

**Study Guide for
Essentials of Statistics for the Social and Behavioral Sciences
by Barry H. Cohen and R. Brooke Lea**

Chapter 1

Distributions

Guidelines for Frequency Distributions

The procedure for constructing a grouped frequency distribution can be summarized in terms of the following steps and guidelines.

Step 1. Choose the width (i) of the class intervals. First, find the range of scores by subtracting the lowest score in your distribution from the highest, and adding one (i.e., range = highest score - lowest score + 1). Then divide the range by a convenient class width (a multiple of 5, or a number less than 5, if appropriate), and round up if there is any fraction, to find the number of intervals that would result. If the number of intervals is between 10 and 20, you have found an appropriate width; otherwise, try another convenient value for i . On the other hand, an external criterion might dictate the size of the class intervals (e.g., assigning letter grades based on predetermined scores).

Step 2. Choose the apparent limits of the lowest interval. The lowest interval must contain the lowest score in the distribution. In addition, it is strongly suggested that the lower limit or the upper limit be a multiple of the chosen interval width.

Step 3. All the class intervals should be the same size. Work upwards from the lowest interval to find the remaining intervals (stop when you reach an interval that contains the highest score). Taking into consideration the real limits of the intervals, make sure that no two intervals overlap, and that there are no gaps between intervals (the real limits are half a unit of measurement above or below the apparent limits).

Step 4. Count the number of scores that fall between the real limits of each interval. The frequency count for each interval is written next to the limits of that interval.

If the number of possible scores between the highest and lowest scores in the distribution is less than 20, a simple (i.e., ungrouped) frequency distribution may suffice. Furthermore, in some cases it may be desirable to have somewhat less than 10 or somewhat more than 20 intervals. Judgment must be used to determine the intervals that result in the most informative description of the data. Once a grouped frequency distribution has been constructed, it is easy to derive the following related distributions:

1. Relative frequency distribution. Divide each entry of the grouped frequency distribution by N (the total number of scores in the distribution = $\sum f$). The sum of the entries should equal 1.0.
2. Cumulative frequency distribution. The cf entry for the lowest interval is just the frequency of that interval. The cumulative frequency for any higher interval is the frequency of that interval plus the cumulative frequency for the next lowest interval. The highest entry should equal N .
3. Cumulative relative frequency distribution. Divide each entry of the cumulative frequency distribution by N . The highest entry should equal 1.0.
4. Cumulative percent frequency distribution. Multiply each entry in the cumulative relative frequency distribution by 100. The highest entry should equal 100%.

We will apply the above procedures to an example. Suppose you are studying eating disorders in women, and want, as a basis for comparison, to know the weights for a random group of college-aged women of average height. The weights, in pounds, for 50 hypothetical women appear in an array (Table 1.1) below.

Table 1.1

171	166	157	153	151	149	147	145	143	142
140	137	137	135	135	134	134	132	131	131
128	128	127	126	126	126	124	124	123	122
122	121	120	120	119	119	118	117	117	116
115	114	114	113	112	110	108	105	102	97

Step 1. The range of scores equals $171 - 97 + 1 = 74 + 1 = 75$. For $i = 10$, the number of intervals would be $75/10 = 7.5$, which rounded up equals only 8. For $i = 5$, $75/5 = 15$. Because 15 is between 10 and 20 we will use $i = 5$.

Step 2. The lowest interval could be 95 - 99, as 95 is a multiple of i , and the width of the interval is (subtracting the real limits) $99.5 - 94.5 = 5$ (96 - 100 could also be the lowest interval, as 100 is a multiple of i , but it seems easier to have the lower limit divisible by i).

Step 3. The next interval above 95 - 99 is 100 - 104, and so on until the highest interval, which is 170 - 174, because this interval contains the highest score (174).

Step 4. Next to each interval, tally marks can be made to indicate how many of the scores in Table 1.1 fall in that interval. The sum of the tally marks is the frequency for that interval, as shown in Table 1.2.

Table 1.2 shows the grouped frequency distribution for the 50 weights, along with relative (rf), cumulative (cf), cumulative relative (crf), percent frequency (pf), and cumulative percent frequency (cpf) distributions.

Table 1.2

Interval	f	rf	cf	crf	pf	cpf
170 - 174	1	.02	50	1.00	2	100%
165 - 169	1	.02	49	.98	2	98
160 - 164	0	0	48	.96	0	96
155 - 159	1	.02	48	.96	2	96
150 - 154	2	.04	47	.94	4	94
145 - 149	3	.06	45	.90	6	90
140 - 144	3	.06	42	.84	6	84
135 - 139	4	.08	39	.78	8	78
130 - 134	5	.10	35	.70	10	70
125 - 129	6	.12	30	.60	12	60
120 - 124	8	.16	24	.48	16	48
115 - 119	7	.14	16	.32	14	32
110 - 114	5	.10	9	.18	10	18
105 - 109	2	.04	4	.08	4	8
100 - 104	1	.02	2	.04	2	4
95 - 99	1	.02	1	.02	2	2

Guidelines for Graphs

1. The Y-axis should be only about two-thirds as long as the X-axis.
2. For frequency distributions, the variable of interest is placed along the X-axis, while the frequency counts (or relative frequency) are represented along the Y-axis.
3. The intersection of the X- and Y-axes is the zero point for both dimensions.
4. The measurement units are equally spaced along the entire length of both axes.
5. Choose a scale to represent the measurement units on the graph (e.g., one pound equals one-tenth of an inch) so that the histogram or polygon fills the space of the graph as much as possible. Indicating a break in the scale on one or both axes may be necessary to achieve this goal.
6. Both axes should be clearly labeled, and each label, including the name of the variable and unit of measurement, should be placed parallel to the axis.

Central Tendency and Variability

Advantages and Disadvantages of the Major Measures of Central Tendency

Advantages of the Mode:

1. Easy to find.
2. Can be used with any scale of measurement.
3. The only measure that can be used with a nominal scale.
4. Corresponds to actual score in the distribution.

Disadvantages of the Mode (the following apply when the mode is used with ordinal or interval/ratio data):

1. Generally unreliable, especially when representing a relatively small population (can change radically with only a minor change in the distribution).
2. Can be misleading; the mode tells you which score is most frequent, but tells you nothing about the other scores in the distribution (radical changes can be made to the distribution without changing the mode).
3. Cannot be easily used in conjunction with inferential statistics.

Advantages of the Median:

1. Can be used with either ordinal or interval/ratio data.
2. Can be used even if there are open-ended categories or undeterminable scores on either side of the distribution.
3. Provides a good representation of a typical score in a skewed distribution; is not unduly affected by extreme scores.
4. Minimizes the sum of the absolute deviations (i.e., the sum of score distances from the median -- ignoring sign -- is less than it would be from any other location in the distribution).

Disadvantages of the Median:

1. May not be an actual score in the distribution (e.g., if there are an even number of scores, or tied scores in the middle of the distribution).
2. Does not reflect the values of all the scores in the distribution (e.g., an extreme score can be moved even further out without affecting the median).
3. Compared to the mean, it is less reliable for drawing inferences about a population from a sample, and harder to use with advanced statistics.

Advantages of the Mean:

1. Reflects the values of all the scores in the distribution.
2. Has many desirable statistical properties
3. Is the most reliable for drawing inferences, and the easiest to use in advanced statistical techniques.

Disadvantages of the Mean:

1. Usually not an actual score in the distribution.
2. Not appropriate for use with ordinal data.
3. Can be misleading when used to describe a skewed distribution.
4. Can be strongly affected by even just one very extreme score (i.e., an outlier).

Advantages and Disadvantages of the Major Measures of Variability

Advantages of the Range:

1. Easy to calculate.
2. Can be used with ordinal as well as interval/ratio data.
3. Encompasses entire distribution.

Disadvantages of the Range:

1. Depends on only two scores in the distribution and is therefore not reliable.
2. Cannot be found if there are undeterminable or open-ended scores at either end of the distribution.
3. Plays no role in advanced statistics.

Advantages of the SIQ Range:

1. Can be used with ordinal as well as interval/ratio data.
2. Can be found even if there are undeterminable or open-ended scores at either end of the distribution.
3. Not affected by extreme scores or outliers.

Disadvantages of the SIQ Range:

1. Does not take into account all the scores in the distribution.
2. Does not play a role in advanced statistical procedures.

Advantages of the Mean Deviation

1. Easy to understand (it is just the average distance from the mean).
2. Provides a good description of variability.
3. Takes into account all scores in the distribution.
4. Less sensitive to extreme scores than the standard deviation.

Disadvantages of the Mean Deviation

1. This measure is smaller when taken around the median than the mean.
2. Is not easily used in advanced statistics.
3. Cannot be calculated with undeterminable or open-ended scores.

Advantages of the Standard Deviation

1. Takes into account all scores in the distribution.
2. Provides a good description of variability.
3. Tends to be the most reliable measure.
4. Plays an important role in advanced statistical procedures.

Disadvantages of the Standard Deviation

1. Very sensitive to extreme scores or outliers.
2. Cannot be calculated with undeterminable or open-ended scores.

In order to illustrate the calculation of the various measures of variability, we will present hypothetical data for the following situation. You are a ninth-grade English teacher, and during the first class of the Fall term, you ask each of your 12 pupils how many books he or she has read over the summer vacation, to give you some idea of their interest in reading. The responses were as follows:

3, 1, 3, 3, 6, 2, 1, 7, 3, 4, 9, 2.

Putting the scores in order will help in finding some of the measures of variability:

1, 1, 2, 2, 3, 3, 3, 3, 4, 6, 7, 9.

