are sampled randomly and independently.
2. The distribution of the two variables is bivariate normal.

**Tests of Differences between Independent Pearson Correlations**

A researcher was interested in whether the relationship between the quantitative portion of the SAT (QSAT) and grades in the first year of college (GPA) is different for engineering majors than it is for humanities majors. One hundred engineering majors and 100 humanities majors were sampled and each student's QSAT and GPA were recorded. The Pearson's correlation between these two measures was 0.61 for the engineering majors and 0.35 for the humanities majors. Is this difference in correlations significant?
1. **The first step** is to specify the null hypothesis and an alternative hypothesis. The null hypothesis and alternative hypotheses for this problem are:

   \[ H_0: \rho_{\text{engineers}} = \rho_{\text{humanities}} \]
   \[ H_1: \rho_{\text{engineers}} \neq \rho_{\text{humanities}} \]

   where \( \rho_{\text{engineers}} \) is the population correlation between QSAT and GPA for engineering majors and \( \rho_{\text{humanities}} \) is the population correlation between QSAT and GPA for the humanities majors.

2. **The second step** is to choose a significance level. Assume the 0.05 level is chosen.

3. The **third step** is to compute the sample correlations. In this experiment, \( r_{\text{engineers}} = 0.61 \) and \( r_{\text{humanities}} = 0.35 \).

4. The **fourth step** is to compute \( p \), the probability (or probability value). It is the probability of obtaining a difference between the statistic

   \[ r_1 - r_2 \]

   and the value specified by the null hypothesis (zero) as large or larger than the difference observed in the experiment. Since the sampling distribution of Pearson's \( r \) is not normal, (click here for illustration) the sample correlations are transformed to Fisher's \( z' \)s.

   The general formula applied to this problem is:

   \[ z = \frac{z'_{1} - z'_{2} - 0}{\sigma_{z'_{1} - z'_{2}}} \]

   where \( z'_{1} \) is the first correlation transformed to \( z' \), \( z'_{2} \) is the second correlation converted to \( z' \), and \( \sigma_{z'_{1} - z'_{2}} \) is the standard error of the difference between \( z' \)s and is equal to:

   \[ \sigma_{z'_{1} - z'_{2}} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \]

   where \( N_1 \) is the sample size for the first correlation and \( N_2 \) is the sample size for the second correlation.

5. (continued) For the experiment on the correlation between QSAT and GPA, the standard error of the difference between \( z' \)s is:

   \[ \sigma_{z'_{1} - z'_{2}} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} = 0.1435. \]

   The correlations of 0.61 and 0.35 can be transformed using the \( r \) to \( z' \) table to \( z' \)s of 0.709 and 0.365 respectively. Therefore,

   \[ z = \frac{0.709 - 0.365}{0.1435} = 2.40. \]

   A \( z \) table can be used to find that the two-tailed probability value for a \( z \) of 2.40 is 0.016.
6. The probability computed in Step 4 is compared to the significance level stated in Step 2. Since the probability value (0.016) is less than the significance level (.05), the effect is significant.

Math Textbooks

Since the effect is significant, the null hypothesis is rejected. It is concluded that the correlation between QSAT and GPA is higher for engineering majors than for humanities majors.

An example of a report of this finding is shown below.

The correlation between QSAT and GPA was significantly higher for the engineering majors \( r = 0.61 \) than for the humanities majors \( r = 0.35 \), \( z = 2.40, p = 0.016 \).

Summary of Computations

1. Specify the null hypothesis and an alternative hypothesis.
2. Compute Pearson’s \( r \) for both samples. (click here for formula).
3. Use the \( r \) to \( z' \) table to convert the \( r \)’s to \( z \)’s.

\[
\sigma_{z_1 - z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}
\]

4. Compute \( z = \frac{z_1' - z_2' - 0}{\sigma_{z_1 - z_2}} \) where \( N_1 \) is the sample size for the first correlation and \( N_2 \) is the sample size for the second correlation.
5. Compute a \( z \) table to find the probability value.

Assumptions

1. The two correlations are from two independent groups of subjects.
2. The \( N \) pairs of scores in each group are sampled randomly and independently.
3. The distribution of the two variables is bivariate normal.

Tests of Proportions

A manufacturer is interested in whether people can tell the difference between a new formulation of a soft drink and the original formulation. The new formulation is cheaper to produce so if people cannot tell the difference, the new formulation will be manufactured. A sample of 100 people is taken. Each person is given a taste of both formulations and asked to identify the original. Sixty-two percent of the subjects correctly identified the new formulation. Is this proportion significantly different from 50%?
1. The first step in hypothesis testing is to specify the null hypothesis and an alternative hypothesis. In testing proportions, the null hypothesis is that \( \pi \), the proportion in the population, is equal to some specific value. In this example, the null hypothesis is that \( \pi = 0.5 \). The alternate hypothesis is \( \pi \neq 0.5 \).

2. The second step is to choose a significance level. Assume the 0.05 level is chosen.

3. The third step is to compute the difference between the sample proportion (\( p \)) and the value of \( \pi \) specified in the null hypothesis. In this example, \( p - \pi = 0.62 - 0.5 = 0.12 \).

The fourth step is to compute \( p \), the probability (or probability value). It is the probability of obtaining a difference between the proportion and the value specified by the null hypothesis as large or larger than the difference observed in the experiment. The general formula for significance testing as applied to this problem If \( p \) is greater than \( \pi \) then the formula is:

\[
z = \frac{p - \pi - 0.5}{\sigma_p}.
\]

If \( p \) is less than \( \pi \) then the formula is:

\[
z = \frac{p - \pi + 0.5}{\sigma_p}.
\]

Note that the correction always makes \( z \) smaller.

\( N \) is the sample size and \( \sigma_p \) is the standard error of a proportion. The formula for a standard error of a proportion is:

\[
\sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}}.
\]

\[
0.5
\]

The term \( \frac{0.5}{N} \) is the correction for continuity. For the present problem,

\[
\sigma_p = \sqrt{\frac{(0.5)(0.5)}{100}} = 0.05.
\]

Therefore,

\[
z = \frac{0.62 - 0.50 - 0.5}{0.05} = 2.3.
\]

A z table can be used to determine that the two-tailed probability value for a \( z \) of 2.3 is 0.0214.
5. The probability computed in Step 4 is compared to the significance level stated in Step 2. Since the probability value (0.0214) is less than the significance level of 0.05, the effect is statistically significant.

6. Since the effect is significant, the null hypothesis is rejected. It is concluded that the proportion of people choosing the original formulation is greater than 0.50.

7. This result might be described in a report as follows:

The proportion of subjects choosing the original formulation (0.62) was significantly greater than 0.50, $z = 2.3$, $p = 0.021$. Apparently at least some people are able to distinguish between the original formulation and the new formulation.

Summary of Computations

1. Specify the null hypothesis and an alternative hypothesis.

2. Compute the proportion in the sample.

3. Compute

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}}.$$

4. If $p > \pi$ then compute

$$z = \frac{p - \pi - 0.5}{\sigma_p}$$

otherwise, compute

$$z = \frac{p - \pi + 0.5}{\sigma_p}$$

5. Use a z table to compute the probability value from $z$.

Assumptions

1. Observations are sampled randomly and independently.

2. The adequacy of the normal approximation depends on the sample size ($N$) and $\pi$. Although there are no hard and fast rules, the following is a guide to needed sample size: If $\pi$ is between 0.4 and 0.6 then an $N$ of 10 is adequate. If $\pi$ is as low as 0.2 or as high as 0.8 then $N$ should be at least 25. For $\pi$ as low as 0.1 or as high as 0.9, $N$ should be at least 30. A conservative rule of thumb is that both $N\pi$ and $N(1 - \pi)$ should be greater than 10.
An experiment is conducted investigating the long-term effects of early childhood intervention programs (such as head start). In one (hypothetical) experiment, the high-school drop out rate of the experimental group (which attended the early childhood program) and the control group (which did not) were compared. In the experimental group, 73 of 85 students graduated from high school. In the control group, only 43 of 82 students graduated. Is this difference statistically significant?

1. The first step in hypothesis testing is to specify the null hypothesis and an alternative hypothesis. When testing differences between proportions, the null hypothesis is that the two population proportions are equal. That is, the null hypothesis is:

   \[ H_0: \pi_1 = \pi_2. \]

   The alternative hypothesis is: \[ H_1: \pi_1 \neq \pi_2. \]

   In this example, the null hypothesis is:

   \[ H_0: \pi_{\text{intervention}} = \pi_{\text{no intervention}} = 0. \]

2. The second step is to choose a significance level. Assume the 0.05 level is chosen.

3. The third step is to compute the difference between the sample proportions. In this example, \[ p_1 - p_2 = \frac{73}{85} - \frac{43}{82} = 0.8588 - 0.5244 = 0.3344. \]

4. The fourth step is to compute \( p \), the probability (or probability value). It is the probability of obtaining a difference between the proportions as large or larger than the difference observed in the experiment. Applying the general formula to the problem of differences between proportions

   \[ z = \frac{p_1 - p_2}{s_{p_1 - p_2}} \]

   where \( p_1 - p_2 \) is the difference between sample proportions and

   \[ s_{p_1 - p_2} \]

   is the estimated standard error of the difference between proportions. The formula for the estimated standard error is:

   \[ s_{p_1 - p_2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} \]

   where \( p \) is a weighted average of the \( p_1 \) and \( p_2 \), \( n_1 \) is the number of subjects sampled from the first population, and \( n_2 \) is the number of subjects sampled from the second population.

(continued) Therefore, \[ z = 0.344/0.0713 = 4.69. \] A z table can be used to find that the two-tailed probability value for a \( z \) of 4.69 is less than 0.0001.

If \( n_1 = n_2 \) then \( p \) is simply the average of \( p_1 \) and \( p_2 \):

\[ p = \frac{p_1 + p_2}{2}. \]
The two p values are averaged since the computations are based on the assumption that the null hypothesis is true. When the null hypothesis is true, both \( p_1 \) and \( p_2 \) are estimates of the same value of \( \pi \). The best estimate of \( \pi \) is then the average of the two p's.

Naturally, if one p is based on more observations than the other, it should be counted more heavily. The formula for \( p \) when \( n_1 \neq n_2 \) is:

\[
p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.
\]

The computations for the example on early childhood intervention are:

\[
p = \frac{(85)(0.859) + (82)(0.524)}{85 + 82} = 0.695.
\]

\[
s_{p_1 - p_2} = \sqrt{\frac{0.695(1 - 0.695)}{85} + \frac{0.695(1 - 0.695)}{82}} = 0.0713.
\]

5. The probability computed in Step 4 is compared to the significance level stated in Step 2. Since the probability value (<0.0001) is less than the significance level of 0.05, the effect is significant.

6. Since the effect is significant, the null hypothesis is rejected. The conclusion is that the probability of graduating from high school is greater for students who have participated in the early childhood intervention program than for students who have not.

7. The results could be described in a report as:

The proportion of students from the early-intervention group who graduated from high school was 0.86 whereas the proportion from the control group who graduated was only 0.52. The difference in proportions is significant, \( z = 4.69, p < 0.001 \).

**Summary of Computations**

1. Compute \( p_1 \) and \( p_2 \).

\[
p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.
\]

2. Compute

\[
s_{p_1 - p_2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}.
\]

3. Compute

\[
z = \frac{p_1 - p_2}{s_{p_1 - p_2}}.
\]

4. Compute

\[
s_{p_1 - p_2} = \sqrt{\frac{(p_1 - p_2)^2}{n_1} + \frac{(p_1 - p_2)^2}{n_2}}.
\]

5. Use a \( z \) table to compute the probability value from \( z \). Note that the correction for continuity is not used in the test for differences between proportions.
Assumptions

1. The two proportions are independent.

2. For the normal approximation to be adequate, $\pi$ should not be too close to 0 or to 1. Values between 0.10 and 0.90 allow the approximation to be adequate.

3. For the normal approximation to be adequate, there should be at least 10 subjects per group.

For apply the exercises at following link:
http://davidmlane.com/hyperstat/hypothesis_testing_se.html
CHAPTER 12

Introduction to ANOVA
Analysis of variance (ANOVA) is used to test hypotheses about differences between two or more means. The t-test based on the standard error of the difference between two means can only be used to test differences between two means. When there are more than two means, it is possible to compare each mean with each other mean using t-tests. However, conducting multiple t-tests can lead to severe inflation of the Type I error rate. (Click here to see why) Analysis of variance can be used to test differences among several means for significance without increasing the Type I error rate. This chapter covers designs with between-subject variables. The next chapter covers designs with within-subject variables.

Consider a hypothetical experiment on the effect of the intensity of distracting background noise on reading comprehension. Subjects were randomly assigned to one of three groups. Subjects in Group 1 were given 30 minutes to read a story without any background noise. Subjects in Group 2 read the story with moderate background noise, and subjects in Group 3 read the story in the presence of loud background noise.

Math Textbooks

The first question the experimenter was interested in was whether background noise has any effect at all. That is, whether the null hypothesis: \( \mu_1 = \mu_2 = \mu_3 \) is true where \( \mu_1 \) is the population mean for the "no noise" condition, \( \mu_2 \) is the population mean for the "moderate noise" condition, and \( \mu_3 \) is the population mean for the "loud noise" condition. The experimental design therefore has one factor (noise intensity) and this factor has three levels: no noise, moderate noise, and loud noise.

Analysis of variance can be used to provide a significance test of the null hypothesis that these three population means are equal. If the test is significant, then the null hypothesis can be rejected and it can be concluded that background noise has an effect.

In a one-factor between-subjects ANOVA, the letter "a" is used to indicate the number of levels of the factor (a = 3 for the noise intensity example). The number of subjects assigned to condition 1 is designated as \( n_1 \), the number of subjects assigned to condition 2 is designated by \( n_2 \), etc.

If the sample size is the same for all of the treatment groups, then the letter "n" (without a subscript) is used to indicate the number of subjects in each group. The total number of subjects across all groups is indicated by "N." If the sample sizes are equal, then \( N = (a)(n) \); otherwise,

\[ N = n_1 + n_2 + \ldots + n_a. \]

Some experiments have more than one between-subjects factor. For instance, consider a hypothetical experiment in which two age groups (8-year olds and 12-year olds) are asked to perform a task either with or without distracting background noise. The two factors are age and distraction.

Assumptions
Analysis of variance assumes normal distributions and homogeneity of variance. Therefore, in a one-factor ANOVA, it is assumed that each of the populations is normally distributed with the same variance ($\sigma^2$). In between-subjects analyses, it is assumed that each score is sampled randomly and independently. Research has shown that ANOVA is "robust" to violations of its assumptions.

This means that the probability values computed in an ANOVA are satisfactorily accurate even if the assumptions are violated. Moreover, ANOVA tends to be conservative when its assumptions are violated. This means that although power is decreased, the probability of a Type I error is as low or lower than it would be if its assumptions were met. There are exceptions to this rule. For example, a combination of unequal sample sizes and a violation of the assumption of homogeneity of variance can lead to an inflated Type I error rate.

**Two estimates of variance**

Analysis of variance tests the null hypothesis that all the population means are equal:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_a$$

by comparing two estimates of variance ($\sigma^2$). (Recall that $\sigma^2$ is the variance within each of the "a" treatment populations.) One estimate (called the Mean Square Error or "MSE" for short) is based on the variances within the samples. The MSE is an estimate of $\sigma^2$ whether or not the null hypothesis is true. The second estimate (Mean Square Between or "MSB" for short) is based on the variance of the sample means. The MSB is only an estimate of $\sigma^2$ if the null hypothesis is true. If the null hypothesis is false then MSB estimates something larger than $\sigma^2$. (A later section discusses exactly what MSB estimates when the null hypothesis is false.)

The logic by which analysis of variance tests the null hypothesis is as follows: If the null hypothesis is true, then MSE and MSB should be about the same since they are both estimates of the same quantity ($\sigma^2$); however, if the null hypothesis is false then MSB can be expected to be larger than MSE since MSB is estimating a quantity larger then $\sigma^2$.

Therefore, if MSB is sufficiently larger than MSE, the null hypothesis can be rejected. If MSB is not sufficiently larger than MSE then the null hypothesis cannot be rejected. How much larger is "sufficiently" larger? That is determined by the statistical analysis explained in a later section.

**How $\sigma^2$ is Estimated by MSE**

The estimation of $\sigma^2$ by MSE in the analysis of variance is exactly the same as the estimation of $\sigma^2$ by MSE in the construction of confidence intervals on a linear combination of means and in the testing of linear combinations of means. It is very similar to the estimation of $\sigma^2$ by MSE in the construction of confidence intervals on the difference between means and in the testing of differences between means. The only difference is that there are "a" groups instead of two groups.