The Range. The range is the highest score minus the lowest. The number of books read ranged from 1 to 9, so range = $9 - 1 = 8$. If the scale is considered continuous (e.g., 9 books is really anywhere between $8 \frac{1}{2}$ and $9 \frac{1}{2}$ books), then range = upper real limit of highest score minus lower real limit of the lowest score = $9.5 - 0.5 = 9$.

Interquartile Range. With such a small set of scores, grouping does not seem necessary to find the quartiles. Because $N = 12$, the 25th percentile is between the third and fourth scores. In this case, the third and fourth scores are both 2, so $Q_1 = 2$. Similarly, the 75th percentile is between the ninth and tenth score, which are 4 and 6, so $Q_3 = 5$. The IQ range = $Q_3 - Q_1 = 5 - 2 = 3$.

Semi-interquartile Range. The SIQ range = $(Q_3 - Q_1)/2$, which is half of the IQ range. For this example, SIQ range = $3/2 = 1.5$.

Mean Deviation. This is the average of the absolute deviations from the mean. The mean for this example is : = $EX/N = 3.67$. The mean deviation is found by applying Formula 1.2 to the data:

$$M. D. = \frac{\sum |X_i - \mu|}{N} = \frac{1}{N} [|-2.67| + |-2.67| + |-1.67| + |-1.67| + |-.67| + |-.67| + |-.67| + |-.67| + |.33| + |2.33| + |3.33| + |5.33|] = \frac{1}{12} [23.66] = 1.97$$

Sum of Squares. This is the sum of the squared deviations from the mean, as found by the numerator of Formula 1.3:

$$\begin{aligned} SS = \sum (X_i - \mu)^2 &= (-2.67)^2 + (-2.67)^2 + (-1.67)^2 + \dots + (2.33)^2 + (3.33)^2 + (5.33)^2 \\ &= 7.13 + 7.13 + 2.79 + 2.79 + .45 + .45 + .45 + .45 + .11 + 5.44 + 11.11 + 28.44 = 66.75 \end{aligned}$$

Variance. This is the average of the squared deviations from the mean, as found by Formula 1.3:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} = \frac{SS}{N} = \frac{66.75}{12} = 5.56.$$

Standard Deviation. This equals the square-root of the variance as given by Formula 1.4. For this example, $F = \sqrt{5.56} = 2.36$.

Unbiased Sample Variance (from Chapter 2). If the 12 students in the English class were considered a sample of all ninth graders, and you wished to extrapolate from the sample to the population, you would use Formula 2.4 to calculate the unbiased variance s^2 :

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{N - 1} = \frac{66.75}{11} = 6.068$$

Unbiased Standard Deviation From Chapter 2. This equals the square-root of the unbiased variance as given by Formula 2.5. For this example, $s = \sqrt{6.068} = 2.46$.

Standardized Scores and the Normal Distribution

If a variable is normally distributed and we know both the mean and standard deviation of the population, it is easy to find the proportion of the distribution that falls above or below any raw score, or between any two raw scores. Conversely, for a given proportion at the top, bottom, or middle of the distribution, we can find the raw score or scores that form the boundary of that proportion. We will illustrate these operations using the height of adult females as our variable, assuming : = 65", and $F = 3$ ".

Using z-scores with the Normal distribution

I. The proportion (and PR) below a given raw-score: (We will not describe a separate procedure for finding proportions above a given raw score, because once a drawing has been made, the procedure should be clear. In any case, one can always find the proportion above a score by first finding the proportion below, and then subtracting from 1.0).

Example A: What proportion of adult women are less than 63" tall?

Step 1. Find the z-score (using Formula 1.6).

$$z = \frac{X - \mu}{\sigma} = \frac{63 - 65}{3} = \frac{-2}{3} = -.67$$

Step 2. Draw the picture (including vertical lines at the approximate z-score, and at the mean). When the z-score is negative, the area below (i.e., to the left of) the z-score is the area beyond z (i.e., towards the tail of the distribution).

Step 3. Look in Table A.1 We look up .67 (ignoring the minus sign) in the z-column, and then look at the entry under "area beyond." The entry is .2514, which is the proportion we are looking for.

Step 4. Multiply the proportion below the score by 100 if you want to find the PR for that score. In this example, PR = .2514 * 100 = 25.14%.

Example B: What proportion of adult women are less than 66" tall?

Step 1. Find the z-score.

$$z = \frac{66 - 65}{3} = \frac{1}{3} = +.33$$

Step 2. Draw the picture.

When the z-score is positive, the area below the z-score consists of two sections: the area between the mean and z, and the area below the mean (the latter always equals .5).

Step 3. Look in Table A.1 The area between the mean and z for z = .33 is .1293. Adding the area below the mean we get .1293 + .5 = .6293, which is the proportion we are looking for.

Step 4. Multiply by 100 to find the PR. In this case, the PR for 66" = .6293 * 100 = 62.93%.

II. The proportion between two raw scores.

Example A: What is the proportion of adult women between 64" and 67" in height?

Step 1. Find the z-scores.

$$z = \frac{64 - 65}{3} = \frac{-1}{3} = -.33 ; \quad z = \frac{67 - 65}{3} = \frac{2}{3} = +.67$$

Step 2. Draw the picture.

When the z-scores are opposite in sign (i.e., one is above the mean and the other is below), there are two areas that must be found: the area between the mean and z for each of the two z-scores.

Step 3. Look in Table A.1 The area between the mean and z is .1293 for z = .33, and .2486 for z = .67. Adding these we get .1293 + .2486 = .3779, which is the proportion we are looking for.

Example B: What proportion of adult women is between 67" and 68" tall?

Step 1. Find the z-scores.

$$z = \frac{67 - 65}{3} = \frac{2}{3} = +.67 ; \quad z = \frac{68 - 65}{3} = \frac{3}{3} = +1.0$$

Step 2. Draw the picture.

When both z-scores are on the same side of the mean, which requires us to find the area between the mean and z for each, and then subtract the two areas.

Step 3. Look in Table A.1 The area between the mean and z is .2486 for z = .67 and .3413 for z = 1.0. Subtracting we get: .3413 - .2486 = .0927, which is the proportion we are looking for. (Note: You cannot subtract the two z-scores first and then look for the area; this will give you the wrong area).

III. The raw score(s) corresponding to a given proportion. (Because we are starting with a raw score and looking for a proportion, the order of the steps is different.)

Example A: Above what height are the tallest 10% of adult women?

Step 1. Draw the picture. Shade in the top 10%; we want to know the z-score for which the area beyond is 10%, or .1000.

Step 2. Look in Table A.1 Looking in the "area beyond" column for .1000, the closest entry is .1003, which corresponds to z = +1.28. (If we were looking for the bottom 10%, z would equal -1.28.)

Step 3. Find the raw score (solving Formula 1.6 for X).

$$X = zF + : = +1.28(3) + 65 = 3.84 + 65 = \underline{68.84}$$

The tallest 10% of adult women are those that are 68.84" or taller.

Example B: Between which two heights are the middle 90% of adult women?

Step 1. Draw the picture. Subtract 90% from 100%, which equals 10%, to find the percentage in the tails. Then divide this percentage in half to find the percentage that should be shaded in each tail. Thus, we need the z-score for which the area beyond is 5% or .0500.

Step 2. Look in Table A.1 We cannot find .0500 in the "area beyond" column, but it is halfway between 1.64 (.0505) and 1.65 (.0495), so we will say that z = 1.645. Note that we want is both z = +1.645 and z = -1.645.

Step 3. Find the raw scores.

$$X = +1.645(3) + 65 = +4.94 + 65 = \underline{69.94}$$

$$X = -1.645(3) + 65 = -4.94 + 65 = \underline{60.06}$$

The middle 90% of adult women are between 60.06" and 69.94."

Definitions of Key Terms

Array - A list of scores in numerical order (usually from highest to lowest).

Simple Frequency Distribution - A list of all possible scores from the highest to the lowest in the group, together with a frequency count for each possible score.

Class Interval - A range of scores that represents a subset of all the possible scores in a group.

Apparent Limits - The lowest score value included in a class interval is called the lower apparent limit, and the highest score value is the upper apparent limit.

Real Limits - If the scores in a distribution have been measured for a continuous variable, the lower real limit is one-half of a measurement unit below the lower apparent limit, and the upper real limit is one-half of a measurement unit above the upper apparent limit.

Grouped frequency distribution - A list of (usually equal-sized) class intervals that do not overlap, and together include all of the scores in a distribution, coupled with the frequency count associated with each class interval.

Cumulative frequency distribution - A distribution in which the entry for each

score (or interval) is the total of the frequencies at or below that score (or interval).

Cumulative percentage frequency distribution - A cumulative distribution, in which each cumulative frequency is expressed as a percentage of the total N (i.e., the cumulative relative frequency is multiplied by 100).

Percentile Rank (PR) - The PR of a score is the percentage of the scores in the distribution that tie or fall below that score (i.e., the PR is the cumulative percentage associated with that score).

Percentile - A score that has a specified PR (e.g., the 25th percentile is the score whose PR is 25). Of most interest are deciles (e.g., 10%, 20%, etc.) or quartiles (25, 50, or 75%).

Bar graph - A graph in which vertical bars represent values on a discrete scale, and adjacent bars are kept separate. When representing a frequency distribution, the heights of the bars are proportional to the frequencies of the corresponding values.

Frequency histogram - A bar graph of a frequency distribution in which the variable of interest is considered to be continuous, and therefore adjacent bars touch (each bar extends horizontally from the lower to the upper real limit of the score or interval represented).

Frequency polygon - A graph of a frequency distribution for a continuous variable, in which a point is drawn above the midpoint of each interval (or score) at a height proportional to the corresponding frequency. These points are connected by straight lines, which are connected to the horizontal axis at both ends of the distribution.

Cumulative frequency polygon - Also called an ogive, this is a frequency polygon, in which the points represent cumulative frequencies, and are drawn above the upper real limit for each interval (or score).

Theoretical Distribution - A distribution based on a mathematical formula, which a particular frequency polygon may be expected to resemble.

Exploratory Data Analysis (EDA) - A set of techniques devised by J.W. Tukey (1977) for the purpose of inspecting one's data before attempting to summarize or draw inferences from it.

Stem-and-leaf Display - One of Tukey's EDA methods which resembles a frequency distribution, but without losing the identity of the raw scores. Leading digits form the vertical stem, while trailing digits form the horizontal rows (or "leaves"). The length of each row is proportional to the frequency corresponding to the leading digit, so that this display resembles a frequency histogram on its side.

Central Tendency: The location in a distribution that is most typical or best represents the entire distribution.

Arithmetic mean: This is the value obtained by summing all the scores in a group (or distribution), and then dividing by the number of scores summed. It is the most familiar measure of central tendency, and is therefore referred to simply as "the mean" or the "average."

Mode: The most frequent category, ordinal position, or score in a population or sample.

Unimodal: describing a distribution with only one major peak (e.g., the normal distribution is unimodal).

Bimodal: describing a distribution that has two roughly equal peaks (this shape usually indicates the presence of two distinct subgroups).

Median: This is the location in the distribution that is at the 50th percentile; half the scores are higher in value than the median while the other half are lower.

Skewed distribution: A distribution with the bulk of the scores closer to one end than the other, and relatively few scores in the other direction.

Positively skewed: describing a distribution with a relatively small number of scores much higher than the majority of scores, but no scores much lower than the majority.

Negatively skewed: opposite of positively-skewed (a few scores much lower than the majority, but none much higher).

Open-ended category: This is a measurement that has no limit on one end (e.g., 10 or more).

Undeterminable score: A score whose value has not been measured precisely but is usually known to be above or below some limit (e.g., subject has three minutes in which to answer a question, but does not answer at all).

Box-and-whisker plot: Also called boxplot, for short, this is a technique for exploratory data analysis devised by Tukey (1977). One can see the spread and symmetry of a distribution at a glance, and the position of any extreme scores.

Hinges: These are the sides of the box in a boxplot, corresponding approximately to the 25th and 75th percentiles of the distribution.

H-spread: The distance between the two hinges (i.e., the width of the box) in a boxplot.

Inner fences: Locations on either side of the box (in a boxplot) that are 1.5 times the H-spread from each hinge. The distance between the upper and lower inner fence is four times the H-spread.

Adjacent values: The upper adjacent value is the highest score in the distribution that is not higher than the upper inner fence, and the lower adjacent value is similarly defined in terms of the lower inner fence of a boxplot. The upper whisker is drawn from the upper hinge to the upper adjacent value, and the lower whisker is drawn from the lower hinge to the lower adjacent value.

Outlier: Defined in general as an extreme score standing by itself in a distribution, an outlier is more specifically defined in the context of a boxplot. In terms of a boxplot, an outlier is any score that is beyond the reach of the whiskers on either side. The outliers are indicated as points in the boxplot.

Range: The total width of the distribution as measured by subtracting the lowest score (or its lower real limit) from the highest score (or its upper real limit).

Interquartile (IQ) Range: The width of the middle half of the distribution as measured by subtracting the 25th percentile (Q1) from the 75th percentile (Q3).

Semi-interquartile (SIQ) Range: Half of the IQ range -- roughly half of the scores in the distribution are within this distance from the median.

Deviation Score: The difference between a score and a particular point in the distribution. When we refer to deviation scores, we will always mean the difference between a score and the mean (i.e., $X_i - \bar{X}$).

Absolute Value: The magnitude of a number, ignoring its sign; that is, negative signs are dropped, and plus signs are left as is. Absolute deviation scores are the absolute values of deviation scores.

Mean Deviation: The mean of the absolute deviations from the mean of a distribution.

Sum of Squares (SS): The sum of the squared deviations from the mean of a distribution.

Population Variance (F^2): Also called the mean-square (MS), this is the mean of the squared deviations from the mean of a distribution.

Population Standard Deviation (F): Sometimes referred to as the root-mean-square (RMS) of the deviation scores, it is the square-root of the population variance.

Definitional Formula: A formula that provides a clear definition of a sample statistic or population parameter, but may not be convenient for computational purposes. In the case of the variance or standard deviation, the definitional formula is also called the deviational formula, because it is based on finding deviation scores from the mean.

Computational Formula: An algebraic transformation of a definitional formula, which yields exactly the same value (except for any error due to intermediate rounding off), but reduces the amount or difficulty of the calculations involved.

Kurtosis: The degree to which a distribution bends sharply from the central peak to the tails, or slopes more gently, as compared to the normal distribution.

Leptokurtic: Referring to a distribution whose tails are relatively fatter than in the normal distribution, and therefore has a positive value for kurtosis (given that the kurtosis of the normal distribution equals zero).

Platykurtic: Referring to a distribution whose tails are relatively thinner than in the normal distribution, and therefore has a negative value for kurtosis.

Mesokurtic: Referring to a distribution in which the thickness of the tails is comparable to the normal distribution, and therefore has a value for kurtosis that is near zero.

Raw score. A number that comes directly from measuring a variable, without being transformed in any way.

Standardized score. A set of raw scores can be transformed into standardized scores by a simple formula, so that the mean and standard deviation are convenient numbers and the shape of the distribution is not changed.

z-score. A standardized score designed to have a mean of zero and a standard deviation of one. The magnitude of the z-score tells you how many standard deviations you are from the mean, and the sign of the z-score tells you whether you are above or below the mean.

Standard Normal Distribution. A normal distribution with a mean of zero and a standard deviation of one. Any normal distribution can be transformed into the standard normal distribution by converting all the scores to z-scores.

Area under the curve. If you consider a range of values within a distribution, constructing vertical lines at each end of the range (or interval), the vertical lines, the horizontal line, and the curve of the distribution will enclose an area that is generally less than the entire distribution. The ratio of that area to the entire distribution (defined as having an area of 1.0) tells you the proportion of scores that fall within the given range of values.

Chapter 2

Hypothesis testing can be divided into a six-step process. The six steps are as follows:

1. State the hypotheses.
2. Select the statistical test and the significance level.
3. Select the sample and collect the data.
4. Find the region of rejection.
5. Calculate the test statistic.
6. Make the statistical decision.

We pose the following question for our example: Do red-headed people have the same IQ as the rest of the population? Such a question can be answered definitively only by measuring the IQ of every red-headed individual in the population, calculating the mean, and comparing that mean to the general population mean. A more practical but less accurate way is to select a random sample of red-headed people and to find the mean of that sample. It is almost certain that the sample mean will differ from the mean of the general population, but that doesn't automatically tell us that the mean for the red-headed population differs from the general population. However, if the mean IQ of the red-headed sample is far enough from the mean of the general population, we can then conclude (at the risk of making a Type I error) that the two population means are different. To determine how far is far enough we need to follow the steps of a one-sample z-test.

Step 1. State the hypotheses.

The easiest way to proceed is to set up a specific null hypothesis that we wish to disprove. If the IQ of the general population is 100, then the null hypothesis would be expressed as follows: $H_0: \mu = 100$. The complementary hypothesis is called the alternative hypothesis, and is the hypothesis we would like to be true. A two-tailed alternative would be written: $H_A: \mu \neq 100$; a one-tailed alternative would be written as: $H_A: \mu < 100$, or $H_A: \mu > 100$. Usually a two-tailed test is performed unless there is strong justification for a one-tailed test.

Step 2. Select the statistical test and the significance level.

We are comparing one sample mean to a population mean, and the standard deviation of the population is known for the variable of interest, so it is appropriate to perform a one-sample z test. Alpha is usually set at .05, unless some special justification requires a larger or smaller alpha.

Step 3. Select the sample and collect the data.

A random sample of the "special" population, in this case red-headed people, is selected. The larger the sample the more accurate will be the results, but practical limitations will inevitably limit the size of the sample. Let us suppose that you have measured the IQ for each of 10 randomly-selected red-headed people, and the 10 values appear in the Table (2.1) below:

Table 2.1

106
111
97
119
88
126
104
93
115
108
EX = 1067

Step 4. Find the region of rejection.

The region of rejection can be found in terms of critical z-scores -- the z-scores that cut-off an area of the normal distribution that is exactly equal to alpha (the region of rejection is equal to half of alpha in each tail of a two-tailed test). Because we have set alpha = .05 and planned a two-tailed test, the critical z-scores are +1.96 and -1.96. The regions of rejection are therefore the

portions of the standard normal distribution that are above (i.e., to the right of) +1.96, or below (i.e., to the left of) -1.96.

Step 5. Calculate the test statistic.

The first step is to calculate the mean of the sample, \bar{X} . We add up the 10 numbers to get the sum of X, which equals 1067 (as indicated in the above table). Then we divide the sum by the size of the sample, N, which equals 10. So $\bar{X} = 1067/10 = 106.7$. Immediately we see that the sample mean is above the population mean. Had we planned a one-tailed test and hypothesized that red-heads would be less intelligent, we could not ethically proceed with our statistical test (to switch to a two-tailed test after you see the data implies that you are using an alpha of .05 in the hypothesized tail and .025 in the other tail, so your actual alpha would be the unconventionally high value of .075).