For simplicity, assume that the sample sizes in each of the "a" samples are equal. The variance ($\sigma^2$) is estimated by $s^2$ in each of the samples. The average of these "a" estimates of $\sigma^2$ is then used as an overall estimate of $\sigma^2$.

Therefore,

$$MSE = \frac{\sum_{i=1}^{a} s_i^2}{a}.$$
The example data from the section on testing of linear combinations of means are reproduced here.

<table>
<thead>
<tr>
<th>Aspirin</th>
<th>Tylenol</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

The means of the Aspirin, Tylenol, and Placebo conditions are 4, 3, and 2 respectively. The sample variances are 1.334, 1.334, and 0.666. The MSE is the mean of the three sample variances and is therefore:

\[
\frac{1.334 + 1.334 + 0.666}{3} = 1.111.
\]

**How \( \sigma^2 \) is Estimated by MSB**

Estimating \( \sigma^2 \) using the sample means is slightly more complicated than using the sample variances. The first step is to use the sample means to estimate the variance of the sampling distribution of the mean (\( \sigma_M^2 \)). In an experiment there are "a" means. The variance of these "a" means is used as the estimate of \( \sigma_M^2 \). For the example, the three sample means are: 4, 3, and 2.

Recall that the formula for \( s^2 \) is:

\[
s^2 = \frac{\sum (X - M)^2}{N - 1}.
\]

In this application, each "X" will be a sample mean and M will be the mean of the sample means. Since there are "a" means, \( N = a \). Letting M designate the mean of the sample means, \( M = (4 + 3 + 2)/3 = 3 \).

\( s_M^2 \) is the estimate of \( \sigma_M^2 \). For the example,

\[
s_M^2 = \frac{(4 - 3)^2 + (3 - 3)^2 + (2 - 3)^2}{3 - 1} = 1.
\]

This means that the estimate of \( \sigma_M^2 \) is 1. But what is really needed is not an estimate of \( \sigma_M^2 \) but an estimate of \( \sigma^2 \). Fortunately, there is a simple relationship between \( \sigma^2 \) and \( \sigma_M^2 \).

Since the sampling distribution of the mean has a standard deviation of: \( \sigma_M = \frac{\sigma}{\sqrt{N}} \), (click here for details) the sampling distribution of the mean has a variance of:

\[
\sigma_M^2 = \frac{\sigma^2}{N}
\]

where N is the number of scores each mean is based upon.
Rearranging the equation,

\[ \sigma^2 = n\sigma_M^2. \]

Therefore, to go from an estimate of \( \sigma_M^2 \) to an estimate of \( \sigma^2 \), you multiply the estimate of \( \sigma_M^2 \) by \( n \). For this example \( n = 4 \), so the estimate of \( \sigma^2 \) is:

\[ \text{MSB} = (4)(1) = 4. \]

In short,

\[ \text{MSB} = ns^2_M \]

where \( n \) is the number of subjects in each group and \( s^2_M \) is the variance of the sample means.

The significance tests associated with analysis of variance are based on the ratio of MSB to MSE. If the ratio is large enough, then the null hypothesis that the population means are equal can be rejected. The next section discusses the distribution of MSB/MSE and how to conduct a significance test.

### The Significance Test in ANOVA

If the null hypothesis is true, then both MSB and MSE estimate the same quantity. If the null hypothesis is false, then MSB is an estimate of a larger quantity (click here to see what it is).

The significance test involves the statistic \( F \) which is the ratio of MSB to MSE: \( F = \frac{\text{MSB}}{\text{MSE}}. \)

If the null hypothesis is true, then the \( F \) ratio should be approximately one since MSB and MSE should be about the same. If the ratio is much larger than one, then it is likely that MSB is estimating a larger quantity than is MSE and that the null hypothesis is false. In order to conduct a significance test, it is necessary to know the sampling distribution of \( F \) given that the null hypothesis is true. From the sampling distribution, the probability of obtaining an \( F \) as large or larger than the one calculated from the data can be determined. This probability is the probability value. If it is lower than the significance level, then the null hypothesis can be rejected. The mathematics of the sampling distribution were worked out by the statistician R. A. Fisher and is called the F distribution in his honor. (Click here for information about the F distribution.)

After an \( F \) is computed, the probability value can be computed from an F table. To use this table, you need to know the two degrees of freedom parameters: \( \text{dfn} \) and \( \text{dfd} \):

\[ \text{dfn} = a-1 \]
\[ \text{dfd} = N-a \]

where \( a \) is the number of groups and \( N \) is the total number of subjects in all groups. The parameter \( \text{dfd} \) is often called degrees of freedom error or dfe for short.

For the example data, \( \text{dfn} = 3-1 = 2 \)
\( \text{dfd} = 12-3 = 9 \), MSB = 4, and MSE = 1.111.

Therefore,
\[ F = \frac{\text{MSB}}{\text{MSE}} = \frac{4}{1.111} = 3.6. \]
An F table can be used to compute that the probability value for an F of 3.6 with 2 and 9 df is 0.071. Therefore the null hypothesis cannot be rejected at the 0.05 level.

**Partitioning the Sums of Squares**

Data from a hypothetical experiment on pain relief described earlier are reproduced below.

<table>
<thead>
<tr>
<th>Aspirin</th>
<th>Tylenol</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

The experiment consists of one factor (Drug condition) which has three levels (Aspirin, Tylenol, and Placebo). There are four subjects in each of the three groups making a total of N = 12 subjects. The scores for these 12 subjects obviously differ from each other: They range from two to five. There are two sources of these differences among scores (1) treatment effects and (2) error.

**Treatment Effects**

It is possible that some of the differences among the 12 scores obtained in the experiment are due to the fact that different subjects were given different drugs. If the drugs differ in their effectiveness, then subjects receiving the more effective drugs will have the higher pain-relief scores. The more different the effectiveness of the drugs, the greater will be the differences among the scores. Since the different drugs represent different ways the subjects were "treated," differences due to differences in drug effectiveness are said to be due to "treatment effects."

**Math Textbooks**

**Error**

All other differences among the subjects' scores are said to be due to "error." Do not take the word "error" literally, however. It does not mean that a mistake was made. For historical reasons, any differences among subjects that cannot be explained by the experimental treatments are called error. It would be better, perhaps, to call this source of differences "unexplained variation," but to be consistent with the terminology in common usage, these differences will be called "error."

A major source of error is the simple fact that even if subjects are treated exactly the same way in an experiment, they differ from each other, often greatly. This stands to reason since subjects differ in both their pre-experimental experiences and in their genetics. In terms of this example, it means that subjects given the same drug will not necessarily experience the same degree of pain relief. Indeed, there are differences among subjects within each of the three groups.

**Total Sum of Squares**

The variation among all the subjects in an experiment is measured by what is called sum of squares total or SST. SST is the sum of the squared differences of each score from the mean of all the scores. Letting GM (standing for "grand mean") represent the mean of all scores, then
STATISTICS
http://davidmlane.com/hyperstat/index.html

SST = Σ(X - GM)^2

where GM = ΣX/N and N is the total number of subjects in the experiment

For the example data:

N = 12

GM = (3+5+3+5+2+2+4+4+2+1+3+2)/12 = 3

SST = (3-3)^2+(5-3)^2+(3-3)^2+(5-3)^2 + (2-3)^2+(2-3)^2+(4-3)^2+(4-3)^2 + (2-3)^2+(1-3)^2+(3-3)^2+(2-3)^2 = 18

Sum of Squares Between Groups
The sum of squares due to differences between groups (SSB) is computed according to the following formula:

SSB = Σn_i(M_i - GM)^2

where n_i is the sample size of the ith group and M_i is the mean of the ith group, and GM is the mean of all scores in all groups.

If the sample sizes are equal then the formula can be simplified somewhat:

SSB = nΣ(M_i - GM)^2

For the example data,

M_1 = (3+5+3+5)/4 = 4
M_2 = (2+4+2+4)/4 = 3
M_3 = (2+1+3+2)/4 = 2
GM = 3
n = 4
SSB = 4[(4-3)^2 + (3-3)^2 + (2-3)^2] = 8

Sum of Squares Error
The sum of squares error is the sum of the squared differences between the individual scores and their group means. The formula for sum of squares error (SSE) for designs with two groups has already been given in the section on confidence interval on the difference between two independent means and in testing differences between two independent means.

The SSE is computed separately for each of the groups in the experiment and then summed.

SSE = SSE_1 + SSE_2 + ... + SSE_a

SSE_1 = Σ(X - M_1)^2 ; SSE_2 = Σ(X - M_2)^2 SSE_a = Σ(X - M_a)^2 where M_1 is the mean of Group 1, M_2 is the mean of Group 2, and M_a is the mean of Group a.

For the example,

SSE_1 = (3-4)^2 + (5-4)^2 + (3-4)^2 + (5-4)^2 = 4
The sums of squares computed in this example are:

\[ \text{SST} = 18 \]
\[ \text{SSB} = 8 \]
\[ \text{SSE} = 10. \]

Notice that \( \text{SST} = \text{SSB} + \text{SSE} \). This is important because it shows that the total sum of squares can be divided into two components: the sum of squares due to treatments (SSB) and the sum of squares that not due to treatments (SSE).

The ANOVA Summary Table

It is convenient to view the results of an analysis of variance in a summary table. The summary table for the example data is shown below.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Ssq</th>
<th>Ms</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>2</td>
<td>8.000</td>
<td>4.000</td>
<td>3.60</td>
<td>0.071</td>
</tr>
<tr>
<td>Error</td>
<td>9</td>
<td>10.000</td>
<td>1.111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>18.000</td>
<td>1.636</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first column contains the source of variation (or source for short). There are two sources of variation: differences among treatment groups and error. The second column contains the degrees of freedom (df). The df for groups is always equal to the number of groups minus one. The df for error is equal to the number of subjects minus the number of groups. The formulas for the degrees of freedom are shown below:

\[ df_{\text{groups}} = a - 1 = 3 - 1 = 2 \]
\[ df_{\text{error}} = N - a = 12 - 3 = 9 \]
\[ df_{\text{total}} = N - 1 = 12 - 1 = 11 \]

\( a \) is the number groups
\( N \) is the total number of subjects.
\( df_{\text{total}} = df_{\text{groups}} + df_{\text{error}} \)

The third column contains the sums of squares. Notice that the sum of squares total is equal to the sum of squares groups + sum of squares error. The fourth column contains the mean squares. Mean squares are estimates of variance and are computed by dividing the sum of squares by the degrees of freedom. The mean square for groups (4.00) was computed by dividing the sum of squares for groups (8.00) by the degrees of freedom for groups (2). The fifth column contains the F ratio. The F ratio is computed by dividing the mean square for groups by the mean square for error. In this example,
F = 4.000/1.111 = 3.60.

There is no F ratio for error or total. The last column contains the probability value. It is the probability of obtaining an F as large or larger than the one computed in the data assuming that the null hypothesis is true. It can be computed from an F table. The df for groups (2) is used as the degrees of freedom in the numerator and the df for error (9) is used as the degrees of freedom in the denominator. The probability of an F with 2 and 9 df as larger or larger than 3.60 is 0.071.

**Computational Methods**

The formulas for sums of squares given in the section on partitioning the variance are not the most efficient computationally. They were chosen because they help to convey the conceptual basis of analysis of variance. This section provides computational formulas. Most likely you will not often have to use these formulas since analysis of variance is usually done by computer. Nonetheless, if you ever need to do an ANOVA with a hand calculator, these formulas may help. Data from a hypothetical experiment on pain relief described earlier are reproduced below.

<table>
<thead>
<tr>
<th>Aspirin</th>
<th>Tylenol</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

The sum of squares total (SST) can be calculated as:

\[
SST = \sum X^2 - CF
\]

where

\[
CF = \frac{\left(\sum X\right)^2}{N}
\]

and N is the total number of subjects.

For these data: \( CF = (3+5+3+5+2+2+4+4+2+1+3+2)^2/12 = 36^2/12 = 108.\)

\[
SST = 3^2+5^2+3^2+5^2+2^2+2^2+4^2+4^2+2^2+1^2+3^2+2^2 - CF
\]

= 126-108 = 18

The sum of squares between (SSB) can be calculated as:

\[
SSB = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \ldots + \frac{T_s^2}{n_s} - CF
\]

where \( T_1^2 \) is the total of the scores in group 1 squared, \( T_2^2 \) is the total of the scores in group 2 squared, etc. \( n_1 \) is the sample size for group 1, \( n_2 \) is the sample size for group 2, etc. For the example data,
STATISTICS

http://davidmlane.com/hyperstat/index.html

\[ T_1^2 = (3 + 5 + 3 + 5)^2 = 16^2 = 256 \]
\[ T_2^2 = (2 + 2 + 4 + 4)^2 = 12^2 = 144 \]
\[ T_3^2 = (2 + 1 + 3 + 2)^2 = 8^2 = 64 \]

SSB = 256/4 + 144/4 + 64/4 - 108 = 8

Since SST = SSB + SSE,
SSE = SST - SSB = 18 - 8 = 10

With equal sample sizes, the formula for SSB can be simplified somewhat:

\[ SSB = \frac{1}{n} \left( T_1^2 + T_2^2 + ... + T_a^2 \right) - CF \]

which in this case is:

\[ SSB = \frac{1}{4} (256 + 144 + 64) - 108 = 8 \]

Introduction to Tests Supplementing a One-factor Between-Subjects ANOVA

The null hypothesis in a one-factor between-subjects ANOVA is that all the population means are equal:

\[ H_0: \mu_1 = \mu_2 = ... = \mu_a. \]

Unfortunately, when the analysis of variance is significant and the null hypothesis is rejected, the only valid inference that can be made is that at least one population mean is different from at least one other population mean. The analysis of variance does not reveal which population means differ from which others. Experimenters usually are interested in more information. They want to know precisely where the differences lie. Consequently, further analyses are usually conducted after a significant analysis of variance. These further analyses almost always involve conducting a series of significance tests. This causes a very serious problem: the more significance tests that are conducted, the greater the chance that at least one of them will produce a Type I error. The probability that a single significance test will result in a Type I error is called the per-comparison error rate (PCER).

The probability that at least one of the tests will result in a Type I error is called the experimentwise error rate (EER). Statisticians differ in their views of how strictly the EER must be controlled. Some statistical procedures provide strict control over the EER whereas others control it to a lesser extent. Naturally there is a tradeoff between the Type I and Type II error rates. The more strictly the EER is controlled, the lower the power of the significance tests.

The remaining sections in this chapter discuss tests used to supplement the finding of a significant ANOVA. Pay particular attention to the tradeoffs among three conflicting goals:

1. to extract all information from the data that is meaningful,
2. to control the EER, and
3. to achieve a high level of power.
All Pairwise Comparisons among Means: Introduction

Treatment conditions are normally included in an experiment to see how they compare to other treatment conditions. Therefore, experimenters frequently wish to compare the mean of each treatment condition with the mean of each other treatment condition. The problem with performing comparisons of all pairs of means (pairwise comparisons) is that there can be quite a large number of comparisons and therefore the danger of a high EER. The figure below shows the number of possible pairwise comparisons as a function of the number of conditions in the experiment. If "a" is the number of conditions, then \((a)(a-1)/2\) pairwise comparisons among condition means are possible. Therefore there are \((a)(a-1)/2\) chances to make a Type I error.

![Graph showing the number of possible pairwise comparisons as a function of the number of conditions.](image)

All Pairwise Comparisons among Means: All t-tests

If there were no need to be concerned about the EER, then, instead of computing an analysis of variance, one could simply compute t-tests among all pairs of means. However, the effect on the EER is not trivial. For example, consider an experiment conducted with eight subjects in each of six treatment conditions. If the null hypothesis were true and all 15 t-tests were conducted using the 0.01 significance level, then the probability that at least one of the 15 tests would result in a Type I error is 0.10. Thus, the EER of 0.10 would be 10 times as high as the PCER of 0.01. Because computing t-tests on all pairs of means results in such a high EER, it is generally not considered an acceptable approach.

All Pairwise Comparisons among Means: Fisher's LSD Procedure

An approach suggested by the statistician R. A. Fisher (called the "least significant difference method" or Fisher's LSD) is to first test the null hypothesis that all the population means are equal (the omnibus null hypothesis) with an analysis of variance. If the analysis of variance is not significant, then neither the omnibus null hypothesis nor any other null hypothesis about differences among means can be rejected. If the analysis of variance is significant, then each mean is compared with each other mean using a t-test. The advantage of this approach is that
there is some control over the EER. If the omnibus null hypothesis is true, then the EER is equal to whatever significance level was used in the analysis of variance. In the example with the six groups of subjects given in the section on t-tests, if the 0.01 level were used in the analysis of variance, then the EER would be 0.01. The problem with this approach is that it can lead to a high EER if most population means are equal but one or two are different.