Because we planned a two-tailed test we can proceed, but to perform the one-sample z test we must know both the mean and the standard deviation of the comparison population. The mean IQ for the general population, which is also the mean specified by the null hypothesis is set at 100 (the average IQ score based on a great deal of data is converted to 100). The standard deviation for IQ (Wechsler test) is set to 15. Now we have all the values we need to use Formula 2.2' to get the z-score:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{106.7 - 100}{\frac{15}{\sqrt{10}}} = \frac{6.7}{4.74} = 1.41$$

In calculating the z-score for groups, there are only a few opportunities to make errors. A common error is to forget to take the square-root of N. Perhaps, an even more common error is to leave N out of the formula entirely, and just divide by the population standard deviation instead of the standard error of the mean. Without the square-root of N, the z-score for groups looks a lot like the ordinary z-score for individuals, and students sometimes complain that they never know which type of z-score to use. The following rule should be kept in mind: if we are asking a question about a group, such as whether a particular group is extreme or unusual, then we are really asking a question about the mean of the group, and should use the z-score for groups (Formula 2.2). Only when you are concerned about one individual score, and where it falls in a distribution, should you use the simple z-score for individuals (Formula 1.6).

Step 6. Make the statistical decision.

If the z-score you calculated above is greater in magnitude than the critical z-score, then you can reject the null hypothesis. As alpha had been set to .05, we would say that the results are significant at the .05 level. However, the calculated (sometimes called "obtained") z-score is for this example less than the critical z, so we cannot reject the null hypothesis. It is sometimes said that we therefore "accept" the null hypothesis, but most researchers seem to prefer stating that we have insufficient evidence to reject the null hypothesis.

For the example mentioned above, rejecting the null hypothesis would have meant concluding that red-headed people have a mean IQ that is different from the general population (based on your data you would, of course, indicate in which direction the difference fell). However, because our calculated z-score was rather small, and therefore fell too near the middle of the null hypothesis distribution, we would have to conclude that we do not have sufficient evidence to say that red-headed people differ in IQ from the rest of the population.

Note: For a review of confidence intervals for the population mean, see the next chapter of this guide, where CI's will be described in terms of the t distribution (even though you don't really need the t distribution when dealing with very large samples, the t distribution always gives you the appropriate values, regardless of sample size).

Definitions of Key Terms

Null Hypothesis: This is a specific hypothesis that we would like to disprove, usually one which implies that your experimental results are actually due entirely

to chance factors involved in random sampling. In the one-sample case the null hypothesis is that the mean of your "special" or experimental population is the same as the comparison population.

Alternative Hypothesis: This hypothesis is usually not stated specifically enough to be tested directly, but it is usually stated in such a way that rejecting the null hypothesis implies that this hypothesis is true. It is related to the research hypothesis which we want to prove.

One-tailed Test: The alternative hypothesis is stated in such a way that only results that deviate in one particular direction from the null hypothesis can be tested. For instance, in the one-sample case, a one-tailed alternative hypothesis might state that the "special" population mean is larger than the comparison population mean. The sample mean can then be tested only if it is larger than the comparison population mean; no test can be performed if the sample mean is smaller than the comparison population mean.

Two-tailed Test: The alternative hypothesis is stated in such a way that sample results can be tested in either direction. Alpha is divided in half; half of the probability is placed in each tail of the null hypothesis distribution.

Type I error: Rejecting the null hypothesis when the null hypothesis is actually true and should not be rejected. This acts like a false alarm, and can lure other experimenters into wasting time exploring an area where there is really no effect to find.

Type II error: Accepting (or failing to reject) the null hypothesis when the null hypothesis is actually false and should be rejected. This is a "miss"; a real effect has been missed, and other experimenters may mistakenly abandon exploring an area where a real effect is going on.

p level: This is the probability that when the null hypothesis is true you will get a result that is at least as extreme as the one you found for your selected sample (or experimental group).

Alpha level: This is the percentage of type I errors you are willing to tolerate. When the p level is less than the alpha level you are willing to take the chance of making a type I error by rejecting the null hypothesis.

Statistical Significance: The p level is low enough (i.e., below alpha) that the null hypothesis can be ignored as an explanation for the experimental results. Statistical significance should not be confused with practical significance or importance. Ruling out chance as an explanation for your results does not imply that your results are strong enough to be useful or meaningful.

Test statistic: A combination of sample statistics that follows a known mathematical distribution and can therefore be used to test a statistical hypothesis. The z-score for groups is an example; it is a sample statistic (i.e., the sample mean) and follows the standard normal distribution.

Critical value: This is the value of the test statistic for which the p level is exactly equal to the alpha level. Any test statistic larger than the critical value will have a p level less than alpha, and therefore lead to rejecting the null hypothesis.

Unbiased Sample Variance (s^2): This formula (i.e., $SS/N-1$), applied to a sample, provides an unbiased estimate of F^2 . When we refer in this text to just the "sample variance" (or use the symbol " s^2 ") it is this formula to which we are referring.

Unbiased Sample Standard Deviation (s): The square-root of the unbiased sample variance. Although not a perfectly unbiased estimate of F , we will refer to s (i.e., $\sqrt{s^2}$) as the unbiased standard deviation of a sample, or just the sample standard deviation.

Degrees of Freedom (df): The number of scores that are free to vary after one or more parameters have been found for a distribution. When finding the variance or standard deviation of one sample, $df = N-1$, because the deviations are taken from the mean, which has already been found.

Chapter 3

One-Group t test

To illustrate the calculation of a one-group t-test and the construction of a confidence interval, we ask you to imagine the following situation. [**Note:** The one-group t test was left out of the Essentials book, due to space limitations (it is very rarely used); it is included here for conceptual completeness.]

An anthropologist has just discovered a previously unknown group of humans native to Antarctica. These "Antarcticans" have adapted to their cold climate, and apparently function quite well at a body temperature somewhat below that which is considered normal for the rest of the human population. The anthropologist wants to know whether the mean body temperature for the population of Antarcticans is really less than the mean body temperature of other humans. Let us suppose that 98.6° F is the population mean (μ) for humans in general, but that we do not have enough information to establish the population standard deviation (σ). We will also suppose that the anthropologist was able to take the temperatures of only 9 Antarcticans. Because the anthropologist has a small sample size, and does not know the population standard deviation, she will have to perform a one-sample t-test.

The appropriate formula for this test requires that we calculate \bar{X} and s before proceeding. If the nine temperature measurements are as follows:

97.5, 98.6, 99.0, 98.0, 97.2, 97.4, 98.5, 97.0, 98.8

The methods of the previous chapter can be used to find that $\bar{X} = 98.0$, and $s = .75$. Now we have all of the values that are needed to compute the t statistic (we will use Formula 2.2', replacing z with t and F with s):

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}} = \frac{98.0 - 98.6}{\frac{.75}{\sqrt{9}}} = \frac{-.6}{.25} = -2.4$$

To complete the t-test, we need to know the critical value from the t distribution, so we need to know the number of degrees of freedom involved. For a one-group test, $df = N - 1$; in this case $df = 9 - 1 = 8$. For a two-tailed test with $\alpha = .05$, the critical t (from Table A.2) = 2.306. Because this is a two-tailed test the region of rejection has two parts -- one in each tail -- greater than +2.306, or less than -2.306. The calculated t of -2.4 is less than -2.306, so the calculated t falls in the rejection zone. Therefore we can conclude that the difference in body temperatures between Antarcticans and other humans is statistically significant (at the .05 level).

It cannot be over-emphasized that in order for this test to be valid, the sample must be truly an independent random sample. If only the younger Antarcticans, or the friendlier ones (could the friendlier ones be "warmer?") are included, for instance, the conclusions from this test would not be valid.

Because body temperature is so tightly constrained by physiological requirements, it would be interesting to find any group of people that maintained a different body temperature, even if the difference was fairly small. So the conclusion of the hypothesis test above would be of some interest in itself. But scientists in several disciplines would probably want more information; many would want to know the mean body temperature of all Antarcticans (μ_A). Based on the study above, the point estimate for μ_A would be $\bar{X} = 98.0$. However, there is a certain amount of error associated with that estimate, and the best way to convey the uncertainty involved is to construct a confidence interval, as shown below.

One-group Confidence Interval

The sample size being small, and the population standard deviation unknown, the appropriate formula is Formula 3.7. \bar{X} has already been calculated ($\bar{X} = 98.0$),

and s_x is the denominator of the t test calculated above ($s_x = .25$). If we choose to construct a 95% CI, we must also find t_{crit} for $df = 8$, and $\alpha = .05$, two-tailed (in general, for a XX% CI, $\alpha = 1 - .XX$, two-tailed). This is the same t_{crit} we used for the hypothesis test above: $t_{crit} = 2.306$. The upper and lower limits for μ are calculated as shown below.

$$\mu_{lower} = \bar{X} - t_{crit} s_{\bar{x}} = 98.0 - (2.306)(.25) = 98.0 - .577 = 97.42$$

$$\mu_{upper} = \bar{X} + t_{crit} s_{\bar{x}} = 98.0 + (2.306)(.25) = 98.0 + .577 = 98.58$$

Based on the above calculations we can say that our confidence is 95% that the mean body temperature for the entire population of Antarciticans is between 97.42 °F and 98.58 °F. Note that the mean for the general population, 98.6 °F, does not fall within the 95% confidence interval, which is consistent with the results of our null hypothesis test.

Of course, any particular confidence interval is either right or wrong -- that is, it either contains the population mean or it does not. We never know which of our confidence intervals are right or wrong, but of all the 95% CI's we construct, we know that about 95% of them will be right (similarly, 99% of our 99% CI's will be right).

Two-group t test

To review the statistical analysis of a two-group experiment, we will describe a hypothetical study from the field of neuropsychology. A researcher believes that a region in the middle of the right hemisphere of the brain is critical for solving paper-and-pencil mazes, and that the corresponding region of the left hemisphere is not involved. She is lucky enough to find 6 patients with just the kind of brain damage that she thinks will disrupt maze-learning. For comparison purposes, she is only able to find 4 patients with similar damage to the left hemisphere. Each of the 10 patients is tested with the same maze. The variable measured is how many trials it takes for each patient to learn the maze perfectly (i.e., execute an errorless run). The data collected for this hypothetical experiment appear in Table 3.1 below:

Table 3.1

Left Damage	Right Damage
5	9
3	13
8	8
6	7
_____	11
_____	6
$\bar{X}_L = 22/4 = 5.5$	$\bar{X}_R = 54/6 = 9$

From Table 3.1 you can see that the results are in the predicted direction: the mean number of trials required to learn the maze was considerably higher for the group with damage to the right side of the brain. Fortunately, the amounts of variability in the two groups are fairly similar, so it is appropriate to perform the pooled-variance t-test. The two variances are: 4.33 (left), and 6.8 (right). So, according to Formula 3.5, the pooled variance is:

$$s_p^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2} = \frac{3(4.33) + 5(6.8)}{3 + 5} = \frac{47}{8} = 5.8$$

Therefore, the pooled-variance t (Formula 3.6) is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}} = \frac{(9 - 5.5)}{\sqrt{5.875 \left(\frac{1}{4} + \frac{1}{6} \right)}} = \frac{3.5}{\sqrt{2.448}} = \frac{3.5}{1.5646} = 2.237$$

We should point out that we deliberately subtracted the sample means in the numerator in the order that would make the t value come out positive. This is frequently done to avoid the awkwardness of working with negative numbers. We already know from looking at the two sample means which is larger; we don't need the sign of the t value to tell us in which direction the results fell.

To find the critical t values we must first calculate the number of degrees of freedom. For the pooled-variance t test, $df = n_1 + n_2 - 2 = 4 + 6 - 2 = 10 - 2 = 8$. For alpha = .05, two-tailed, the critical values are -2.306 and +2.306. Because our calculated t (2.237) is smaller than the critical t, we cannot reject the null hypothesis, and we cannot say that our results are significant at the .05 level.

Two-group Confidence Interval

Although you would probably not bother to calculate a confidence interval for the above experiment because of the small sample sizes and lack of statistical significance, we will find the 95% confidence interval just for review. We will use Formula 3.8, and plug in the values from the above example:

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{crit} s_{\bar{x}_1 - \bar{x}_2} = 3.5 \pm (2.306)(1.5646) = 3.5 \pm 3.61$$

Note that for the 95% confidence interval the critical t values are the same that were used for the .05, two-tailed hypothesis test. Also note that the value for $s_{\bar{x}_1 - \bar{x}_2}$ in the above formula is the denominator in our calculation of t above. The 95% confidence interval extends from -.11 to +7.11 maze-learning trials. The fact that the 95% confidence interval contains zero tells us that we would not be able to reject zero as the null hypothesis, using alpha = .05, two-tailed. This is consistent with the result of our hypothesis test.

Matched t test

For the following example, imagine that each subject views the same videotape of a husband and wife arguing. Afterwards, the subject is asked to rate the likability (on a 0 to 10 scale) of both the husband and the wife. This is an example of simultaneous repeated-measures. Each subject's ratings are sorted according to whether the subject is rating someone of the same gender (e.g., a man rating the likability of the husband), or the opposite gender (e.g., the same man rating the likability of the wife). In this example, six individuals (3 of each gender) were chosen at random, each providing two different ratings. The two ratings of each subject appear in Table 3.2 below (only six subjects are included in this example in order to minimize the amount of calculation).

Table 3.2

<u>Subject</u>	<u>Same</u>	<u>Opposite</u>	<u>D</u>
1	9	5	+4
2	5	5	0
3	8	3	+5
4	4	5	-1
5	6	3	+3
6	7	9	<u>-2</u>
			+9

The scores can be subtracted in either order (as long as you are consistent for every pair, of course), so you may want to choose the order that minimizes the minus signs. In finding the sum of the difference scores, you may want to add the positive and negative numbers separately, and then add these two sums at the end. For instance, in Table 3.2 the positive numbers are +4, +5 and +3, adding to +12. The negative numbers are -1 and -2, adding to -3. Finally, adding +12 and -3, we

find that $ED = +9$. Dividing this sum by N , which is 6 for this example, we find that $\bar{D} = 1.5$. The (unbiased) standard deviation of the difference scores, s_D , is 2.88 (if you use a calculator to obtain the SD, make sure you enter the negative difference scores with minus signs, or you will get the wrong answer). The matched t value can now be found by using Formula 3.12:

$$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{N}}} = \frac{1.5}{\frac{2.88}{\sqrt{6}}} = \frac{1.5}{1.18} = 1.28$$

Because the N is so small we must use the t distribution to represent our null hypothesis. The $df = N - 1$, which equals $6 - 1 = 5$, so the critical t ($\alpha = .05$, two-tailed) is 2.571 (see Table A.2). Because the calculated t (1.28) is well below this value, the null hypothesis -- that the mean of the difference scores is zero -- cannot be rejected. This lack of statistical significance is also obvious from the 95% confidence interval for the difference of the populations, as given by Formula 3.13.

$$\mu_{lower} = \bar{D} - t_{crit} s_{\bar{D}} = 1.5 - 2.571(1.18) = 1.5 - 3.03 = -1.53$$

$$\mu_{upper} = \bar{D} + t_{crit} s_{\bar{D}} = 1.5 + 3.03 = 4.53$$

Zero is clearly contained in the 95% CI, so the null hypothesis of zero difference between the two populations cannot be rejected.

Assumptions: The assumptions underlying the matched t -test are that the difference scores are: 1) normally distributed, and 2) random and independent of each other.

When to use the matched t -test

Repeated measures. The two basic types are: *simultaneous* and *successive*. The successive repeated-measures experiment has two major sub-types: *the before-after design* and *the counterbalanced design*. The before-after design usually requires a control group in order to draw valid conclusions. The counterbalanced design does not need a control group, but it is susceptible to carry-over effects; if the carry-over effects are not symmetrical, repeated measures should not be used.

Matched-Pairs. The two main sub-types are: *experimental* and *natural*. The experimenter may create pairs based either on a relevant pre-test, or on other available data (e.g., gender, age, IQ, etc.). Another way to create pairs is to have two subjects rate or "judge" the same stimulus. Rather than creating pairs, the experimenter may use naturally occurring pairs.

Definitions of Key Terms

t distribution: Actually, there is a family of t distributions that differ according to the number of degrees of freedom. Each t distribution looks like the normal distribution except with fatter tails. As df increases, the t distribution more closely approximates the normal distribution.

One-sample t -test: This test is similar to the one-sample z -test, except that the sample standard deviation is used in place of the population standard deviation, and the critical values are found from the t distribution. It is used when the sample size is not very large.

Point estimation: Estimating a population parameter with a single value, such as using the sample mean as an estimate of the population mean.

Interval estimation: Using a range of values to estimate a population parameter.

Confidence Interval: These intervals are constructed so that a certain percentage of the time the interval will contain the specified population parameter. For instance, a 95% CI for the population mean will not always contain the population

mean, but 95% of such intervals will.

Standard error of the difference: This is short for "standard error of the difference between means". It is the value you would get if you kept selecting two random samples at a time, finding the difference between the means of the two samples, and after piling up a huge number (preferably, infinite) of these difference scores, calculating the standard deviation.

Large-sample test for a difference between means: This test is valid only when the two samples are quite large (40 subjects per group, at the very least). The two sample variances are not pooled. The critical values are found from the standard normal distribution.

Pooled-variance: This is short for "pooled-variance estimate of the population variance". It is a weighted-average of the two sample variances, which produces the best estimate of the population variance when the two populations being sampled have the same variance.

Pooled-variances t-test: This t-test is based on the use of the pooled-variance to estimate the population variance, and, strictly speaking, is only valid when the two populations have the same variance.

Separate-variances t-test: This t-test is appropriate when the samples are not very large, not equal in size, and it cannot be assumed that the two populations have the same variance. The sample variances are not pooled, but rather each is separately divided by its own sample size. The critical values must be found through special procedures.

Homogeneity of variance: This means that the two populations have the same variance. This is one of the assumptions that underlies the use of the pooled-variance t-test.

Heterogeneity of variance: This means that the two populations do not have the same variance. This is the conclusion that is made if a test for homogeneity of variance is statistically significant.

RM or Matched t-test. Also called t-test for (the difference of) correlated (or dependent) means (or samples). Compared to the independent groups t-test, this t-test tends to have a smaller denominator (i.e., standard error) to the extent that the two sets of measures are positively correlated.

Direct-difference method. Difference scores are computed for each pair of scores, and then a one-group t-test is performed on the difference scores, usually against the null hypothesis that the mean of the difference scores in the population (μ_D) is zero.

Repeated-measures design. Also called a within-subjects design. Each subject is measured twice on the same variable, either before and after some treatment (or period of time), or under different conditions (simultaneously or successively presented).

Order effects: When each subject is measured twice under successive conditions, order effects can increase or decrease the measurements according to whether a condition is administered first or second. Practice and fatigue are examples of order effects.

Counterbalanced design. When a successive repeated-measures design involves two conditions, counterbalancing requires that half the subjects receive the conditions in one order, while the other half receive the conditions in the reverse order. This will average (or "balance") out simple order effects.

Matched-pairs design. Instead of measuring the same subject twice, subjects are paired off based on their similarity on some relevant variable, and then randomly assigned to the two conditions. Naturally occurring pairs can be used, but this will limit the conclusions that can be drawn.