In the example, if a seventh treatment condition were included and the population mean for the seventh condition were very different from the other six population means, an analysis of variance would be likely to reject the omnibus null hypothesis. So far, so good, since the omnibus null hypothesis is false. However, the probability of a Type I error in one or more of the 15 t-tests computed among the six treatments with equal population means is about 0.10. Therefore, the LSD method provides only minimal protection against a high EER.

Homogeneity of variance is typically assumed for Fisher's LSD procedure. Therefore, MSE, the estimate of variance, is based on all the data, not just on the data for the two groups being compared. In order to make the relationship between Fisher's LSD and other methods of computing pairwise comparisons clear, the formula for the studentized t (t_s) rather than the usual formula for t is used. This makes no difference in the outcome since, for Fisher's LSD procedure, the critical value of t is computed as if their were only two means in the experiment, a situation in which t and t_s result in identical probability values, although t_s will be 1.414 times t, the critical value of ts will also be 1.414 times the critical value of t. It makes no difference in the results.

All Pairwise Comparisons among Means: Tukey’s HSD Procedure

The "Honestly Significantly Different" (HSD) test proposed by the statistician John Tukey is based on what is called the "studentized range distribution." To test all pairwise comparisons among means using the Tukey HSD, compute t for each pair of means using the formula:

\[
t_s = \frac{M_i - M_j}{\sqrt{\frac{\text{MSE}}{n_h}}}\]

where \(M_i - M_j\) is the difference between the ith and jth means, \(\text{MSE}\) is the Mean Square Error, and \(n_h\) is the harmonic mean of the sample sizes of groups i and j.

The critical value of \(t_s\) is determined from the distribution of the studentized range. The number of means in the experiment is used in the determination of the critical value, and this critical value is used for all comparisons among means. Typically, the largest mean is compared with the smallest mean first. If that difference is not significant, no other comparisons will be significant either, so the computations for these comparisons can be skipped.

The Tukey HSD procedure keeps the EER at the specified significance level. This is a great advantage. This advantage comes at a cost, however: the Tukey HSD is less powerful than other methods of testing all pairwise comparisons.
The Newman-Keuls method, like the Tukey HSD, is based on the studentized range distribution. Consider an experiment in which there are five treatment conditions. First the means are rank ordered from smallest to largest. Then, the smallest mean is compared to the largest mean using the studentized t. If the test is not significant, then no pairwise tests are significant and no more testing is done. So far, the Newman-Keuls method is exactly the same as the Tukey HSD. If the difference between the largest mean and the smallest mean is significant, then the difference between the smallest mean (Mean 1) and the second largest mean (Mean 4) as well as the difference between the largest mean (Mean 5) and the second smallest mean (Mean 2) are tested. Unlike the Tukey HSD, these comparisons are done using a critical value based on only four means rather than all five. The rationale is that the comparison of Mean 1 to Mean 4 only spans four means so the lower critical value associated with four rather than five means is used. The basic idea is that when a comparison that spans k means is significant, comparisons that span k-1 means within the original span of k means are performed.

The critical value of t associated with k-1 means is used for these comparisons. If a comparison spanning k means is not significant, then no further comparisons within the span of k means is performed. For instance, assume that the following five sample means were obtained from an experiment in which there were 5 subjects in each group: 10, 11, 14, 18, 25. Further assume that MSE = 30. The critical values for the studentized t with 20 degrees of freedom (20 = k x (k-1) = 5 x 4) at the 0.05 level are shown below (Use a table of the studentized range to determine critical values.):

<table>
<thead>
<tr>
<th>Span</th>
<th>Critical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.95</td>
</tr>
<tr>
<td>3</td>
<td>3.58</td>
</tr>
<tr>
<td>4</td>
<td>3.96</td>
</tr>
<tr>
<td>5</td>
<td>4.23</td>
</tr>
</tbody>
</table>

The first comparison is between 10 and 25. The critical value of t is based on a span of 5 means and is equal to 4.23. The calculated value of t for this comparison is 6.12, so the comparison is significant.

Since the comparison at span 5 is significant, the following two comparisons are made at span 4 using the lower critical value of t of 3.96:

Mean 1 versus Mean 4
Mean 2 versus Mean 5

The t for Mean 1 versus Mean 4 is 3.27, which is not significant. Therefore, no further comparisons within the span are performed and the following comparisons are deemed non significant:

Mean 1 versus Mean 2
Mean 1 versus Mean 3
Mean 1 versus Mean 4
Mean 2 versus Mean 3
Mean 2 versus Mean 4
Mean 3 versus Mean 4
The t for Mean 2 versus Mean 5 is 5.71 which is higher than the critical value for a span of four of 3.96. Since the comparison of Means 2 and 5 is significant, comparisons of span 3 within this span that were not previously ruled out are performed. Since the comparison of Means 2 and 4 is ruled out by the failure of the comparison of Means 1 and 4 to be significant, the only comparison that is done is between Means 3 and 5. The value of t is 4.49 which is higher than the critical value for span 3 of 3.58.

After the significant difference at span 3, differences not previously ruled out at span 2 are tested. The comparison of Means 3 and 4 has been ruled out by the failure of the comparison of Means 1 and 4 to be significant. Therefore, the only comparison left to be performed is between Means 4 and 5. The t is 2.86 which is less than the critical value for a span of 2 of 2.95, so the difference is not significant. The Newman-Keuls procedure has the advantage of being more powerful than the Tukey HSD. It is better at controlling the EER than the Fisher's LSD. However, there are patterns of population means that can lead to an inflated EER. For instance, if six population means were: 10, 10, 100, 100, 1,000, and 1,000 then comparisons among sample means at span 2 would almost certainly be performed. The null hypothesis is true for three of these comparisons: Mean 1 versus Mean 2 Mean 3 versus Mean 4 Mean 5 versus Mean 6 Since the PCER for these comparisons is 0.05, the EER is above 0.05.

All Pairwise Comparisons among Means: Duncan's Procedure

Duncan's procedure is similar to the Newman-Keuls procedure. The difference is that the critical values are much more liberal. The goal of Duncan's procedure is to keep the EER at

\[ 1 - (1 - \alpha_{pc})^{a-1} \]

where "a" is the number of means and is the PCER. This level of EER can be unacceptably high, however. With six means and a per comparison error rate of 0.05, the EER is 0.23. Duncan's procedure is only very slightly more conservative than Fisher's LSD. In practice, they almost always lead to the same conclusions.

All Pairwise Comparisons among Means: Recapitulation and Recommendations

The Fisher's LSD, Duncan's, Newman-Keuls, and Tukey HSD procedures all use the same formula for computing t. They differ in the critical value. The following figure shows the critical values for an experiment with 10 treatment conditions and four subjects per condition as a function of the number of means spanned in the comparison.

Note that the number of means spanned does not affect the critical value for either Tukey's
HSD or Fisher's LSD. The former uses the most conservative (highest) critical value for all comparisons whereas the latter uses the least conservative (lowest) critical value for all comparisons. The Newman-Keuls and the Duncan procedures increase the critical value as a function of span, although Duncan's does not increase it much.

There is no "correct" procedure to use; the various procedures trade off power for control of the EER in different ways. There is a consensus that the Duncan's and Fisher's LSD procedures result in too high an EER and should not be used. The choice between the Newman-Keuls and Tukey HSD is a close call. You have to decide how important it is to control the EER completely. If you want to be sure that you have controlled the EER, then the Tukey HSD should be used. Most statisticians now consider the Newman-Keuls unacceptably liberal because of situations in which it allows an inflated EER. However, it is up to the researcher to weigh the balance between power and controlling the EER.

Be careful not to accept the null hypothesis for comparisons for which the null hypothesis is not rejected. Otherwise you may be left with contradictory conclusions. You may find, for example, that \( \mu_1 \) is significantly different from \( \mu_3 \) but that \( \mu_1 \) is not significantly different from \( \mu_2 \) and that \( \mu_2 \) is not significantly different from \( \mu_3 \).

If the null hypothesis were accepted for the two nonsignificant comparisons, you would conclude that:

1. \( \mu_1 \neq \mu_3 \)
2. \( \mu_1 = \mu_2 \) and
3. \( \mu_2 = \mu_3 \)

which is clearly not possible. The proper conclusion is that either \( \mu_1 \neq \mu_2, \mu_2 \neq \mu_3 \) or neither \( \mu_1 \) nor \( \mu_3 \) is equal to \( \mu_2 \). More data are needed to decide among these three alternatives.

All Pairwise Comparisons among Means: Example Calculations

The table shown below contains the data from a hypothetical experiment with four groups of four subjects each.

<table>
<thead>
<tr>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>2.0</td>
<td>5.0</td>
<td>5.0</td>
<td>7.0</td>
</tr>
<tr>
<td>3.0</td>
<td>4.0</td>
<td>4.0</td>
<td>7.0</td>
</tr>
<tr>
<td>3.0</td>
<td>6.5</td>
<td>7.0</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Means: 3.0 4.875 5.0 7.0

The MSE is 1.1823. The critical values as a function of the number of steps for the four methods (as determined from a studentized range table) are shown below.

<table>
<thead>
<tr>
<th>Number of steps</th>
<th>PROCEDURE</th>
<th>Fisher's LSD</th>
<th>Duncan</th>
<th>Newman-Keuls</th>
<th>Tukey HSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.08</td>
<td>3.08</td>
<td>3.08</td>
<td>3.08</td>
<td>4.20</td>
</tr>
<tr>
<td>3</td>
<td>3.08</td>
<td>3.08</td>
<td>3.23</td>
<td>3.08</td>
<td>4.20</td>
</tr>
<tr>
<td>4</td>
<td>3.08</td>
<td>3.77</td>
<td>3.33</td>
<td>3.08</td>
<td>4.20</td>
</tr>
</tbody>
</table>
Substituting the values for $n_h$ and MSE,

$$t_s = \frac{M_i - M_j}{\sqrt{\frac{1.1823}{4}} \cdot 0.544}$$

lower    higher
mean     mean        $t_s$
G1 vs G2:  3.000    4.875     3.449
G1 Vs G3:  3.000    5.000     3.679
G1 Vs G4:  3.000    7.000     7.357
G2 Vs G3:  4.875    5.000     0.230
G2 Vs G4:  4.875    7.000     3.909
G3 Vs G4:  5.000    7.000     3.679

The table shows the value of $t$ for each of the six possible pairwise comparisons. For Fisher's LSD, all values of $t$ except the comparison of Group 2 and Group 3 are above the critical value of 3.08 and are therefore significant.

Duncan's test begins with a comparison of G1 and G4 using a critical value of 4.20. The test is significant so the comparisons G1 vs G3 and G2 vs G4 are tested using a critical value of 3.77. The comparison G1 vs G3 is not significant. This means that the comparisons of G1 vs G2 and G2 vs G3 are deemed not significant as well. The comparison of G2 vs G4 is significant, so the comparison G3 vs G4 is done using the critical value of 3.08. The comparison is significant. Two comparisons that were significant with the Fisher's LSD are not significant with the Newman-Keuls test (G1 vs G3 and G1 vs G2). The Tukey HSD uses a critical value of 4.20 for all comparisons. The only significant difference is between Group 1 and Group 4.
Comparing Means with a Control

The purpose of some experiments is to compare the mean of each of several experimental groups with the mean of a control group. For example, a researcher may wish to find out whether any of four new methods of teaching arithmetic is better than the traditional method.

Subjects are taught with one of the four new methods or with the traditional method (the control group). The experimenter then wishes to see which experimental group means (if any) are significantly different from the control group mean. Naturally, one could use a procedure such as Tukey's HSD to compare each mean with each other mean. The problem with using the Tukey's HSD or any other method designed to compare each mean with each other mean is that these methods overcorrect for the number of comparisons made. If each mean is compared to each other mean, then a total of \((a)(a-1)/2\) comparisons among "a" means would be tested. If there are "a" means (including the control) then there are only \((a-1)\) comparisons between experimental means and the control mean to be tested.

The procedure for comparing each experimental mean with the control mean is called "Dunnett's test" after the statistician who developed it. Dunnett's test controls the EER and is more powerful than tests designed to compare each mean with each other mean. Dunnett's test is conducted by computing a t-test between each experimental group and the control group using the formula:

\[
t_d = \frac{M_i - M_c}{2 \text{MSE}} \sqrt{n_h}
\]

where \(M_i\) is the mean of the \(i\)th experimental group, \(M_c\) is the mean of the control group, MSE is the mean square error as computed from the analysis of variance, and \(n_h\) is the harmonic mean of the sample sizes of the experimental group and the control group. The degrees of freedom (df) for the test are equal to \(N-a\) where \(N\) is the total number of subjects in all groups and "a" is the number of groups (including the control).

The critical value of \(t_d\) depends on the number of means in the experiment. Critical values can be obtained from a table of Dunnett's test.

In a hypothetical experiment described earlier, the effectiveness of Aspirin, Tylenol, and a Placebo were measured. The data are reproduced below.

<table>
<thead>
<tr>
<th>Aspirin</th>
<th>Tylenol</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

The analysis of variance summary table is shown below:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Ssq</th>
<th>Ms</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>2</td>
<td>8.0</td>
<td>4.00</td>
<td>3.60</td>
<td>0.071</td>
</tr>
<tr>
<td>Error</td>
<td>9</td>
<td>10.0</td>
<td>1.111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>18.0</td>
<td>1.636</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The means of the Aspirin, Tylenol, and Placebo groups are: 4, 3, and 2 respectively. Comparing the Aspirin group to the Placebo control,

\[ t_a = \frac{4 - 2}{\sqrt{\frac{2(1.111)}{4}}} \]

\[ = 2.68. \]

A table for Dunnett's test can be used to determine that the critical value for the 0.05 level (two-tailed) is 2.61. Therefore, this comparison is significant at the 0.05 level.

Comparing the Tylenol group to the Placebo control,

\[ t_d = \frac{3 - 2}{\sqrt{\frac{2(1.111)}{4}}} \]

\[ = 1.34 \] which is not significant.

**Specific Comparisons among Means: Overview**

Frequently experimenters wish to make more complex comparisons than simply comparisons between pairs of means. In a hypothetical experiment described earlier, the researcher wished to compare the average pain relief experienced by the aspirin and Tylenol groups with relief experienced by the placebo group. The null hypothesis was:

\[ H_0 : \frac{\mu_{\text{aspirin}} + \mu_{\text{Tylenol}}}{2} - \mu_{\text{placebo}} = 0 \]

In other words, the null hypothesis was that the average amount of relief experienced by subjects taking either of the two drugs was the same as the relief experienced by subjects taking a placebo. This null hypothesis can be specified in terms of a linear combination of means:

\[ H_0 : \sum a_i \mu_i = 0 \] where \( a_1 = 0.5 \), \( a_2 = 0.5 \), and \( a_3 = -1 \).

**Planned versus Unplanned Comparisons**

It makes a tremendous difference whether or not a comparison among means is planned prior to viewing the data. If the comparison is planned in advance, it can be tested using the procedures spelled out in another section.

A very different and much less powerful procedure must be used if the comparison is unplanned. It may seem that it should not matter whether or not a comparison is planned in advance. After all, the the same data are analyzed regardless of whether or not the comparison is planned. It does matter, however, because if an experimenter looks at the data and then chooses a comparison, he or she will almost certainly choose to compare the means that differ the most. This is tantamount to doing all possible comparisons among means, a procedure that, by capitalizing on chance, produces an inflated Type I error rate.
The Scheffé test is a procedure that allows one to test unplanned comparisons among means without inflating the Type I error rate. Scheffé’s test gives the freedom to test any and all comparisons that looks interesting. However, this great flexibility has a cost: Scheffé’s test normally has very low power.

**Testing Multiple Comparisons**

Even if the comparisons are planned in advance, testing more than one comparison results in an increase in the EER. One way to control the EER is to lower the significance level for the PCER. Naturally, the more comparisons that are conducted, the more the PCER has to be lowered. A second way to control the EER is to use Scheffé’s test. Scheffé’s test controls the EER no matter how many comparisons are conducted.

**Computing Tests of Comparisons**

**Planned Comparisons**

A method for computing planned comparisons among means is described in another section. That method is generalized here so that it can be used with unequal as well as with equal sample sizes.