Chapter 4

Correlation

In order to review the calculation of Pearson's r , we will describe a hypothetical study to determine whether people who come from large immediate families (i.e., have many siblings) tend to create large immediate families (i.e., produce many children). In this example, each "subject" is actually an entire immediate family (parents and their children). The "X" variable is the average number of siblings for the two parents (e.g., if the mother has 1 brother and 1 sister, and the father has 2 brothers and 2 sisters, $X = 3$), and the "Y" variable is the number of children in the selected family. The families should be selected in an independent random manner (in this case, the entire family is being selected as though a single subject). For this hypothetical example, 10 families were selected, so $N = 10$. The data in Table 4.1 consist of the mean number of parental siblings, X , and the number of children for each of the 10 selected families, Y , as well as the (XY) cross-products required for computing the correlation.

Table 4.1

<u>X</u>	<u>XY</u>	<u>Y</u>
3	9	3
1	2	2
4	20	5
2	4	2
1.5	1.5	1
2	8	4
4	12	3
2.5	10	4
2	4	2
<u>1</u>	<u>1</u>	<u>1</u>
23	71.5	27

As long as we are satisfied with assessing the degree of linear correlation, the appropriate test statistic is Pearson's r . First, we will calculate the means and standard deviations, so that we can use Formula 4.2. The mean of X equals $23/10 = 2.3$, and the mean of Y equals $27/10 = 2.7$. Using the biased formula for SD, you can verify that $F_x = 1.0296$, and $F_y = 1.2688$. Plugging these values into Formula 4.2, we get:

$$r = \frac{\frac{\Sigma XY}{N} - \mu_x \mu_y}{\sigma_x \sigma_y} = \frac{\frac{71.5}{10} - (2.3)(2.7)}{(1.0296)(1.2688)} = \frac{7.15 - 6.21}{1.3064} = \frac{.94}{1.3064} = .71955$$

If you have already calculated the unbiased SD's ($s_x = 1.08525$, $s_y = 1.3375$), you need to use Formula 4.3. The value for r will come out to 1.0444 divided by $(1.08525)(1.3375)$, which equals $1.0444 / 1.4515 = .71952$. As you can see both formulas produce the same value for Pearson's r , except for a slight difference due to rounding off. Because the value for r will never be larger than 1.0, one must be careful not to round off too much on the intermediate steps -- for instance, in dealing with the standard deviation, at least four digits after the decimal point should be retained (as in the example above).

The r we calculated can be tested for statistical significance by computing a t value using Formula 4.4'.

$$\frac{\sqrt{N-2}}{1-r^2} = \frac{.71955 \sqrt{10-2}}{\sqrt{1-.518}} = \frac{.71955 (2.828)}{\sqrt{.482}} = \frac{2.035}{.694}$$

To look up the critical value for t , we need to know that the number of degrees of freedom equals $N - 2$; for this example $df = 10 - 2 = 8$. Looking in Table A.2 under $\alpha = .05$, two-tailed, we find that the critical $t = 2.306$. Because $2.93 > 2.306$, the null hypothesis (i.e., $D = 0$) can be rejected.

Our statistical conclusion, it must be noted, is only valid if we have met the

assumptions of the test, which are as follows:

1. Independent random sampling. In particular, any restriction in the range of values sampled can threaten the accuracy of our estimate of D.
2. Normal distributions. Each variable should be inspected separately to see that it follows an approximately normal distribution.
3. Bivariate normal distribution. The scatterplot of the data should be inspected for unusual circumstances, such as bivariate outliers or curvilinearity.

Regression

We will begin reviewing regression with another example of linear correlation. Imagine that a career counselor is interested in how well college grade point averages (GPA's) predict annual salaries five years after graduation. From her records, she selects at random six individuals who graduated five years ago, and finds out their final cumulative GPA's (X), as well as their current annual salaries (Y). The data are shown in the Table 4.2 (highest possible GPA is 4.0, which means A's in all courses; annual salary is expressed in thousands of dollars), along with columns for the squared values and cross-products.

Table 4.2

<u>X</u>	<u>XY</u>	<u>Y</u>
3.1	99.2	32
2.5	67.5	27
3.6	108.0	30
2.2	52.8	24
3.3	92.4	28
<u>2.7</u>	<u>59.4</u>	<u>22</u>
17.4	479.3	163

Assuming that the two variables are linearly related, it can be useful to find the linear regression equation that predicts annual salary based on the subject's college GPA. The first step is to find the slope, which can be found from Pearson's r and the standard deviations. Assuming that we had already calculated the unbiased standard deviations ($s_x = .5254$ and $s_y = 3.7103$), the next step is to calculate Pearson's r using Formula 4.3:

$$r = \frac{\frac{1}{N-1} (\Sigma XY - N \bar{X} \bar{Y})}{s_x s_y} = \frac{\frac{1}{5} [479.3 - 6(2.9)(27.17)]}{(.5254)(3.7103)} = \frac{1.32}{1.949} = .6771$$

Now we can use Formula 4.6, to find that the slope equals:

$$b_{yx} = r \frac{s_y}{s_x} = .6771 \left(\frac{3.7103}{.5254} \right) = .6771 (7.062) = 4.782$$

The second step in determining the regression equation is to find the Y-intercept using Formula 4.7. To use this formula we need the slope found above, and the means of both variables, which are: $\bar{X} = \Sigma X/N = 17.4/6 = 2.9$; and $\bar{Y} = \Sigma Y/N = 163/6 = 27.17$.

$$a_{yx} = \bar{Y} - b_{yx} \bar{X} = 27.17 - 4.7826 (2.9) = 27.17 - 13.87 = 13.3$$

Finally, the regression equation is found by plugging the slope and Y-intercept into Formula 4.8, as shown below:

$$Y' = b_{yx}X + a_{yx} = 4.78X + 13.3$$

The equation above can be used to predict the salary for any GPA, even those not in our original sample. For instance, let us consider a GPA less than any in the sample -- one that would barely allow graduation -- $X = 1.8$. The predicted annual salary five years after graduation would be:

$$Y' = 4.786 (1.8) + 13.3 = 8.615 + 13.3 = 21.9 \text{ (i.e., \$21,900)}$$

The most common uses for linear regression are:

- 1) Predicting future performance from measures taken previously.
- 2) Statistically "removing" the effects of a confounding or unwanted variable.
- 3) Evaluating the linear relationship between the quantitative levels of a truly independent (i.e., manipulated) variable and a continuous dependent variable.
- 4) Testing a theoretical model that predicts values for the slope and Y-intercept of the regression line.

Definitions of Key Terms

Perfect linear correlation: When positive, perfect correlation means each subject has the same z-score on both variables. Perfect negative correlation means each subject has the same z-score, in magnitude, on both variables, but the z-scores are always opposite in sign.

Positive correlation: As the magnitude of one variable increases, the second variable tends to increase, as well, and as one decreases, the other tends to decrease.

Negative Correlation: As the magnitude of one variable increases, the second variable tends to decrease, and decreases in the first variable tend to be associated with increases in the second variable.

Linear transformation: Occurs when one variable is changed into another variable by the adding, subtracting, multiplying, and/or dividing of constants, but by no other types of mathematical operations. After a linear transformation, each subject has the same z-score as before the transformation.

Scatterplot (or scattergraph): A graph in which one variable is plotted on the X-axis, while the other variable is plotted on the Y-axis. Each subject (or observation) is represented by a single dot on the graph.

Pearson's correlation coefficient (r): Measures the degree of linear relationship between two variables. Ranges from -1 for perfect negative correlation to 0 for a total lack of linear relationship to +1 for perfect positive correlation. Also called "Pearson's product-moment correlation coefficient."

Population correlation (D): The Pearson correlation coefficient calculated on an entire population.

Truncated (or restricted) range: Occurs when a sample fails to include the full range of values of some variable that are represented in the population. This usually reduces the magnitude of the sample r as compared to D.

Curvilinear correlation: Occurs when two variables are related in such a way that the scatterplot appears as a curve instead of a straight line. There are coefficients of curvilinear correlation that are sensitive to such relationships.

Bivariate Outliers: These are data points that need not be extreme on either variable separately, but rather represent an unusual combination of values of the two variables. A few or even just one bivariate outlier can greatly influence the correlation coefficient.

Bivariate distribution: This is a distribution that represents the relative likelihood for each possible pair of values for two variables. In order to test Pearson's r for significance, it must be assumed that the two variables follow a bivariate normal distribution.

Covariance: Measures the tendency of two variables either to vary together or to vary consistently in opposite directions. Must be divided by the product of the standard deviations of the two variables in order to yield a correlation coefficient that varies in magnitude between 0 and 1.

Regression towards the mean: The statistical tendency for extreme values on one variable to be associated with less extreme (i.e., closer to the mean) values on a second less-than-perfectly correlated variable.

Regression line: When regressing Y on X, this is the straight line that minimizes the squared differences between the actual Y values, and the Y values predicted by the line. Regressing X on Y leads to a different line (unless correlation is perfect) that minimizes the squared differences in the direction of the X-axis.

Slope of the regression line: When regressing Y on X, the slope is the amount by which the Y variable changes when the X variable changes by one unit.

Y-intercept of the regression line: When regressing Y on X, the Y-intercept is the value of Y when X is zero.

Residual: In linear regression, this is the portion of an original score that is left over after a prediction has been subtracted from it.

Variance of the estimate ($F^2_{\text{est } y}$): Also called residual variance. When regressing Y on X, it is the variance of the Y values from the regression line, or equivalently, the variance of the residuals, after each Y value has been subtracted from its predicted value.

Coefficient of determination: This is the proportion of the total variance that is "explained" by linear regression. It is the ratio of explained to total variance, and equals r^2 .

Coefficient of nondetermination: This is the proportion of the total variance that is "unexplained" by linear regression. It is the ratio of unexplained to total variance, and equals $1 - r^2$, which is sometimes symbolized as k^2 .