First compute

\[
t = \frac{L}{s_L}
\]

where \( L = \sum M_i a_i \) and

\[
s_L = \sqrt{\frac{\sum a_i^2}{n_i} \text{MSE}}
\]

\( a_i \) is the coefficient applied to the \( i \)th mean

\( n_i \) is the sample size of the \( i \)th group

\( M_i \) is the \( i \)th mean

\( \text{MSE} \) is from the analysis of variance. It equals \( \frac{\text{SSE}}{\text{dfe}} \)

The \( t \) is based on \( \text{dfe} = N-a \) degrees of freedom where \( N \) is the total number of subjects and "a" is the number of groups.

**Math Textbooks**

Consider a hypothetical experiment on the effect of background music on reading comprehension. There were five groups of subjects who read in the presence of either:

Group 1: classical music (soft)
Group 2: classical music (loud)
Group 3: rock music (soft)
Group 4: rock music (loud)
Group 5: no background music

The reading comprehension scores are shown below:
Assume the experimenter had planned in advance to test whether classical music and rock music have different effects. The coefficients to test the difference between classical and rock music (ignoring volume) are:

1, 1, -1, -1, 0.

These coefficients compare the first two groups with the second two groups. The means of the five groups are: 91.2, 89.5, 86.0, 84.4, and 94.0 respectively. To compute $L$, the coefficients are multiplied times the means. Therefore,

$$L = (1)(91.2) + (1)(89.5) + (-1)(86.0) + (-1)(84.4) + (0)(94) = 10.3$$

As can be seen from the analysis of variance summary table shown below, MSE = 21.722.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Ssq</th>
<th>Ms</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>4</td>
<td>274.913</td>
<td>68.728</td>
<td>3.16</td>
<td>.039</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>391.000</td>
<td>21.722</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>665.913</td>
<td>30.269</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$\sum \frac{a_i^2}{n_i} = (1)^2/5 + (1)^2/4 + (-1)^2/5 + (-1)^2/5 + (0)^2/4 = .85$$

Therefore,

$$s_L = \sqrt{(0.85)(21.722)} = 4.297.$$ (Click here for formula for $s_L$)

Finally,

$$t = \frac{L}{s_L} = \frac{10.3}{4.297} = 2.397.$$ (See the summary table). From a $t$ table, it can be determined that the probability value associate with a $t$ of 2.397 with 18 df is 0.028. Comparisons among means can be tested with $F$ as well as with $t$. The formula for $F$ is:

$$F = \frac{MSB}{MSE}$$

where MSB (the mean square between) is equal to the sum of squares between (SSB) divided by the degrees of freedom numerator (dfn): $MSB = SSB/dfn$.

The formula for SSB for a comparison is:

$$SSB = \frac{L^2}{\sum \frac{a_i^2}{n_i}}$$
where L, a, and ni are as defined previously. The dfn for a planned comparison is always one. Therefore, MSB = SSB.

MSE, as defined previously, is taken from the analysis of variance.

The degrees of freedom for the F test are:

dfn = 1
dfe = N-a.

An F with one dfn is equal to t². Therefore, the F test of a comparison is equal to the square of the t-test of the comparison described previously. For the example data,

\[
MSB = SSB = \frac{10.3^2}{0.85} = 124.812.
\]

From the ANOVA table, MSE = 21.722.

\[
F = \frac{124.812}{21.722} = 5.746.
\]

The F of 5.746 = t² = 2.397².

The probability value for F (as determined from an F table using dfn = 1 and dfe = 18) is 0.028, which is the same as for t.

**Unplanned Comparisons**

Scheffé's test is used for unplanned comparisons. This test is the same as the F test for planned comparisons just discussed except that dfn = a-1. Therefore,

\[
MSB = SSB/dfn = SSB/(a-1)
\]

As in the section on planned comparisons, MSE is from the analysis of variance. It equals: SSE/dfe.

The section on planned comparisons has a computational example of a planned comparison. For that example, assume that the comparison between the classical music and the rock music was not planned in advance. This makes it necessary to do Scheffé's test. In the example, the values: SSB = 124.812 and MSE = 21.722 have already been computed.

\[
MSB = \frac{124.812}{dfn} = \frac{124.812}{(5-1)} = 31.203
\]

\[
F = \frac{31.203}{21.722} = 1.44
\]

An F table shows that the probability of an F with 4 and 18 degrees of freedom being 1.44 or larger is 0.26. Compare this value with the probability value of 0.028 obtained when the comparison was assumed to be planned in advance. Scheffé's test is not a powerful test. If at all possible, you should plan your comparison(s) among means in advance.

**Synonyms**

Planned comparisons are sometimes called "a priori" comparisons. Unplanned comparisons are sometimes called "post hoc" comparisons and at other times are called "a posteriori" comparisons.
Multiple Comparisons

If more than one comparison among means is conducted at a given PCER, the EER will be higher than the PCER. The following inequality can be used to control the EER:

\[ \text{EER} \leq (c)(\text{PCER}) \]

where \( c \) is the number of comparisons performed. For example, if six comparisons are performed at the 0.05 significance level (PCER = 0.05), then the EER is less than or equal to \((6)(0.05) = 0.30\). If a researcher wishes to perform six comparisons and keep the EER at the 0.05 level, the \(0.05/6 = 0.0083\) significance level should be used for each comparison. That way, the EER would be less than or equal to \((6)(0.0083) = 0.05\).

In general, to keep the EER at or below 0.05, the PCER should be:

\[ \text{PCER} = 0.05/c \]

More generally, to insure that the EER is less than or equal to \(\alpha\), use

\[ \text{PCER} = \alpha/c. \]

Adjusting the PCER in this manner is called either the "Bonferroni adjustment" or "Dunn's procedure" after the statisticians who developed it. Since Scheffé's test can also be used to test multiple comparisons, it is important to use the Scheffé's test if it is more powerful than the Bonferroni adjustment. Scheffé's test will be more powerful if more than three comparisons are planned among three means or more than seven comparisons are planned among four means. In almost all realistic situations, the Bonferroni adjustment is more powerful.

As in the case of pairwise comparisons, there is a tradeoff between controlling the EER and power. The Bonferroni adjustment provides control over the EER at a substantial cost in power. Some statisticians argue that it is not always necessary to control the EER. The "experiment," after all, is an arbitrary unit of analysis. Why is it necessary to control the error rate in one experiment but not in a whole series of experiments. For example, if a researcher conducted a series of five related experiments, few statisticians would recommend that the probability of a Type I error in any of the comparisons in any of the experiments be controlled. Nonetheless, the experiment is a convenient unit of analysis. The general consensus is that steps should be taken to control the EER.

There is room for argument, however, about whether it should be strictly controlled, or whether it should be allowed to rise slightly.

One consideration is the definition of a family of comparisons. Let's say you conducted a study in which you were interested in whether there was a difference between male and female babies in the age at which they started crawling. After you finished analyzing the data, a colleague of yours had a totally different research question: Do babies who are born in the winter differ from those born in the summer in the age they start crawling? Should the EER be controlled or should it be allowed to be greater than 0.05? A compelling argument can be made that there is no reason you should be penalized (by lower power) just because your colleague used the same data to address a different research question.
Orthogonal Comparisons

When comparisons among means provide independent information, the comparisons are called "orthogonal." If an experiment with four groups were conducted, then a comparison of Groups 1 and 2 would be orthogonal to a comparison of Groups 3 and 4. There is nothing in the comparison of Groups 1 and 2 that provides information about the comparison of Groups 3 and 4. These two comparisons are orthogonal.

Now consider the following two comparisons: Group 1 with Group 2 and Group 1 with the mean of Groups 2 and 3. These two comparisons are clearly not orthogonal: both involve a comparison of Groups 1 and 2, although the second comparison also involves Group 3. If Group 1 is larger than Group 2, then it is probably (but not necessarily) larger than the mean of Groups 2 and 3. The information conveyed by the two comparisons overlaps; the comparisons are not independent.

There is a simple rule for determining if two comparisons are orthogonal: they are orthogonal if and only if \( \Sigma a_i b_i = 0 \)

where \( a_i \) is the ith coefficient of the first comparison and \( b_i \) is the ith coefficient of the second comparison.

Again consider the comparisons:

Group 1 with Group 2
Group 1 with the average of Groups 2 and 3

The coefficients for the first comparison are: 1, -1, 0, 0.
The coefficients for the second comparison are: 1, -.5, -.5, 0.

\[ \Sigma a_i b_i = (1)(1) + (-1)(-.5) + 0 + 0 \neq 0. \]

Therefore, these two comparisons are not orthogonal. For the comparisons:

Group 1 with Group 2
Group 3 with Group 4

the coefficients are: 1, -1, 0 and 0, 0, 1, -1. Therefore,

\[ \Sigma a_i b_i = (1)(0) + (-1)(0) + (0)(1) + (0)(-1) = 0 \]

and the comparisons are orthogonal.

If there are "a" groups in an experiment, then it is possible to make up a-1 mutually orthogonal comparisons among the means. There are many ways to make up a set of a-1 orthogonal comparisons. There is no way to make up a set of more than a-1 comparisons. Two sets of 3 mutually orthogonal comparisons among four means are shown below.

<table>
<thead>
<tr>
<th></th>
<th>Set 1</th>
<th>Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean 2</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>Mean 3</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Mean 4</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>
The value of \( \Sigma a_i b_i \) for any pair of comparisons in either set is zero. For instance, for Comparisons 2 and 3 of Set 1,

\[
\Sigma a_i b_i = (1)(1) + (-1)(-1) + (1)(-1) + (-1)(1) = 1 + 1 + (-1) + (-1) = 0.
\]

If a set of \( a-1 \) orthogonal comparisons is constructed and the sum of squares for each comparison is computed, then the sum of the \( a-1 \) sums of squares will be equal to the sum of squares between in the one factor analysis of variance. For example, consider the data described in the section on partitioning the sums of squares. The means for the three groups are:

\[
M_1 = 4 \\
M_2 = 3 \\
M_3 = 2, \\
n = 4, and,
\]
as previously computed, \( SSB = 8 \).

The SSB can also be computed by making up two orthogonal comparisons among the three means and adding the sums of squares associated with each. A set of orthogonal comparisons is:

<table>
<thead>
<tr>
<th>Mean 1</th>
<th>2</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean 2</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Mean 3</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

The formula for each sum of square is:

\[
SSB = \frac{L^2}{\sum \frac{a_i^2}{n_i}}
\]

For this example,

\[
L_1 = (2)(4)+ (-1)(3) + (-1)(2) = 3
\]

\[
\sum \frac{a_i^2}{n_i} = \frac{(2)^2}{4} + \frac{(-1)^2}{4} + \frac{(-1)^2}{4} = 1.5
\]

\( SSB_1 = 3^2/1.5 = 6 \)

\[
L_2 = (0)(4)+ (1)(3) + (-1)(2) = 1
\]

\[
\sum \frac{a_i^2}{n_i} = \frac{(0)^2}{4} + \frac{(1)^2}{4} + \frac{(-1)^2}{4} = 0.5
\]

\( SSB_2 = 1^2/5 = 2 \)

\( SSB = SSB_1 + SSB_2 = 6 + 2 = 8 \) which is the same value obtained in the analysis of variance.

It is important to know whether the comparisons you are conducting are orthogonal or not. However, it is not critical that you limit yourself to orthogonal comparisons. It is much more important to test the comparisons that make sense in terms of your experimental hypotheses.
Trend Analysis

If an experiment contains a **quantitative independent variable**, then the shape of the function relating the levels of this quantitative independent variable to the **dependent variable** is often of interest. For instance, consider a hypothetical experiment on the effect of the magnitude of reward on the time it takes rats to run down an alley. Running time as a function of the number of food pellets provided as the reward is shown in the graph below. The graph shows that running time decreased as reward size increased.

![Graph of running time vs. number of pellets](image)

Trend analysis can be used to test different aspects of the shape of the function relating the independent variable (Number of pellets in this example) and the dependent variable (running time). Trend analysis consists of testing one or more components of trend. These components are tested using **specific comparisons**. The linear component of trend is used to test whether there is an overall increase (or decrease) in the dependent variable as the independent variable increases. The graph below shows that running time decreased as magnitude of reward increased.

![Graph of running time vs. number of pellets](image)

A test of the linear component of trend is a test of whether this decrease in running time is **significant**. If there were a perfectly linear **relationship** between magnitude of reward and running time, then no components of trend other than the linear would be present. The figure shows, however, that the relationship is not perfectly linear: The slope of the function decreases as magnitude of reward increases. Running time decreases four seconds as the number of pellets increases from two to four, two seconds as the number of pellets increases from four to six, and only one second as the number of pellets increases from six to eight.
The quadratic component of trend is used to test whether the slope increases (or decreases) as the independent variable increases. Components of trend beyond the quadratic are not usually psychologically interpretable. The cubic component, for example, tests whether the slope changes twice (decreasing and then increasing or increasing and then decreasing) as the independent variable increases.

Trend analysis is computed as a set of orthogonal comparisons using a particular set of coefficients. Coefficients for the linear, quadratic, and cubic components of trend are given below:

2 means
Lin  -1  1

3 means
Lin  -1  0  1
Quad  1  -2  1

4 means
Lin  -3  -1  1  3
Quad  1  -1  -1  1
Cubic  3  -3  3  1

5 means
Lin  -2  -1  0  1  2
Quad  2  -1  -2  -1  2
Cubic  1  2  0  -2  1

6 means
Lin  -5  -3  -1  1  3  5
Quad  5  -1  -4  -4  -1  5
Cubic  -5  7  4  4  7  5

7 means
Lin  -3  -2  -1  0  1  2  3
Quad  5  0  -3  -4  -3  0  5
Cubic  1  1  0  -1  -1  1

8 means
Lin  -7  -5  -3  -1  1  3  5  7
Quad  7  1  -3  -5  -3  1  7  7
Cubic  -7  5  7  3  -3  -7  -5  7

9 means
Lin  -4  -3  -2  -1  0  1  2  3  4
Quad  28  7  -8  -17  -20  -17  -8  7  28
Cubic  -14  7  13  9  0  -9  -13  -7  14

10 means
Lin  -9  -7  -5  -3  -1  1  3  5  7  9
Quad  6  2  -1  -3  -4  -4  -3  -1  2  6
Cubic  -42  14  35  31  12  -12  -31  -35  -14  42
Assume that the MSE for the magnitude of reward experiment were 5 and that there were 10 subjects per group. The group means are:

<table>
<thead>
<tr>
<th>Reward size</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>

The table on the previous page shows that the coefficients for the linear component of trend with four groups are:

-3, -1, 1, 3.

Applying the formulas for a comparison presented elsewhere:

\[ L = \sum M_i a_i = (10)(-3) + (6)(-1) + (4)(1) + (3)(3) = -23 \]

\[ t = \frac{-23}{3.162} = -7.27 \]

The degrees of freedom for the t is equal to the degrees of freedom error in the analysis of variance which is: \( N - a = 40 - 4 = 36 \). A t table can be used to find that the probability value (p) is less than .0001. Therefore, the decrease in running time associated with the increase in magnitude of reward is significant. Rats run faster if the reward is larger.

The quadratic component of trend is used to test whether the decreasing slope that shows up as a flattening out of the function is significant. The table on a previous page shows that the coefficients for the quadratic component of trend with four groups are: 1, -1, -1, 1.

\[ L = \sum M_i a_i = (10)(1) + (6)(-1) + (4)(-1) + (3)(1) = 3 \]

\[ t = \frac{3}{1.414} = 2.12, \text{ df } = 36, \text{ and } p = .041. \] The quadratic component of trend is significant

---

**Formal Model**
In an analysis of variance design, each score in the "a" populations can be defined in terms of the following formal model:

\[ y_{ij} = \mu_g + \alpha_j + \varepsilon_{ij} \]

where:

- \( y_{ij} \) is the score of the ith subject in the jth population,
- \( \mu_g \) is the mean of all the population means,
- \( \alpha_j \) equals \( \mu_j - \mu_g \) where \( \mu_j \) is the mean of the jth population,
- \( \alpha_j \) represents the effect of being in the jth population, where \( \sum \alpha_j = 0 \),
- \( \varepsilon_{ij} \) is the sum of all other effects on the ith person in the jth population and is referred to as error.

The error within each population is assumed to be normally distributed and have a mean of zero. The error variances within the "a" populations are assumed to be equal.

**Math Textbooks**

The formal model shows that each score is assumed to be the sum of three components: the overall height or elevation of the scores (\( \mu_g \)), the elevation or depression resulting from being in the jth population (\( \mu_j \)), and error (\( \varepsilon_{ij} \)). In terms of this model, the null hypothesis tested by ANOVA is that all "a" values of \( \alpha_j = 0 \). That is, each score is equal to the mean of the population means plus error.

Mathematical statisticians use the formal model in many of their derivations. An example is the derivation of the expected value of MSB.

**Expected Mean Squares for a One-factor between-subject ANOVA**

As stated in another section, both the numerator and the denominator of the F ratio estimate the population variance when the null hypothesis is true. Since they are both unbiased estimates, the expected value of both MSB and MSE is \( \sigma^2 \). Symbolically, \( E[MSB] = E[MSE] = \sigma^2 \). This section covers the expected value of MSB and MSE when the null hypothesis is false.

The MSB is based upon the sample means: the greater the variance of the sample means, the greater the MSB. Therefore, when the null hypothesis is false and the population means are not equal, the expected value of MSB is greater than when the null hypothesis is true. It stands to reason that the more the population means differ from each other, the greater the expected value of MSB. Mathematical statisticians have derived the following formula for the expected value of MSB:

\[
E[MSB] = \sigma^2 + \frac{n \sum (\mu_i - \bar{\mu})^2}{a - 1}
\]

where \( \sigma^2 \) is the population variance, \( \mu_i \) is the ith population mean, \( \bar{\mu} \) is the mean of the population means, \( n \) is the number of subjects in each group, and "a" is the number of population means.

**Math Textbooks**

For example, if \( \sigma^2 = 100 \), \( n = 10 \), \( a = 3 \), \( \mu_1 = 5 \), \( \mu_2 = 6 \), and \( \mu_3 = 10 \) then \( \bar{\mu} = 7 \) and

\[
E[MSB] = 100 + \frac{10[(5 - 7)^2 + (6 - 7)^2 + (10 - 7)^2]}{3 - 1} = 170.
\]
The expected value of MSB is $\sigma^2$ when the null hypothesis is true but is equal to a larger value when the null hypothesis is false. $E[MSE] = \sigma^2$ whether or not the null hypothesis is true.

For the example, $E[MSE] = 100$. Whenever the null hypothesis is false, $E[MSB] > E[MSE]$ and relatively large values of $F = MSB/MSE$ can be expected. Although you will rarely need to know the expected value of MSB or MSE, it is important to see that both expected values are the same when the null hypothesis is true and that the expected value of MSB is larger when the null hypothesis is false.

**Reporting Results**

Results should be described as simply and as free of statistical jargon as possible. It is best to start by presenting a graph or table that portrays the descriptive statistics. Then, describe the relevant findings in simple plain English. Finally state which effects were statistically significant. A report of a the hypothetical study comparing the effect of background noise might read as follows:

<table>
<thead>
<tr>
<th>Condition</th>
<th>M</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Noise</td>
<td>8.90</td>
<td>2.47</td>
</tr>
<tr>
<td>Moderate Noise</td>
<td>5.60</td>
<td>2.50</td>
</tr>
<tr>
<td>Loud Noise</td>
<td>2.60</td>
<td>2.22</td>
</tr>
</tbody>
</table>

As can be seen in Table 1, background noise had a substantial effect on performance: the louder the background noise, the lower the performance.
Box plots further illustrating the differences are shown in Figure 1. An analysis of variance was conducted and the effect of noise was significant, F(2,27) = 5.94, p = 0.007. The Tukey HSD procedure revealed that all pairwise differences among means were significant, p < 0.05.

The first sentence sums up the findings. The reader is referred to the table and the figure to see just how large the effect of noise was. The rest of the report contains the results of the statistical analyses. It shows that the differences were not due to chance. If possible, the section on the statistical analyses should be written so that it can be skipped without loss of continuity. The symbol "F(2,27) = 5.94" means that the F was computed to be 5.94, that the degrees of freedom numerator were 2, and the degrees of freedom denominator were 27.

In general, the analysis of variance summary table is not reported, although it is sometimes included in an appendix. When planned comparisons are reported, make sure to specify that they were planned. If the rationale for a planned comparison is not obvious, it should be spelled out. Otherwise, the reader will have difficulty believing that it was really planned.

For apply the exercises at following link:
http://davidmlane.com/hyperstat/intro_ANOVA.html
CHAPTER 13

Factorial between-subject ANOVA

Basic Definitions
Factorial Design

When an experimenter is interested in the effects of two or more independent variables, it is usually more efficient to manipulate these variables in one experiment than to run a separate experiment for each variable. Moreover, only in experiments with more than one independent variable is it possible to test for interactions among variables. Consider a hypothetical experiment on the effects of a stimulant drug on the ability to solve problems. There were three levels of drug dosage: 0 mg, 100 mg, and 200 mg. A second variable, type of task, was also manipulated. There were two types of tasks: a simple well-learned task (naming colors) and a more complex task (finding hidden figures in a complex display). The mean time to complete the task for each condition in the experiment is shown below:

<table>
<thead>
<tr>
<th>Dose</th>
<th>Simple Task</th>
<th>Complex Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 mg</td>
<td>32</td>
<td>80</td>
</tr>
<tr>
<td>100 mg</td>
<td>25</td>
<td>91</td>
</tr>
<tr>
<td>200 mg</td>
<td>21</td>
<td>96</td>
</tr>
</tbody>
</table>

As you can see, each level of dosage is paired with each level of type of task. The number of conditions (six) is therefore the product of the number of levels of dosage (three) and type of task (two).

Experimental designs in which every level of every variable is paired with every level of every other variable are called factorial designs. In this example, every level of dosage (0 mg, 100 mg, and 200 mg) is paired with every level of type of task (simple and complex). Also note that this implies that every level of type of task is paired with every level of dosage. This experiment on dosage and type of task is described as a Dosage (3) x Type of task (2) between-subjects factorial design.

Factorial designs can also contain more than two variables. For example, an Age (2) x Dosage (3) x Type of task (2) factorial design would consist of the following $2 \times 3 \times 2 = 12$ conditions.
Main Effect

The main effect of an independent variable is the effect of the variable averaging over all levels of other variables in the experiment. The means from the hypothetical experiment described in the section on factorial designs are reproduced below.

<table>
<thead>
<tr>
<th>Dose</th>
<th>Simple Task</th>
<th>Complex Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 mg</td>
<td>32</td>
<td>80</td>
</tr>
<tr>
<td>100 mg</td>
<td>25</td>
<td>91</td>
</tr>
<tr>
<td>200 mg</td>
<td>21</td>
<td>96</td>
</tr>
</tbody>
</table>

The main effect of type of task is assessed by computing the mean for the two levels of type of task averaging across all three levels of dosage. The mean for the simple task is: \((32 + 25 + 21)/3 = 26\) and the mean for the complex task is: \((80 + 91 + 95)/3 = 86.67\). The main effect of type of task, therefore, involves a comparison of the mean of the simple task (26) with the mean of the complex task (86.67).

Analysis of variance provides a significance test for the main effect of each variable in the design. If the main effect of type of task were significant, then the null hypothesis that there is no difference between the simple and complex tasks would be rejected. Similarly, if the main effect of drug dosage were significant, then the null hypothesis that there is no effect of drug dosage would be rejected.

Interaction

Two independent variables interact if the effect of one of the variables differs depending on the level of the other variable. The means from the hypothetical experiment described in the section on factorial designs are reproduced below.

<table>
<thead>
<tr>
<th>Dose</th>
<th>Simple Task</th>
<th>Complex Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 mg</td>
<td>32</td>
<td>80</td>
</tr>
<tr>
<td>100 mg</td>
<td>25</td>
<td>91</td>
</tr>
<tr>
<td>200 mg</td>
<td>21</td>
<td>96</td>
</tr>
</tbody>
</table>

Notice that the effect of drug dosage differs depending on whether the task is simple or complex. For the simple task, the higher the dosage, the shorter the time to complete the task. For the complex task, the higher
the dosage, the longer the time to complete the task. Thus, there is an interaction between dosage and task complexity. It is usually much easier to interpret an interaction from a graph than from a table. A graph of the means for the interaction between task complexity and drug dosage is shown below.

The dependent variable (response time) is shown on the Y axis. The levels of drug dosage are shown on the X axis. The two levels of task complexity are graphed separately.

A look at the above graph shows that the effect of dosage differs as a function of task complexity. It also shows that the effect of task complexity differs as a function of drug dosage: The larger the drug dosage, the greater the difference between the simple task and the complex task. An interaction does not necessarily imply that the direction of an effect is different at different levels of a variable. There is interaction as long as the magnitude of an effect is greater at one level of a variable than at another. In the example, the complex task always takes longer than the simple task. There is an interaction because the magnitude of the difference between the simple and complex tasks is different at different levels of the variable drug dosage.

Two variables interact if a particular combination of variables leads to results that would not be anticipated on the basis of the main effects of those variables. For instance, it is known that both drinking alcohol and smoking increase the chance of throat cancer. However, people who both drink and smoke have a much higher chance of getting cancer than would be predicted if one knew only how much more likely smokers are than nonsmokers to get throat cancer and how much more likely drinkers are than nondrinkers to get throat cancer. The combination of smoking and drinking is particularly dangerous: these drugs interact.
This definition of interaction in terms of a particular combination of variables is consistent with the previously-given definition that there is an interaction if the effect of one variable differs depending on the level of another variable. In the tobacco and alcohol example, the effect of smoking on the probability of getting cancer is greater for people who drink than for people who do not drink: the effect of smoking differs depending on whether drinkers or nondrinkers are being considered. Similarly, the effect of drinking differs depending on whether smokers or nonsmokers are being considered.

Interactions can be described in a variety of ways. Examples of graphs of interactions and possible verbal descriptions of each follow.

1. The difference between the treatment and control conditions was greater for subjects performing Task 1 than for subjects performing Task 2.

2. There was a greater difference between Task 1 and Task 2 for subjects in the treatment condition than there was for subjects in the control condition.

3. Tasks 1 and 2 were performed about equally well in the control condition but Task 1 was performed considerably better than Task 2 in the treatment condition.

1. On the well-learned task, increased magnitude of reward was associated with better performance whereas on the novel task, increased magnitude of reward was associated with reduced performance.

2. Under low magnitude of reward, the novel task was performed better than the well-learned task. Under high magnitude of reward, the well-learned task was performed better than the novel task.

3. The difference between the novel task and the well-learned task changed from positive for low magnitude of reward to slightly negative for medium magnitude of reward to very negative for high magnitude of reward.
1. Overall, Condition B₂ led to better performance than did either B₁ or B₃. This effect was much more pronounced for subjects performing Task 1 than for subjects performing Task 2.

2. The difference between Tasks 1 and 2 was greatest for subjects in Condition B₂.

3. The combination of Task 1 and Condition B₂ led to especially high performance.

In Example 4 there is no interaction. The effect of task is the same at all three levels of B and the effect of B is the same for both tasks. Notice that the two lines are parallel. When there is no interaction, the lines will always be parallel.

**Higher Order Interactions**

So far, all the interactions that have been described are called "two-way" interactions. They are two-way interactions because they involve the interaction of two variables. A three-way interaction is an interaction among three variables. There is a three-way interaction whenever a two-way interaction differs depending on the level of a third variable. Consider the two figures on the left side of this page. The upper figure shows the interaction between task and condition (B) for well-rested subjects; the lower figure shows the same interaction for sleep-deprived subjects. The forms of these interactions are different. For the well-rested subjects, the difference between Tasks 1 and 2 is largest under condition B₂ whereas for the sleep-deprived subjects the difference between Tasks 1 and 2 is smallest under condition B₂. The two-way interactions are therefore different for the two levels of the variable "sleep deprivation." This means that there is a three-way interaction among the variables sleep deprivation, task, and condition.

Four-way interactions occur when three-way interactions differ as a function of the level of a fourth variable. Four-way and higher interactions are usually very difficult to interpret and are rarely meaningful.
A two-factor analysis of variance consists of three significance tests: a test of each of the two main effects and a test of the interaction of the variables. An analysis of variance summary table is a convenient way to display the results of the significance tests. A summary table for the hypothetical experiment described in the section on factorial designs and a graph of the means for the experiment are shown below.

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>df</th>
<th>Squares</th>
<th>Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>1</td>
<td>47125.3333</td>
<td>47125.3333</td>
<td>384.174</td>
<td>0.000</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>42.6667</td>
<td>21.3333</td>
<td>0.174</td>
<td>0.841</td>
</tr>
<tr>
<td>TD</td>
<td>2</td>
<td>1418.6667</td>
<td>709.3333</td>
<td>5.783</td>
<td>0.006</td>
</tr>
<tr>
<td>ERROR</td>
<td>42</td>
<td>5152.0000</td>
<td>122.6667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>47</td>
<td>53738.6667</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sources of Variation**
The summary table shows four sources of variation: (1) Task, (2) Drug dosage, (3) the Task x Drug dosage interaction, and (4) Error.

**Degrees of Freedom**
The degrees of freedom total is always equal to the total number of numbers in the analysis minus one. The experiment on task and drug dosage had eight subjects in each of the six groups resulting in a total of 48 subjects. Therefore, df total = 48 - 1 = 47.

The degrees of freedom for the main effect of a factor is always equal to the number of levels of the factor minus one. Therefore, df task = 2 - 1 = 1 since there were two levels of task (simple and complex). Similarly, df dosage = 3 - 1 = 2 since there were three levels of drug dosage (0 mg, 100 mg, and 200 mg).

The degrees of freedom for an interaction is equal to the product of the degrees of freedom of the variables in the interaction. Thus, the degrees of freedom for the Task x Dosage interaction is the product of the degrees of freedom for task (1) and the degrees of freedom for dosage (2). Therefore, df Task x Dosage = 1 x 2 = 2.

The degrees of freedom error is equal to the degrees of freedom total minus the degrees of freedom for all the effects. Therefore, df error = 47 - 1 - 2 - 2 = 42.

**Sums of Squares**
Computational formulas for the sums of squares will not be given since it is assumed that
complex analyses will not be done by hand.

**Mean Squares**
As in the case of a one-factor design, each mean square is equal to the sum of squares divided by the degrees of freedom. For instance, Mean square dosage = 42.6667/2 = 21.3333 where the sum of squares dosage is 42.6667 and the degrees of freedom dosage is 2.

**F Ratios**
The F ratio for an effect is computed by dividing the mean square for the effect by the mean square error. For example, the F ratio for the Task x Dosage interaction is computed by dividing the mean square for the interaction (709.3333) by the mean square error (122.6667). The resulting F ratio is: \( F = 709.3333/122.6667 = 5.783 \).

**Probability Values**
To compute a probability value for an F ratio, you must know the degrees of freedom for the F ratio. The degrees of freedom numerator is equal to the degrees of freedom for the effect. The degrees of freedom denominator is equal to the degrees of freedom error. Therefore, the degrees of freedom for the F ratio for the main effect of task are 1 and 42, the degrees of freedom for the F ratio for the main effect of drug dosage are 2 and 42, and the degrees of freedom for the F for the Task x Dosage interaction are 2 and 42.

An F distribution calculator can be used to find the probability values. For the interaction, the probability value associated with an F of 5.783 with 2 and 42 df is 0.006.

**Drawing Conclusions**
When a main effect is significant, the null hypothesis that there is no main effect in the population can be rejected. In this example, the effect of task was significant. Therefore it can be concluded that, in the population, the mean time to complete the complex task is greater than the mean time to complete the simple task (hardly surprising). The effect of dosage was not significant. Therefore, there is no convincing evidence that the mean time to complete a task (in the population) is different for the three dosage levels.
The significant Task x Dosage interaction indicates that the effect of dosage (in the population) differs depending on the level of task. Specifically, increasing the dosage slows down performance on the complex task and speeds up performance on the simple task. The effect of increasing the dosage therefore depends on whether the task is complex of simple.

There will always be some interaction in the sample data. The significance test of the interaction lets you know whether you can infer that there is an interaction in the population.

**Analysis of Three-factor Designs**

A three-factor analysis of variance consists of seven significance tests: a test for each of the three main effects, a test for each of the three two-way interactions, and a test of the three-way interaction. Consider a hypothetical experiment investigating how well children and adults remember. There are three factors in the experiment:

1. Type of test: Subjects are either asked to recall or to recognize the stimuli. In the recall test, subjects are asked to state the names of as many of the stimuli as they can. On each recognition test trial, subjects are asked to pick out the one stimulus that had been presented from a set of four stimuli.
The experiment can therefore be described as an Age (2) x Type of stimulus (2) x Type of test (2) factorial design. There were five subjects in each of the eight conditions. The data are shown on the next page.