Homoscedasticity: When this condition is true for linear regression, it means that the variance around the regression line at one location (i.e., one particular X value) is the same as at any other location.

Point-biserial r: A Pearson's correlation coefficient that is calculated when one of the variables has only two possible values. It can be used to measure the strength of association between the independent and dependent variables in a two-group experiment.

Omega-squared (Γ^2): The proportion of variance accounted for in one variable by another (usually discrete) variable in a population.

Chapter 5

One-way ANOVA

To illustrate the calculation of a one-way independent-groups ANOVA, we will describe a hypothetical experiment involving four groups. Imagine that a social psychologist wants to know if information about a person affects the way someone judges the attractiveness of that person. This is a very general question, and it has many aspects that can be tested. For this example, we will focus on the perceptions of women, and create four specific conditions. In each condition a female subject looks at a photograph of a male and rates the male for attractiveness on a scale from 1 to 10 (all subjects view the same photograph, which is chosen for being about average in attractiveness). There are four experimental conditions, depending on the short "biography" that accompanies the photo: successful biography, failure biography, neutral biography, or no biography at all. The effects of these four conditions can be explored by performing a one-way independent-groups ANOVA.

In this example, only six subjects will be assigned to each condition to reduce the amount of calculation. In order to illustrate the case of unequal sample sizes we will imagine that one subject from each of two conditions had to be eliminated (e.g., it was later found that the subject actually knew the person in the photograph, used the scale incorrectly, etc.). The resulting n's are: $n_1 = 5$,

$n_2 = 5, n_3 = 6, n_4 = 6$. The attractiveness rating given by each subject and means and (unbiased) SD's for each condition, appear in the Table 5.1:

Table 5.1

	<u>Control</u>	<u>Neutral</u>	<u>Failure</u>	<u>Success</u>
	4	5	4	6
	5	6	5	5
	4	5	3	7
	3	4	4	6
	6	4	2	5
			5	6
Mean =	4.4	4.8	3.833	5.833
SD =	1.14	.837	1.17	.753

Because the sample sizes are not all equal, we need to use the general formula for one-way ANOVA, Formula 5.4. To use this formula, we need to first calculate the grand mean. \bar{X}_G is equal to the sum of all the scores in Table 5.1 divided by the total N, so $\bar{X}_G = 104 / 22 = 4.727$. The numerator of Formula 5.4 is MS_{bet} ; inserting the means from Table 5.1 and the grand mean, the following result is obtained (k=4):

$$\frac{5(4.4-4.727)^2 + 5(4.8-4.727)^2 + 6(3.833-4.727)^2 + 6(5.833-4.727)^2}{4 - 1}$$

$$\frac{5(.107) + 5(.00533) + 6(.8) + 6(1.224)}{3} = \frac{.535 + 2.8}{3}$$

So $MS_{bet} = 12.7 / 3 = 4.23$. The next step is to find MS_w using the denominator of Formula 5.4:

$$\frac{4(1.14)^2 + 4(.837)^2 + 5(1.17)^2 + 5(.753)^2}{(5 - 1) + (5 - 1) + (6 - 1) + (6 - 1)} = \frac{5.2+2.8}{18}$$

So $MS_w = .98$. Finally, we calculate the F ratio: $F = MS_{bet}/MS_w = 4.23/.982 = 4.31$.

In order to find the appropriate critical F value, we need to know the df for both the numerator and denominator of our F ratio: $df_{bet} = k - 1 = 4 - 1 = 3$, and $df_w = N_T - k = 22 - 4 = 18$. Therefore, we look down the column labeled "3" in Table A.3 until we hit the row labeled "18" to find that the critical F = 3.16. Because the calculated F (4.31) is greater than the critical F, the null hypothesis can be rejected at the .05 level. Our hypothetical researcher can reject the hypothesis that all of the population means are equal, but without further testing (see next section), she cannot say which pairs of population means are different from each other.

Summary Table

The results of the above ANOVA can be summarized in the following "source" table:

Table 5.2

<u>Source</u>	<u>SS</u>	<u>df</u>	<u>MS</u>	<u>F</u>	<u>p</u>
Between-groups	12.7	3	4.23	4.31	<.05
Within-groups	17.67	18	.98		
Total	30.36	21			

Multiple Comparisons

When dealing with four groups, Tukey's HSD is favored over Fisher's procedure, because of its tighter control over the experiment-wise Type I error rate. However, Tukey's test is based on equal sample sizes. Fortunately, with small discrepancies in size, there is little error involved in calculating the harmonic mean of the sample sizes, and then pretending that all of the samples have this size. The harmonic mean of 5, 5, 6, and 6 is found as follows:

$$\frac{4}{\frac{1}{5} + \frac{1}{5} + \frac{1}{6} + \frac{1}{6}} = \frac{4}{.2 + .2 + .167 + .167} = \frac{4}{.734} =$$

We still look up q in Table A.4 with 18 as the df for the error term (and, of course, the number of groups is 4). Inserting q and the value found above for n into Formula 5.12, we find that:

$$HSD = q_{crit} \sqrt{\frac{MS_w}{n}} = 4.0 \sqrt{\frac{.982}{5.45}} = 4.0 (.4245) = 1.7$$

Only the means for the failure and success conditions, which differ by 2.0, have a difference larger than HSD (1.7), so only these two conditions differ significantly according to Tukey's test.

It is instructive to see what conclusions we would have reached had we used the less conservative LSD procedure. Looking up the (two-tailed) .05 critical t for 18 df, and again using 5.45 as our sample size, LSD can be found using Formula 5.11:

$$LSD = t_{crit} \sqrt{\frac{2MS_w}{n}} = 2.101 \sqrt{\frac{2(.982)}{5.45}} = 2.101 (.60) = 1.26$$

With LSD, success differs significantly not only from failure, but from the control condition, as well (success - control = 1.43 > 1.26). According to Hayter's modification of the LSD test, the significance of your ANOVA allows you to test your pairwise comparisons with an HSD based on $k-1$, rather than k ; for this example, q for $k-1$ is 3.61, instead of 4.00. Therefore, the modified LSD is (see HSD calculation above): 3.61 (.4245) = 1.53. This is not as liberal as the ordinary LSD test, and in this example, leads to the same conclusion as HSD (success - control = 1.43 < 1.53).

With 4 conditions, 6 pairwise comparisons are possible ($4*3/2 = 12/2 = 6$). If we had planned to test just 5 of the 6 (perhaps, we thought it unnecessary to test the difference between neutral and control), we might consider adjusting our alpha for each comparison according to the Bonferroni correction. To keep α_{EW} to .05, α_{PC} is set to $.05/5 = .01$. To perform the Bonferroni test for our five pairs we can use the LSD formula (5.11) with the critical at the .01 level. Because $t_{.01}(18) = 2.878$, the difference required by the Bonferroni test is $2.878 (.60) = 1.73$. This is even stricter (i.e., larger) than HSD (1.7), and had we planned all 6 tests, it would have been harder to get a significant pairwise comparison. The Bonferroni adjustment is too conservative, if you plan to test all or nearly all of the possible pairs of means. For this example, the Bonferroni test has more power than HSD only if you can narrow down your planned tests to 4 or less out of the possible 6.

Complex Comparisons

Suppose that you had planned to compare the failure condition with the average of the three others. The L for this comparison would be $3.833 - 5.011 = -1.178$. You can test this L for significance with Formula 5.14, if you use the harmonic mean of the sample sizes (found above) as n (the results would be a bit more accurate

with the formula for unequal sample sizes as given in Cohen, 2000).

$$\frac{.45(1.178)^2 / \left(\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + (-1)^2 + \left(\frac{1}{3}\right)^2 \right)}{.982} = \frac{7.563}{1.333} / .982$$

The critical F for this linear contrast is not the same as for the original ANOVA, because the contrast has only one df in its numerator; $F_{.05}(1, 18) = 4.41$. In this case, the contrast is significant, because $5.78 > 4.41$. The result would not have been so fortunate if for some strange reason you had planned to compare the average of the success and control conditions with the average of neutral and failure. L for this comparison is $5.1165 - 4.3165 = .8$. The F ratio for testing this contrast comes out to:

$$\frac{5.45(.8)^2 / \left(\left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right)}{.982} = \frac{3.488}{1.0} / .982 = 3.55$$

For this contrast, the calculated F is less than the critical F ($3.55 < 4.41$), so you would have lost the gamble.

It should be noted that if the failure vs. others contrast had not been planned, but rather noticed after the data were inspected, the appropriate critical value would come from **Scheffé's test**: $F_s = df_{bet} F_{ANOVA}$. For this example, $F_s = 3(3.16) = 9.48$. By this stringent criterion, the failure vs. others contrast would not be significant at the .05 level ($5.78 < 9.48$). Such is the advantage of planning contrasts before collecting (or inspecting) the data.

Recommendations

Tukey's HSD test is recommended (or the more powerful but less known modification of LSD by Hayter) when: a) there are more than three groups; b) all the groups are the same size, or nearly so; c) only pairwise comparisons are performed; and d) the pairwise comparisons are chosen after seeing the results. For only three groups, Fisher's protected t-tests (or the Newman-Keuls test) are sufficiently conservative, and more powerful. If complex comparisons are chosen after seeing the results, the Scheffé test should be used. If a limited number of comparisons can be planned, an a priori procedure, such as the Bonferroni test, should be used.

Definitions of Key Terms

One-way ANOVA: An analysis of variance in which there is only one independent variable.

Factor: A factor is an independent variable, and has at least two levels. Each group in a one-way ANOVA corresponds to a different level of one factor.

Mean-square-between (MS_{bet}): When the null hypothesis is true, MS_{bet} is an estimate of the population variance based on the variance of the sample means and the sample sizes. When the null hypothesis is not true, the size of MS_{bet} depends on both the magnitude of the treatment effect and error variance.