<table>
<thead>
<tr>
<th>Recognition</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>Adults</td>
</tr>
<tr>
<td>Pictures</td>
<td>Words</td>
</tr>
<tr>
<td>92</td>
<td>70</td>
</tr>
<tr>
<td>90</td>
<td>70</td>
</tr>
<tr>
<td>92</td>
<td>71</td>
</tr>
<tr>
<td>91</td>
<td>70</td>
</tr>
<tr>
<td>90</td>
<td>72</td>
</tr>
</tbody>
</table>

The ANOVA summary table is shown below. "T" stands for Type of task (recall or recognition), "A" stands for Age (children or adults) and "S" stands for type of stimulus (pictures or words). All main effects and interactions are significant, p < 0.01.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>1</td>
<td>1000.00</td>
<td>1000.00</td>
<td>273.038</td>
<td>0.000</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1960.00</td>
<td>1960.00</td>
<td>535.154</td>
<td>0.000</td>
</tr>
<tr>
<td>TA</td>
<td>1</td>
<td>518.40</td>
<td>518.40</td>
<td>141.543</td>
<td>0.000</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>902.50</td>
<td>902.50</td>
<td>246.416</td>
<td>0.000</td>
</tr>
<tr>
<td>TS</td>
<td>1</td>
<td>202.50</td>
<td>202.50</td>
<td>55.290</td>
<td>0.000</td>
</tr>
<tr>
<td>AS</td>
<td>1</td>
<td>220.90</td>
<td>220.90</td>
<td>60.314</td>
<td>0.000</td>
</tr>
<tr>
<td>TAS</td>
<td>1</td>
<td>28.90</td>
<td>28.90</td>
<td>7.891</td>
<td>0.008</td>
</tr>
<tr>
<td>Error</td>
<td>32</td>
<td>117.20</td>
<td>3.6625</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>4950.40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The means are graphed above. The left graph is for the recognition task; the right graph is for the recall task. The effects in the design can be described as follows:
1. Task: The percent correct was higher for the recognition task than for the recall task.
2. Age: Adults performed better than children.
3. Task x Age interaction: The effect of age was greater for the recall task than for the recognition task.
4. Type of stimulus: Memory was better for pictures than for words.
5. Task x Type of stimulus interaction: The difference between pictures and words was larger for the recognition task than for the recall task.
6. Age x Type of stimulus interaction: The effect of age was larger for the words than it was for the pictures.
7. Task x Age x Type of stimulus interaction: The Age x Type of stimulus interaction was larger for the recognition task than for the recall task. With the recall task, the difference between children and adults was only slightly smaller for pictures than for words. With the recognition task, the difference between children and adults was much smaller for pictures than for words. There was essentially no difference between children and adults for the pictures whereas there was a large difference between children and adults for the words.
8. Hypothetical data from a two-factor design with three levels of Factor A and two levels of factor B are shown below:

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>Marginal means</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>7.08</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>8.25</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Mean = 5</td>
<td>Mean = 8</td>
<td>Mean = 8.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>6.50</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Mean = 5.25</td>
<td>Mean = 6.75</td>
<td>Mean = 7.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal Means</td>
<td>5.125</td>
<td>7.375</td>
<td>7.875</td>
<td>6.79</td>
</tr>
</tbody>
</table>

The ANOVA summary table is shown below:
STATISTICS
http://davidmlane.com/hyperstat/index.html

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SSq</th>
<th>Ms</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>34.333</td>
<td>17.17</td>
<td>9.29</td>
<td>0.0017</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>2.042</td>
<td>2.04</td>
<td>1.10</td>
<td>0.3070</td>
</tr>
<tr>
<td>AB</td>
<td>2</td>
<td>2.333</td>
<td>1.167</td>
<td>0.63</td>
<td>0.5431</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>33.250</td>
<td>1.847</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>71.958</td>
<td>3.129</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The marginal mean for A₁ (5.125) is the average of A₁B₁ and A₁B₂ which is: (5.00 + 5.25)/2 = 5.125. Similarly, the marginal mean for A₂ is the average of A₂B₁ and A₂B₂, etc. The significant main effect of A means that the marginal means for A₁, A₂, and A₃ are not all equal. It does not reveal which population means are different from which others. Procedures for following up a significant main effect such as this one are analogous to the procedures used to follow up a significant effect in a one-factor ANOVA. These procedures are:

1. All pairwise comparisons among means
2. Specific comparisons among means
3. Comparing all means with a control

The formulas used in these procedures all involve the sample size (n).

When applying these formulas to multi-factor experiments, n refers to the number of scores in each of the means being compared.

For example, in the formula for the standard error of a comparison:

\[ s_L = \sqrt{n \cdot MSE} \]

n is the sample size of each group. In the present example, there are three groups, A₁, A₂, and A₃ and there are eight subjects ub each of these groups. Therefore, n = 8. All other aspects of the formulas are equivalent

**Supplementing Interaction: When no Follow-up is Needed**

A significant interaction indicates that the effect of one variable differs depending on the level of another variable. Often, this is exactly the information that a researcher needs, and no follow-up analyses are necessary. For example, consider a hypothetical experiment designed to test whether hyperactive and non-hyperactive children are affected differently by the presence of distracting stimuli. The experiment has two factors: child's classification (hyperactive or non-hyperactive) and distraction (distraction present and distraction absent).
A graph of the means for the four conditions is shown below.

The graph shows that the effect of distraction is greater for the hyperactive children than it is for the non-hyperactive children. If the interaction were significant, then the researcher would be able to conclude that distracting stimuli are more disruptive to the hyperactive children than they are to the non-hyperactive children. No additional statistical tests would be needed to justify this conclusion. Some researchers feel that the finding of a significant interaction requires one to do supplemental analyses. No supplemental analyses are required if the researcher wishes simply to conclude that the effect of one variable differs depending on the level of another variable since that is exactly what the interaction is testing.

**Supplementing Interaction: Simple Effects**

The presence of interaction limits the generalizeability of main effects. This is because it is difficult to make a general statement about a variable's effect when the size of the effect depends on the level of a second variable. For example, consider the hypothetical data shown in the figure below. The experiment is on the effect of condition (treatment versus control) on performance for each of two tasks. Since the effect of condition is different for the two tasks, it is not possible to make a general statement about the treatment's effectiveness.
A significant main effect of condition indicates that, on average, the treatment condition leads to better performance than does the control condition. Since the effect of condition is different for the two treatments, a significant main effect of condition does not necessarily imply an effect of condition for both tasks. It might be that, in the population, there is an effect of condition for Task 1 but not for Task 2. If the researcher wished to know whether there was an effect of condition for both tasks, he or she could not rely on the test of the main effect. Instead, the researcher would test the significance of the effect of condition separately for the two tasks. The effect of a variable at a specific level of another variable is called a "simple effect" of the variable. In this example there are two simple effects of condition: the effect of condition for Task 1 and the effect of condition for Task 2. The presence of interaction means that the main effect is not representative of the simple effects. If you wish to know whether a variable has an effect at each level of a second variable, you should test the simple effects.

Some researchers have the misconception that testing simple effects is necessary for understanding the interaction. This is not true since the interaction is interpretable in its own right. The rationale for testing simple effects, rather, is simply that sometimes it is important to know at what levels of a second variable the variable in question has a significant effect. Testing simple effects is done following an interaction not to help understand the interaction, but rather to see where the effect of the variable is significantly different from zero.

Components of Interaction

Consider the graph shown below.

![Graph with A1, A2, A3 on the x-axis and Number Correct on the y-axis with two curves, B1 and B2.]  

Overall, there is an interaction since the effect of B depends on the level of A. However, only one piece or component of the interaction is present: The effect of B is greater at A3 than it is at A1 or A2. No other part of the interaction is present since the effect of B is the same at A1 as it is at A2.

It is possible to test components of interaction for significance by using the procedures described in the section on specific comparisons among means. The test of whether the effect of B is significantly different at A3 then it is at the other two levels is done with the comparison among means shown on the next page.

In this comparison, the difference between B1 and B2 at A3 is compared with the average of the difference between B1 and B2 at A1 and the difference between B1 and B2 at A2:

$$A_3B_1 - A_3B_2 = \frac{(A_1B_1 - A_1B_2) + (A_2B_1 - A_2B_2)}{2}$$
The coefficients for the comparison can be determined by distributing the negative sign and rearranging the expression. The coefficients are:

\[ \begin{align*}
A_1B_1 & : -0.5 \\
A_1B_2 & : 0.5 \\
A_2B_1 & : -0.5 \\
A_2B_2 & : 0.5 \\
A_3B_1 & : 1.0 \\
A_3B_2 & : -1.0
\end{align*} \]

Once the coefficients for a component of the interaction are determined, the comparison is tested just like any other specific comparison among means.

As with other comparisons among means, different procedures are used if the comparison is planned than if it is unplanned. There are as many components to an interaction as the interaction has degrees of freedom. In the present example, the interaction has two degrees of freedom.

The remaining component of the interaction is whether the difference between \( B_1 \) and \( B_2 \) is different at \( A_1 \) than it is at \( A_2 \). This can be expressed as:

\[ (A_1B_1 - A_1B_2) - (A_2B_1 - A_2B_2) \]

The coefficients for this comparison are:

\[ \begin{align*}
A_1B_1 & : 1 \\
A_1B_2 & : -1 \\
A_2B_1 & : -1 \\
A_2B_2 & : 1 \\
A_3B_1 & : 0 \\
A_3B_2 & : 0
\end{align*} \]

For the example, this component of interaction is zero.

### Reporting Results in Factorial Between-Subjects ANOVA

Results should be described as simply and as free of statistical jargon as possible. Begin with a presentation of descriptive statistics. The descriptive statistics may be presented numerically, graphically, or both. The results of the analysis of variance should be discussed with reference to a graph of the group means. First note whether or not there is an interaction. Describe the relevant outcomes and back up any claims with the results of statistical tests. Do not let the statistical analysis become the focus of the discussion. Instead, focus the discussion on the graph of the means and use the statistical analysis as a way to substantiate the effects you point out in the graph. For example, consider the following hypothetical experiment on age differences in memory for words and pictures.

Sixteen 8-year-old children and 16 12-year-old children were shown a set of stimuli and later given a test to see how well they could recognize the stimuli that had been presented. Half children at each age level were presented with word stimuli; the other half were presented with pictorial stimuli. The percentage correct on the test was recorded for each child.
The results of the experiment might be written up as follows:

**Results**

Measures of central tendency and variability for the four experimental groups are shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Eight-Year Olds</th>
<th>Twelve-Year Olds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Words</td>
<td>Pictures</td>
</tr>
<tr>
<td>Mean</td>
<td>63.38</td>
<td>84.12</td>
</tr>
<tr>
<td>Trimean</td>
<td>61.22</td>
<td>84.50</td>
</tr>
<tr>
<td>SD</td>
<td>5.83</td>
<td>7.45</td>
</tr>
<tr>
<td>Semi IQR</td>
<td>2.75</td>
<td>5.00</td>
</tr>
</tbody>
</table>

In all four groups, the mean and trimean are essentially the same. Therefore, only the means were considered in further analyses.

An Age (2) x Type of stimuli (2) analysis of variance was used to test differences between means for significance. Figure 1 shows the mean percent correct on the recognition test as a function of age and type of stimuli. An inspection of figure 1 reveals that the effect of age was much larger for the word stimuli than it was for the pictorial stimuli. The difference was reflected in a significant Age x Type of stimuli interaction, $F(1,28) = 6.37, p = 0.018$. It is also evident from figure 1 that, overall, 12-year olds performed better than 8-year olds, $F(1,28) = 9.35, p < 0.01$.

![Figure 1. Mean Percent Correct as a function of type of stimulus and age.](image)

An analysis of simple effects showed that this age effect was significant for the word stimuli, $F(1,28) = 15.56, p < 0.01$, but not for the pictorial stimuli, $F(1,28) = 0.14, p = 0.71$. Therefore, there is no evidence that 12-year olds differ from 8-year olds in their ability to recognize pictures. Finally, pictures were recognized better than words, $F(1,28) = 35.05, p < 0.01$. An analysis of simple effects showed that the advantage of words over pictures was significant for both the 8-year olds, $F(1,28) = 26.46, p < 0.01$, and for the 12-year olds, $F(1,28) = 5.77, p = 0.023$.

For apply the exercises at following link:
Chapter 14

Within-subjects ANOVA
Within-Subjects Design

Within-subject designs are designs in which one or more of the independent variables are within-subject variables. Within-subject designs are often called repeated-measures designs since within-subjects variables always involve taking repeated measurements from each subject. Within-subject designs are extremely common in psychological and biomedical research.

Advantages of Within-Subject Designs

Subjects inevitably differ from one another. In an experiment on children's memory, some children will remember more than others; in an experiment on depression, some subjects will be more depressed than others; in an experiment on weight control, some subjects will be heavier than others. It is simply a fact of life that subjects differ greatly. In between-subject designs, these differences among subjects are uncontrolled and are treated as error. In within-subject designs, the same subjects are tested in each condition. Therefore, differences among subjects can be measured and separated from error. For example, consider the following data:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Control</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>57</td>
</tr>
</tbody>
</table>

Every subject did better in the experimental condition than in the control condition. Even though the advantage of the experimental condition is small, it is very likely real since it is very consistent across the four subjects. The important point is that this small but consistent difference can be detected in the face of large overall differences among the subjects. Indeed, the difference between conditions is very small relative to the differences among subjects. It is because the conditions can be compared within each of the subjects that allows the small difference to be apparent. Differences between subjects are taken into account and are therefore not error.

Removing variance due to differences between subjects from the error variance greatly increases the power of significance tests. Therefore, within-subjects designs are almost always more powerful than between-subject designs. Since power is such an important consideration in the design of experiments, within-subject designs are generally preferable to between-subject designs.

The Problem of Carryover Effects

Carryover effects can pose difficult problems for within-subject designs. Sometimes the problems are so severe that a within-subjects design is invalid and a between-subjects design must be used. For example, consider the study of incidental learning. In incidental learning, subjects are often presented with stimuli and asked to answer questions about them. They might be asked to count the number of letters in each word or to judge the pleasantness of each word. Subsequently, a surprise memory test is given. An experimenter wishing to test
whether incidental learning was better when subjects counted the letters or made pleasantness judgments would not be able to use a within-subjects design. Once subjects had been given one memory test, the second memory test would not be a surprise.

It is possible to use a within-subjects design even if carryover effects are present as long as the carryover effects are (a) not severe and (b) are symmetric. Consider an experiment comparing the time it takes to read a list of color names with the time it takes to name colors. If a within-subjects design is used, then all subjects are tested in both conditions. Carryover effects are certainly possible in this design. The second task performed may be performed better because of some kind of practice effect or because subjects have become primed to say color names. Alternatively, the second task performed may be performed worse because subjects have become tired or bored. However, in either case, the carryover effects would likely be symmetric.

Counterbalancing can be used to control for symmetric carryover effects. In this experiment, this means simply that half of the subjects would be given the color-name-reading task before the color-naming task and the other half of the subjects would be given the color naming task first.

**Assumptions of Within-subject Designs**

Within-subjects ANOVA assumes that the scores in all conditions are normally distributed. It is also assumed that each subject is sampled independently from each other subject. Naturally, it is not assumed that the scores of a given subject are independent of each other since the whole point of the analysis is that they are dependent.

In addition to the assumption of normality, within-subject analysis of variance is based on assumptions about the variances of the measurements and the correlations among the measurements. Taken together, these assumptions are called the assumption of sphericity. Although a complete description of sphericity is beyond the scope of this text, there is sphericity if (a) the population variances of the repeated measurements are equal and (b) the population correlations among all pairs of measures are equal. Other complex and unusual patterns of variances and correlations can also produce sphericity. Violation of the assumption of sphericity is serious: It results in an increase in the Type I error rate.

Since real data rarely meet the assumption of sphericity, this assumption cannot be safely ignored. In 1954 a statistician named Box developed an index of the degree to which the assumption of sphericity is violated. The index, called epsilon (\(\varepsilon\)) ranges from 1.0, meaning no violation to 1/df where df is the degrees of freedom in the numerator of the F ratio. In the Geisser-Greenhouse correction, the sample value of epsilon can be used to correct the probability value for violations of sphericity by multiplying both the degrees of freedom numerator and denominator by the sample value of \(\varepsilon\). (Note that the corrected df are used only to compute the p value, not to divide the sum of squares in order to find the mean square.) The corrected probability value will always be higher (less significant) than the uncorrected value except when the effect has one degree of freedom in which case \(\varepsilon\) will be 1.0 and the corrected and uncorrected probability values will be the same. Statistical packages often refer to this correction as the Geisser-Greenhouse correction because of an article of theirs in 1957 that discussed this issue.

The "Geisser-Greenhouse correction" is known to be somewhat conservative. An alternative correction developed by Huynh and Feldt is less conservative and is often computed by
standard statistical packages.

Although the assumption of sphericity has been discussed for many years, it is still often ignored in practice. This is unfortunate since an uncorrected probability value from an analysis variance with within-subject variables is very rarely valid.

ANOVA with 1 Within-Subject Variable

A one-factor within-subjects analysis of variance tests the null hypothesis that all the population means are equal: $H_0: \mu_1 = \mu_2 = \ldots = \mu_a$

Sources of Variation
In a between-subjects ANOVA, variance due to differences among subjects goes into the error term. In within-subjects ANOVA, differences among subjects can be separated from error. "Subjects" is therefore a source of variation in within-subjects designs. The analysis of variance summary table for the data given in the section on the advantages of within-subjects designs is shown below. Notice the three sources of variation: Subjects, Condition, and Error.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Ssq</th>
<th>Ms</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>3</td>
<td>1888.375</td>
<td>629.458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>1</td>
<td>15.125</td>
<td>15.125</td>
<td>121.00</td>
<td>0.002</td>
</tr>
<tr>
<td>Error</td>
<td>3</td>
<td>0.375</td>
<td>0.125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>1903.875</td>
<td>271.982</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Degrees of Freedom
The degrees of freedom for subjects is always N-1 where N is the number of subjects. For this problem: N-1 = 4 - 1 = 3.
The degrees of freedom for the independent variable (Condition) is always equal to the number of levels of the independent variable minus 1. For this example, the degrees of freedom is 2 - 1 = 1 since there are two conditions (experimental and control).

The degrees of freedom error is always the product of the degrees of freedom for Subjects and the degrees of freedom for the independent variable. For this example, the degrees of freedom error is equal to 3 x 1 = 3. The degrees of freedom total is equal to the total number of numbers in the analysis minus 1. For this example, there are eight numbers so the degrees of freedom is equal to 8 - 1 = 7. Notice that the degrees of freedom for Subjects, Conditions, and Error add up to the degrees of freedom total 3 + 1 + 3 = 7.

The formulas for degrees of freedom are summarized below:

- df Subjects = N - 1 = 4 - 1 = 3
- df Conditions = a - 1 = 2 - 1 = 1
- df error = (N -1)(a-1) = 3
- df total = aN - 1 = 2 x 4 -1 = 8-1 = 7
a is the number of conditions; N is the total number of subjects.

**Error as Interaction**

The error term in within-subject ANOVA designs is equal to the Subjects x Condition interaction. As such, it is a measure of the degree to which the effect of conditions is different for different subjects. Remember that when variables interact, the effect of one variable differs depending on the level of another variable. The error term is a measure of the degree to which the effect of the variable "Condition" is different for different levels of the variable "Subjects." (Each subject is a different level of the variable "Subjects." ) Low interaction (low error) means that the effect of conditions is consistent across subjects. High interaction (high error) means that the effect of conditions is not consistent across "subjects."

**Sums of Squares**

Computational formulas for the sums of squares will not be given since it is assumed that complex analyses will not be done by hand.

**Mean Squares**

As always, a mean square is computed by dividing the sum of squares by the degrees of freedom. Generally speaking, no mean square is computed for the variable "subjects" since it is assumed that subjects differ from one another thus making a significance test of "subjects" superfluous.

**Relationship between F and t**

The F from a within-subjects design with two levels of the independent variable is equal to the t² from a test of differences between dependent means. This relationship is the same as the relationship between F and t in between-subject designs.

**ANOVA with Two Within-Subject Variables**

Consider the following hypothetical experiment: A total of four rats are run in a maze three trials a day for two days. The number of wrong turns on each trial are shown below.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>

**Sources of Variation**

The sources of variation in this design are: Subjects, Days, Trials, the Days x Trials interaction, and Error. The ANOVA summary table is shown below. Notice that there are three error terms: one for the effect of Days (MSE = 1.38), one for the effect of Trials (MSE = 0.61), and one for the interaction (MSE = 0.17).
**Degrees of Freedom**

The degrees of freedom for Subjects is equal to the number subjects minus one. In this example, this is $N - 1 = 4 - 1 = 3$ since there are 4 subjects.

The degrees of freedom for each main effect is equal to the number of levels of the variable in question minus one. Therefore, since there are two days, the effect of Days has $2 - 1 = 1$ degrees of freedom. There are three trials, so the effect of Trials has $3 - 1 = 2$ degrees of freedom.

The degrees of freedom for the interaction of the two variables is equal to the product of the degrees of freedom for the main effects of these variables. Since Days has 1 df and Trials has 2 df, the Days x Trials interaction has $1 \times 2 = 2$ df. The formulas are summarized below:

\[
\begin{align*}
\text{df Subjects} &= N - 1 = 4 - 1 = 3 \\
\text{df Days} &= d - 1 = 2 - 1 = 1 \\
\text{df error(Days)} &= (N -1)(d-1) = 3 \\
\text{df Trials} &= t - 1 = 3 - 1 = 2 \\
\text{df error(Trials)} &= (N -1)(t-1) = 6 \\
\text{df Days x Trials} &= (d-1)(t-1) = 2 \\
\text{df error (Days x Trials)} &= (N-1)(d-1)(t-1) = 6 \\
\text{df total} &= dtN - 1 = (2)(3)(4) - 1 = 23
\end{align*}
\]

$d$ is the number of days; $t$ is the number of trials (each day); and $N$ is the total number of subjects.

**Sums of Squares**

Computational formulas for the sums of squares will not be given since it is assumed that complex analyses will not be done by hand.

**Mean Squares**

As always, a mean square is computed by dividing the sum of squares by the degrees of freedom. Generally speaking, no mean square is computed for the variable "subjects" since it is assumed that subjects differ from one another thus making a significance test of "subjects" superfluous. F Ratios The F ratio for each effect is the mean square for the test divided by the mean square error for the effect. Since there is a separate error term for each effect, a different mean square error is used in each test.

**Tests Supplementing Within-Subjects ANOVA**

Tests supplementing a within-subjects ANOVA are analogous to tests supplementing to a between-subjects ANOVA. The only differences lie in the computational details which are taken care of for you by standard statistical packages. One important computational difference is that when specific comparisons among means are made on within-subjects variables, an error term for each specific comparison is calculated. The same error term is used for all comparisons when the variable is between-subjects.

**ANOVA with Between- and Within- Subject Variables**
It is common for designs to have a combination of between- and within-subject variables. In this section designs with one between- and one within-subject variable are discussed.

Consider an experiment in which four 8-year-old and four and 12-year-old subjects are given five trials on a motor learning task. There are two variables: age and trials. Age is a between-subject variable since each subject is in either one age group or the other. Trials is a within-subject variable since each subject performs on all five trials.

Sources of Variation
The sources of variation are: age, trials, the Age x Trials interaction, and two error terms. One error term is used to test the effect of age whereas a second error term is used to test the effects of trials and the Age x Trials interaction.

Degrees of Freedom
The degrees of freedom for age is equal to the number of ages minus one. That is: $2 - 1 = 1$. The degrees of freedom for the error term for age is equal to the total number of subjects minus the number of groups: $8 - 2 = 6$. The degrees of freedom for trials is equal to the number of trials - 1: $5 - 1 = 4$. The degrees of freedom for the Age x Trials interaction is equal to the product of the degrees of freedom for age (1) and the degrees of freedom for trials (4) = $1 \times 4 = 4$. Finally, the degrees of freedom for the second error term is equal to the product of the degrees of freedom of the first error term (6) and the degrees of freedom for trials (4): $6 \times 4 = 24$.

The formulas for the degrees of freedom for each effect are summarized below:

- $df \text{ Age} = k - 1 = 2 - 1 = 1$
- $df \text{ error(Age)} = N - k = 8 - 2 = 6$
- $df \text{ Trials} = t - 1 = 5 - 1 = 4$
- $df \text{ Age x Trials} = df \text{ Age} \times df \text{ Trials} = 1 \times 4 = 4$
- $df \text{ error(Trials and Age x Trials)} = df \text{ error(Age)} \times df \text{ Trials} = 6 \times 4 = 24$
- $df \text{ total} = Nt - 1 = 8 \times 5 - 1 = 39$

$k$ is the number of levels of the between-subject variable; $t$ is the number of trials (each day); $N$ is the total number of subjects.

Sums of Squares
Computational formulas for the sums of squares will not be given since it is assumed that complex analyses will not be done by hand.

Mean Squares
As always, a mean square is computed by dividing the sum of squares by the degrees of freedom.

F Ratios
The F ratio for each effect is the mean square for the test divided by the mean square error for the effect.

For apply the exercises at following link:
http://davidmlane.com/hyperstat/within-subjects.html
CHAPTER 16

Chi Square
Testing Differences between $p$ and $\pi$

Another section shows how to use a test based on the normal distribution to see whether a sample proportion ($p$) differs significantly from a population proportion ($\pi$). This section shows how to conduct a test of the same null hypothesis using a test based on the chi square distribution.

The two tests always yield identical results. The advantage of the test based on the chi square distribution is that it can be generalized to more complex situations. In the other section, an example was given in which a researcher wished to test whether a sample proportion of 62/100 differed significantly from an hypothesized population value of 0.5. The test based on $z$ resulted in a $z$ of 2.3 and a probability value of 0.0107. To compute the significance test using chi square, the following table is formed:

<table>
<thead>
<tr>
<th></th>
<th>Succeeded</th>
<th>Failed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62 (50)</td>
<td>38 (50)</td>
</tr>
</tbody>
</table>

The number of people falling in a specified category is listed as the first line in each cell (62 succeeded, 38 failed). The second line in each cell (in parentheses) contains the number expected to succeed if the null hypothesis is true. Since the null hypothesis is that the proportion that succeed is 0.5, $(0.5)(100) = 50$ are expected to succeed and $(0.5)(100) = 50$ are expected to fail.

The formula for chi square is:

$$\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

where $\chi^2$ is the symbol for the chi square, $E$ is an expected cell frequency, and $O$ is an observed cell frequency. The $\Sigma$ symbol is summation notation and means to sum up the quantity over both cells. Therefore, the formula says that for each cell you

1. take the absolute value of the difference between the expected cell frequency and the observed cell frequency
2. subtract 0.5 (the correction for continuity)
3. square the result,
4. divide by the expected frequency,
5. and finally, sum up the values across the cells.

For the present data,

$$\chi^2 = \frac{(|50 - 62| - 0.5)^2}{50} + \frac{(|50 - 38| - 0.5)^2}{50} = \frac{11.5^2}{50} + \frac{11.5^2}{50} = 5.29.$$  

The degrees of freedom for chi square is equal to the number of categories minus one. For this section in which there are always just two categories (success and failure for the present example), the degrees of
freedom is always one. A chi square table can be used to find that the two-tailed probability value for a chi square of 5.29 with one degree of freedom is 0.0214.

At the beginning of this section it was stated that the chi square test for proportions was equivalent to the one based on the normal distribution. It turns out that chi square will always equal z². For the present example, the value of z was 2.3 and the value of chi square was 5.29. Note that $2.3^2 = 5.29$. The probability values for a z of 2.3 and a chi square of 5.29 are identical ($p = 0.0214$).

**Reporting Results**

These results could be reported as follows:

The proportion of subjects choosing the original formulation (0.62) was significantly greater than 0.50, $\chi^2(1, N = 100) = 5.29$, $p = 0.021$. Apparently at least some people are able to distinguish between the original formulation and the new formulation.

The symbol "$(1, N = 100)$" means that the chi square is based on 1 df and that the total number of subjects is 100.

**Assumptions**

1. Observations are sampled randomly and independently.

2. For the chi square test to be accurate, $\pi$ cannot be too close to 0 or to 1 and N cannot be too small. A conservative rule of thumb is that $N\pi$ as well as $N(1-\pi)$ should both be at least 10.

**More than Two Categories**

A skeptical view of psychotherapy can be summed up as follows: one third better, one third worse, one third the same. In other words, one third of the patients improve as a result of psychotherapy, one third get worse, and one third stay the same.

Consider a hypothetical study of 100 patients who underwent psychotherapy. Assume that 45 were classified as having improved, 30 as having gotten worse, and 25 as having stayed the same. Is this outcome significantly different from that expected under the one third better, one third worse, one third the same?

The test is conducted by constructing a table similar to the one used to test the difference between $p$ and $\pi$.

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better</td>
<td>45 (33.333)</td>
</tr>
<tr>
<td>Worse</td>
<td>25 (33.333)</td>
</tr>
<tr>
<td>Same</td>
<td>30 (33.333)</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>
The first column contains the three categories into which patients could be classified. The top line in each cell in the second column contains the number of patients in the category; the second line contains the number expected under the null hypothesis. These "expected" frequencies are computed by multiplying the total number of subjects (100) by the proportion expected to be in the category (0.33333).

### More than Two Categories

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better</td>
<td>45 (33.333)</td>
</tr>
<tr>
<td>Worse</td>
<td>25 (33.333)</td>
</tr>
<tr>
<td>Same</td>
<td>30 (33.333)</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

The chi square statistic is then computed using the formula:

$$\chi^2 = \sum \frac{(E - O)^2}{E}$$

The degrees of freedom are equal to the the number of categories -1. For this example,

$$\chi^2 = \frac{(45 - 33.333)^2}{33.333} + \frac{(25 - 33.333)^2}{33.333} + \frac{(30 - 33.333)^2}{33.333} = 6.5$$

Since there are three categories, the degrees of freedom equal 3 - 1 = 2. From a chi square table, it can be determined that the probability value is 0.039. Therefore, the null hypothesis of one third better, one third worse, one third the same can be rejected.

### Reporting Results

The results of this experiment could be described as follows:

The proportions of subjects falling into the three outcome categories (better, worse, the same) were 0.45, 0.25, 0.30 respectively. These proportions differed significantly from the expected 0.333, 0.333, 0.333 proportions, $\chi^2(2, N = 100) = 6.50$, $p = 0.039$.

As shown, the Chi Square statistic is reported with its degrees of freedom and sample size.

### Summary of Computations

1. Make a table of expected and observed frequencies.
2. Compute Chi Square using the formula:

$$\chi^2 = \sum \frac{(E - O)^2}{E}$$
3. Degrees of freedom = number of categories minus one.

The correction for continuity is not used when there are more than two categories. Some authors claim it should be used whenever an expected cell frequency is below 5. Research in statistics has shown that this practice is not advisable.

Assumptions

1. Observations are sampled randomly and independently.

2. The formula for chi square yields a statistic that is only approximately distributed as Chi Square. For the Chi Square approximation to be sufficiently accurate, the total number of subjects should be at least 20.

Introduction to the Chi Square Test of Independence

Contingency tables are used to examine the relationship between subjects' scores on two qualitative or categorical variables. For example, consider the hypothetical experiment on the effectiveness of early childhood intervention programs described in another section. In the experimental group, 73 of 85 students graduated from high school. In the control group, only 43 of 82 students graduated. These data are depicted in the contingency table shown below.

<table>
<thead>
<tr>
<th></th>
<th>Graduated</th>
<th>Failed to Graduate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>73</td>
<td>12</td>
<td>85</td>
</tr>
<tr>
<td>Control</td>
<td>43</td>
<td>39</td>
<td>82</td>
</tr>
<tr>
<td>Total</td>
<td>116</td>
<td>51</td>
<td>167</td>
</tr>
</tbody>
</table>

The cell entries are cell frequencies. The top left cell with a "73" in it means that 73 subjects in the experimental condition went on to graduate from high school; 12 subjects in the experimental condition did not. The table shows that subjects in the experimental condition were more likely to graduate than were subjects in the control condition. Thus, the column a subject is in (graduated or failed to graduate) is contingent upon (depends on) the row the subject is in (experimental or control condition).

If the columns are not contingent on the rows, then the rows and column frequencies are independent. The test of whether the columns are contingent on the rows is called the chi square test of independence. The null hypothesis is that there is no relationship between row and column frequencies.

Computing the Chi Square Test of Independence

The first step in computing the chi square test of independence is to compute the expected frequency for each cell under the assumption that the null hypothesis is true. To calculate the expected frequency of the
first cell in the example (experimental condition, graduated), first calculate the proportion of subjects that graduated without considering the condition they were in. The table below shows that of the 167 subjects in the experiment, 116 graduated.

<table>
<thead>
<tr>
<th></th>
<th>Graduated</th>
<th>Failed to Graduate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental</strong></td>
<td>73</td>
<td>12</td>
<td>85</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>43</td>
<td>39</td>
<td>82</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>116</td>
<td>51</td>
<td>167</td>
</tr>
</tbody>
</table>

Therefore, 116/167 graduated. If the null hypothesis were true, the expected frequency for the first cell would equal the product of the number of people in the experimental condition (85) and the proportion of people graduating (116/167). This is equal to (85)(116)/167 = 59.042. Therefore, the expected frequency for this cell is 59.042. The general formula for expected cell frequencies is:

\[
E_{ij} = \frac{T_i \times T_j}{N}
\]

where \( E_{ij} \) is the expected frequency for the cell in the ith row and the jth column, \( T_i \) is the total number of subjects in the ith row, \( T_j \) is the total number of subjects in the jth column, and \( N \) is the total number of subjects in the whole table.

The calculations are shown below.

\[
E_{11} = \frac{85 \times 116}{167} \quad E_{12} = \frac{85 \times 51}{167} \\
E_{21} = \frac{82 \times 116}{167} \quad E_{22} = \frac{82 \times 51}{167}
\]

Once the expected cell frequencies are computed, it is convenient to enter them into the original table as shown below. The expected frequencies are in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Graduated</th>
<th>Failed to Graduate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental</strong></td>
<td>73 (59.042)</td>
<td>12 (25.958)</td>
<td>85</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>43 (56.958)</td>
<td>39 (25.042)</td>
<td>82</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>116</td>
<td>51</td>
<td>167</td>
</tr>
</tbody>
</table>

The formula for chi square test for independence is

\[
\chi^2 = \sum \frac{(E - O)^2}{E}
\]
For this example,

\[ E_{11} = \frac{85 \times 116}{167}, \quad E_{12} = \frac{85 \times 51}{167}, \]

\[ E_{21} = \frac{82 \times 116}{167}, \quad E_{22} = \frac{82 \times 51}{167}. \]

\[ \chi^2 = 22.01. \]

The degrees of freedom are equal to \((R-1)(C-1)\) where \(R\) is the number of rows and \(C\) is the number of columns. In this example, \(R = 2\) and \(C = 2\), so \(df = (2-1)(2-1) = 1\). A chi square table can be used to determine that for \(df = 1\), a chi square of 22.01 has a probability value less than 0.0001.

In a table with two rows and two columns, the chi square test of independence is equivalent to a test of the difference between two sample proportions. In this example, the question is whether the proportion graduating from high school differs as a function of condition. Whenever the degrees of freedom equal one (as they do when \(R = 2\) and \(C = 2\)), chi square is equal to \(z^2\). Note that the test of the difference between proportions for these data results in a \(z\) of 4.69 which, when squared, equals 22.01.

The same procedures are used for analyses with more than two rows and/or more than two columns. For example, consider the following hypothetical experiment: A drug that decreases anxiety is given to one group of subjects before they attempted to play a game of chess against a computer. The control group was given a placebo. The contingency table is shown below.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Win</th>
<th>Lose</th>
<th>Draw</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>12 (14.29)</td>
<td>18 (14.29)</td>
<td>10 (11.43)</td>
<td>40</td>
</tr>
<tr>
<td>Placebo</td>
<td>13 (10.71)</td>
<td>7 (10.71)</td>
<td>10 (8.57)</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>20</td>
<td>70</td>
</tr>
</tbody>
</table>

The expected frequencies are shown in parentheses. As in the previous example, each expected frequency is computed by multiplying the row total by the column total and dividing by the total number of subjects. For example, the expected frequency for the "Drug-Lose" condition is the product of the row total (40) and the column total (25) divided by the total number of subjects (70): \((40)(25)/70 = 14.29\).

The chi square is calculated using the formula:

\[ \chi^2 = \sum \frac{(E - O)^2}{E} \]

\[ \chi^2 = \frac{(14.29 - 12)^2}{14.29} + \frac{(14.29 - 18)^2}{14.29} + \ldots + \frac{(8.57 - 10)^2}{8.57} = 3.52. \]
The df are \((R-1)(C-1) = (2-1)(3-1) = 2\). A chi square table shows that the probability of a chi square of 3.52 with 2 degrees of freedom is 0.172. Therefore, the effect of the drug is not significant.

Summary of Computations

1. Create a table of cell frequencies.
2. Compute row and column totals.
3. Compute expected cell frequencies using the formula:
   \[
   E_{ij} = \frac{T_i \times T_j}{N}
   \]
   where \(E_{ij}\) is the expected frequency for the cell in the \(i\)th row and the \(j\)th column, \(T_i\) is the total number of subjects in the \(i\)th row, \(T_j\) is the total number of subjects in the \(j\)th column, and \(N\) is the total number of subjects in the whole table.
4. Compute Chi Square using the formula:
   \[
   \chi^2 = \sum \frac{(E - O)^2}{E}
   \]
5. Compute the degrees of freedom using the formula: \(df = (R-1)(C-1)\) where \(R\) is the number of rows and \(C\) is the number of columns.
6. Use a chi square table to look up the probability value.

Note that the correction for continuity is not used in the chi square test of independence.

Assumptions of the Chi Square Test of Independence

A key assumption of the chi square test of independence is that each subject contributes data to only one cell. Therefore the sum of all cell frequencies in the table must be the same as the number of subjects in the experiment.

Consider an experiment in which each of 12 subjects threw a dart at a target once using his or her preferred hand and once using his or her non-preferred hand. The data are shown below:

<table>
<thead>
<tr>
<th></th>
<th>Hit</th>
<th>Missed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preferred hand</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Non-preferred hand</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

It would not be valid to use the chi square test of independence on these data since each subject contributed data to two cells: one cell based on their performance with their preferred hand and one cell based on their performance with their non-preferred hand. The total of the cell frequencies in the table is 24 but the total number of subjects is only 12.
The formula for chi square yields a statistic that is only approximately a chi square distribution. In order for the approximation to be adequate, the total number of subjects should be at least 20.

Some authors claim that the correction for continuity should be used whenever an expected cell frequency is below 5. Research in statistics has shown that this practice is not advisable. For example, see: Bradley, Bradley, McGrath, & Cutcomb

**Reporting Results from the Chi Square Test of Independence**

The results for the example on the effects of the early childhood intervention example could be reported as follows:

The proportion of students from the early-intervention group who graduated from high school was 0.86 whereas the proportion from the control group who graduated was only 0.52. The difference in proportions is significant, $\chi^2(1, N = 167) = 22.01, p < 0.001$.

The experiment on the effect of the anti-anxiety drug on chess playing could be reported as:

The number of subjects winning, losing, and drawing as a function of drug condition is shown in Figure 1. Although subjects receiving the drug performed slightly worse than subjects not receiving the drug, the difference was not significant, $\chi^2(2, N = 70) = 3.52, p = 0.17$

For apply the exercises at following link: 
http://davidmlane.com/hyperstat/chi_square.html
Chapter 17

Distribution Free-Tests
Overview of Distribution-Free Tests

Most inferential statistics assume normal distributions. Although these statistical tests work well even if the assumption of normality is violated, extreme deviations from normality can distort the results. Usually the effect of violating the assumption of normality is to decrease the Type I error rate, although this may sound like a good thing, it often is accompanied by a substantial decrease in power. Moreover, in some situations the Type I error rate is increased.

There is a collection of tests called distribution-free tests that do not make any assumptions about the distribution from which the numbers were sampled. Thus the name, "distribution-free." The main advantage of distribution-free tests is that they provide more power than traditional tests when the samples are from highly-skewed distributions. Since alternative means of dealing with skewed distributions such as taking logarithms or square roots of the data are available, distribution-free tests have not achieved a high level of popularity. Distribution-free tests are nonetheless a worthwhile alternative. Moreover, computer analysis has made possible new and promising variations of these tests. Distribution-free tests are sometimes referred to as nonparametric tests because, strictly speaking, they do not test hypotheses about population parameters. Nonetheless, they do provide a basis for making inferences about populations and are therefore classified as inferential statistics.

Some researchers take the position that measurement on at least an interval scale is necessary for techniques such as the analysis of variance that result in inferences about means. These researchers use distribution-free tests as alternatives to the usual parametric tests. As discussed in another section, most researchers do not accept the argument that tests such as the analysis of variance can only be done with interval-level data.

Randomization Tests

Most distribution-free tests are based on the principle of randomization. The best way to understand the principle of randomization is in terms of a specific example of a randomization test. Assume that four numbers are sampled from each of two populations. The numbers are shown below.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Mean</td>
<td>10</td>
</tr>
</tbody>
</table>

The first step is to compute the difference between means. For these data, the difference is six. The second step is to compute the number of ways these eight numbers could be divided into two groups of four. The general formula is:

$$W = \frac{N!}{n_1! n_2!}$$

where W is the number of ways, N is the total number of numbers (8 in this case), n_1 is the size of the first group (4 in this case) and n_2 is the size of the second group (4 in this case). Therefore, $W = 8!/(4! 4!) = 70$. 
This means that there are 70 ways in which eight numbers can be divided into two groups of four.

The third step is to determine how many of these W ways of dividing the data result in differences between the means as large or larger than the difference obtained in the actual data. An examination of the data shows that there are only two ways the eight numbers can be divided into two groups of four with a larger difference between means than the difference of six found in the actual arrangement. These two ways are shown below.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Mean</td>
<td>10.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Mean</td>
<td>10.25</td>
</tr>
</tbody>
</table>

Thus, including the original data, there are three ways in which the eight numbers can be arranged so that the difference between means is six or more.

To compute the probability value for a one-tailed test of the difference between groups, divide this value of three by the W of ways of dividing the data into two groups of four. The probability value is therefore: $p = \frac{3}{70} = 0.0429$.

For a two-tailed test, the three cases in which Group 2 had a mean that was greater than Group 1 by six or more would be considered. This would make the two-tailed probability: $p = \frac{6}{70} = 0.0857$.

In summary, a randomization test proceeds from the data actually collected. It compares a computed statistic (the difference between means in this example) with the value of that statistic for other arrangements of the data. The probability value is simply the proportion of arrangements leading to a value of the statistic as large or larger than the value obtained from the actual data.

Consider one more example of a randomization test. Suppose a researcher wished to know whether or not there were a relationship between two variables: X and Y. The most common way to test the relationship would be using Pearson’s r.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

The test based on the principle of randomization would proceed as follows. First, Pearson’s correlation would be computed for the data as they stand. The value is: $r = 0.9556$. Next, the number of ways the X and Y numbers could be paired is calculated (Note that X's do not
become Y’s and Y’s do not become X’s. It is the pairings that change.) The formula for the number of ways that the numbers can be paired is simply:
\[ W = N! \]

where \( N \) is the number of pairs of numbers. For this example, \( N = 5 \) and \( W = 120 \). This means there are 120 ways the numbers can be paired. Of these pairings, only one would produce a higher correlation than 0.9556. It is shown on the next page. Therefore, there are two ways of arranging the data that result in correlations of 0.9556 or higher.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>.981</td>
</tr>
</tbody>
</table>

This makes the one-tailed probability equal to: \( p = \frac{2}{120} = 0.017 \). Naturally, the two-tailed probability is 0.034.

Randomization tests have only one major drawback: they are impractical to compute with moderate to large sample sizes. For example, the number of ways 45 scores can be equally divided among three groups is which is an astronomical number. Even high-speed computers are incapable of the necessary calculations.

\[ W = \frac{45!}{15! \times 15! \times 15!} = 5.34 \times 10^{19} \]

Fortunately, there is an alternative computational method called "resampling." Instead of calculating all possible ways the data could be arranged, one can take numerous random samples from the set of possible arrangements. For the problem of dividing 45 numbers into three groups of 15 each, one could randomly select 10,000 of the possible arrangements and determine the proportion of these for which the effect size in the sample is exceeded. This proportion is then an accurate estimate of the probability level. This procedure is currently not in frequent use due, in part, to the general unavailability of programs to do this type of computation.

**Rank Randomization Tests**

The major problem with randomization tests is that they are very difficult to compute. Rank randomization tests are performed by first converting the scores to ranks and then computing a randomization test. The primary advantage of rank randomization tests is that there are tables that can be used to determine significance. The disadvantage is that some information is lost when the numbers are converted to ranks. Therefore, rank randomization tests are generally less powerful than randomization tests based on the original numbers. Consider the following data:

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>
that were used as an example in the section on randomization tests. A rank randomization test on these data begins by converting the numbers to ranks. The converted data are:

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

The sum of the ranks of the first group is 24 and the sum of the ranks of the second group is 12. The difference is therefore 12. The problem is to figure out the number of ways that data can be arranged so that the difference between summed ranks is 12 or more. The bottom of this page shows the four arrangements of the eight ranks that lead to a difference of 12 or more. Since there are \( W = \frac{8!}{(4!\ 4!)} = 70 \) ways of arranging the data, the one-tailed probability is: \( \frac{4}{70} = 0.057 \).

8 4  
7 3  
6 2  
5 1  
8 5  
7 3  
6 2  
4 1  
8 6  
7 3  
5 2  
4 1  
8 5  
7 4  
6 2  
3 1  

The two-tailed probability is therefore: \( \frac{8}{70} = 0.114 \).

Notice that this probability value is slightly higher than the 0.0857 obtained with the randomization test for these same data. This rank randomization test for differences between two groups has several names. It is most often called the Mann-Whitney U test or the Wilcoxon Rank Sum test. In practice, tables are available to look up the probability value based on the sum of the ranks of the group with the lower mean rank. For these data, the sum of the ranks is: \( 2 + 6 + 1 + 3 = 12 \). A table would show that for a one-tailed test to be significant at the 0.05 level, the sum of the ranks must be \( \leq 11 \). Since 12>11, the test is not significant at this level. This agrees with the calculated probability value of 0.057. The purpose of this section is to present the concepts rather than the details of the computations. Therefore, if you wish to actually perform this test, you should find a textbook with appropriate tables. An online table and further discussion can be found here.

Rank randomization tests are also available for correlation and for testing differences between more than two groups (analogous to one-way ANOVA). The rank randomization
test for correlation is called Spearman's rho. The rank randomization test for differences among groups is called the Kruskal and Wallis Test. In summary, rank randomization tests are based on the principle of randomization. They are much easier to compute than randomization tests. However, they are less powerful than randomization tests and the more typically performed parametric tests such as analysis of variance. For this reason, they are not as widely used.

**Sign Test**

Suppose a researcher were interested in whether a drug helped people fall asleep. Subjects took the drug one night and a placebo another night with the order counterbalanced. The number of minutes it took each of 8 subjects to fall asleep is shown below.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>21</td>
<td>36</td>
</tr>
<tr>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>18</td>
<td>29</td>
</tr>
<tr>
<td>11</td>
<td>22</td>
</tr>
</tbody>
</table>

If the drug has no effect, then the population proportion (π) of people falling asleep faster in the drug condition is 0.50. Therefore, the null hypothesis is that π = 0.50. Of the 8 subjects, 7 fell asleep faster with the drug.

Using the formula for the binomial distribution, the probability of obtaining 7 or more "successes" with N = 8 and π = 0.50 can be calculated to be 0.0352 as follows:

\[
P(7) = \frac{8!}{7!(8 - 7)!} \cdot 0.5^7 (1 - 0.5)^{8-7} = (8)(0.00391) = 0.03125
\]

\[
P(8) = \frac{8!}{8!(8 - 8)!} \cdot 0.5^8 (1 - 0.5)^{8-8} = 0.00391
\]

Therefore, \( P(7) + P(8) = 0.03125 + 0.00391 = 0.0352 \). The one-tailed probability value is therefore 0.03516; The two-tailed probability value is \( 2)(0.03516) = 0.0732 \).

Technically speaking, the sign test is a test of whether the median is a particular value. In this example, the test is whether the median difference between conditions is zero. The sign test is generally less powerful than a t test of the difference between dependent means. For the example data, the t test would be: \( t(7) = -3.21 \). The two-tailed probability value for this t is 0.015.

For apply the exercises at following link:
http://davidmlane.com/hyperstat/dist_free.html
Reference

The Basic Reference


Helpful Sites

English Sites

- www.statistics.com
- www.uwsp.edu/psych/stat/index.htm
- www.stattrek.com
- www.statisticssolutions.com

Arabic Sites

- www.arabicstat.com/board/index.php
- www.minshawi.com/vb/

Helpful Books

English Books

- Elementary Statistics.
- Using SPSS for windows.

Arabic Books

- مقدمة في الطرق الإحصائية / د.جلال الصياد و د.محمد حبيب
- مبادئ الإحصاء / د.جلال الصياد و أ.عامل سمرة