Mean-square-within (MS_w): This is an estimate of the population variance based on the variances within each sample of the experiment. It serves as the denominator in an independent-groups ANOVA, and is often referred to as the error term.

F Distribution: a mathematical distribution that is followed by the ratio of two independent estimates of the same population variance. The shape of the distribution depends on the degrees of freedom for both the numerator and the denominator.

F-Ratio: A ratio of two independent estimates of the same population variance that follows one of the F distributions. It can be used to test homogeneity of variance or an ANOVA.

Grand Total: The sum of all the measurements in the study, regardless of group.

Grand Mean: The mean of all the measurements in the study, regardless of group. It can be found by dividing the grand total by N_T .

Summary Table: Displays the sum of squares (SS), mean-squares (MS) and degrees of freedom (df), according to source (i.e., between-groups, within-subjects, or total of both) in an ANOVA. The F-ratio and p value (or significance level) are usually displayed, as well.

Eta-squared: The square of this correlation coefficient is used to represent the proportion of variance accounted for by the independent variable in the results of a particular ANOVA. It is the same as r_{pb}^2 in the two-group case.

Experimentwise alpha (" α_{EW} "): The probability of making at least one type I error among all of the statistical tests that are used to analyze an experiment (sometimes the term familywise alpha, which has a more restricted definition, is preferred).

Alpha per comparison (" α_{pc} "): The alpha level used for each particular statistical test, in a series of comparisons.

Pairwise comparison: A statistical test involving only two sample means.

Complex comparison: A statistical test involving a weighted combination of any number of sample means.

A priori comparison: A comparison that is planned before the experiment is run (or before looking at the data).

A posteriori comparison: More often called a post hoc (i.e., after the fact) comparison, this comparison is chosen after seeing the data.

Fisher's protected t-tests: t-tests that are performed only after the one-way ANOVA has attained statistical significance. If homogeneity of variance can be assumed, MS_w is used to replace s^2 in the two-group t-test formula.

Fisher's Least Significant Difference (LSD): The difference between two sample means that produces a calculated t that is exactly equal to the critical t. Any larger difference of means is statistically significant. This test is applicable if the overall ANOVA is significant, and all the groups are the same size.

Tukey's Honestly Significant Difference (HSD): Like LSD, this is the smallest difference between two sample means that can be statistically significant, but this difference is adjusted (i.e., made larger) so that " α_{EW} " remains at the initially set level, regardless of the number of groups, and whether the complete null or only a partial null hypothesis is true.

Studentized range statistic: The distribution of the largest difference of means when more than two equal-sized samples from the same population are compared. The critical value from this distribution depends on the number of groups compared, and the size of the groups.

Conservative: Statisticians use this term to refer to procedures that are strict in keeping the type I error rate down to a preset limit.

Liberal: A term that refers to statistical procedures that are relatively lax about type I errors, and therefore produce fewer type II errors.

Newman-Keuls test: Like Tukey's HSD, this test is based on the studentized range statistic, but it uses different critical values for different pairwise comparisons depending on the "range" that separates two sample means.

Dunnett's test: This test applies to the special case when one particular sample mean (usually a control group) is being compared to each of the other means. When this test is applicable, it is usually the most powerful.

Scheffé's test: This post hoc comparison test maintains strict control over experimentwise alpha even when complex comparisons are involved. It is unnecessarily conservative when only pairwise comparisons are being performed.

Bonferroni-Dunn test: This test is only appropriate for planned (a priori) comparisons, and is based on the Bonferroni inequality. The desired " α_{EW} " is divided by the number of planned comparisons to find the appropriate " α_{pc} ".

Linear Contrasts: Weighted combinations of means in which the weights (or coefficients) sum to zero.

Orthogonal comparisons (or contrasts): These are comparisons that are independent of each other. As many as $k - 1$ (where k is the number of groups) comparisons can be mutually independent. Testing a set of orthogonal comparisons can be used as an alternative to the one-way ANOVA.

Chapter 6

Power for the two-group t test

Power analysis can be separated into two categories: fixed and flexible sample sizes.

1. When the sample sizes are fixed by circumstance (e.g., a certain number of patients available with a particular condition), the effect size is estimated in order to find the power. If the power is too low, the experiment would not be performed as designed. Conversely, one can decide on the lowest acceptable power, find the corresponding delta, and then put the fixed sample size into the equation to solve for d . The d that is found in this way is the lowest d that yields an acceptable level of power with the sample sizes available. If the actual effect size in the proposed experiment is expected to be less than this d , power would be too low to justify performing the experiment.

2. When there is a fair amount of flexibility in determining sample size, d is estimated and the delta corresponding to the desired level of power is found. The appropriate equation can then be used to find the sample sizes needed to attain the desired power level with the effect size as estimated. This procedure can also be used to set limits on the sample size. One finds the smallest d that is worth testing, and puts that d into the equation along with the delta corresponding to the lowest acceptable power level. The sample size that is thus found is the largest that can be justified. Using any more subjects would produce too great a chance of obtaining statistically significant results with a trivial effect size. Conversely, the largest d that can reasonably be expected for the proposed experiment is put into the same equation. The sample size that is found is the bare minimum that is worth employing. Using any fewer subjects would mean that the power would be less than acceptable even in the most optimistic case (d as large as can be expected).

The use of the power tables and formulas is reviewed below for both categories of power analysis.

1. A psychologist is studying the relationship between type of crime and sociopathy in prison inmates. For the comparison of arsonists to burglars, there are unfortunately only 15 of each available for testing. What is your power for this experiment? First, d must be estimated. Suppose that $d = .3$ can be expected for the sociopathy difference between arsonists and burglars. Now, delta can be found from Formula 6.4:

$$\delta = \sqrt{\frac{N}{2}} \quad d = \sqrt{\frac{15}{2}} \cdot .3 = (2.74) (.3) = .82$$

Finally, an alpha level must be selected in order to use Table A.5. Assuming alpha = .05, two-tailed, a delta of .8 (rounding off) corresponds to a power level of .13. This is a very low power level, and it would certainly not be worth running

the experiment. If d were really only .3, the sample sizes being used would only produce significant results about 13 times in 100 experiments.

If the lowest acceptable power were judged to be .7, we can see from Table A.5 that we need delta to be about 2.48 (a delta of 2.5 corresponds to power = .71 in the table). Using the sample size we are stuck with in this problem, we can solve Formula 6.4 for d in order to find the minimal effect size.

$$d = \sqrt{\frac{2}{N}} (\delta) = \sqrt{\frac{2}{15}} (2.48) = (.365)(2.48) = .91$$

The d calculated, .91, is the lowest d that will yield acceptable power with the available sample sizes (at the chosen alpha level). If there is little chance that the sociopathy difference could lead to an effect size this large, then there is little point to conducting the experiment as planned.

However, if it is possible to match the arsonists and burglars into pairs on some basis (e.g., the risk involved in their particular crime), power can be increased. Given that you are stuck with $n = 15$, and $d = .3$, we can calculate how high the correlation would have to be between the two groups of subjects to attain power = .7. We have already calculated delta for this situation without matching; $\delta_{ind}^* = .82$. The delta required for power = .7 (.05, two-tailed) is 2.48. We can find the necessary degree of population correlation needed to attain this delta ($\delta_{matched}^*$) by using the equation on page 133 of the Essentials text.

$$\delta_{matched} = \sqrt{\frac{1}{1 - \rho}} \delta_{ind} \text{ so, } \sqrt{\frac{1}{1 - \rho}} = \frac{\delta_{matched}}{\delta_{indep}} = \frac{2.48}{.82} = 3.024$$

Therefore, $1 / (1 - D) = 3.024^2 = 9.147$, so $D = 1 - (1 / 9.147) = 1 - .11 = .89$. The correlation produced by this matching would have to be as high as .89 in the population in order to guarantee adequate power. Such a high degree of matching in this type of experiment would be very difficult to achieve, but a moderate amount of matching would be helpful in reducing the number of subjects needed for adequate power when d is only .3.

2. Another psychologist is studying the difference between men and women in recalling the colors of objects briefly seen. A very large number of college students of both genders is readily available for testing. How many subjects should be used? First, we need to decide on a level of power. Let us say that power = .85 is desired. From Table A.5, it can be seen that this level of power for a test involving alpha = .05, two-tailed, corresponds to delta = 3.0. Next, d must be estimated. Assume that from previous experiments, d is expected to be of medium size, so that $d = .5$. Finally we apply Formula 6.5:

$$N = 2 \left(\frac{\delta}{d} \right)^2 = 2 \left(\frac{3.0}{.5} \right)^2 = (2) 6^2 = (2) (36) = 72$$

Seventy-two males and seventy-two females are required in order to have power equal to .85, with an effect size that equals .5.

If we have no estimate for d , another strategy is to decide on the smallest d worth testing. For the present example, it might be decided that if the effect size is less than .1, there is no point to finding statistical significance. In this case, any effect size less than .1 might be considered about as unimportant as an effect size of zero. If power = .85 is desired (and therefore delta is still 3.0) for $d = .1$, the number of subjects is again found from Formula 6.5:

$$N = 2 \left(\frac{3.0}{.1} \right)^2 = (2) (30)^2 = (2) (900) = 1800$$

A total of 1800 males and 1800 females would be required to attain a power

equal to .85 if d were only equal to .1. It is not likely that so many subjects would be used for such an experiment, but the above calculation demonstrates that there would certainly be no point to using more than that number of subjects. Using more subjects would lead to a high level of power for effect sizes so small, that they would be considered trivial.

Finally, one could ask: What is the largest that d might be for the difference in color memory? Suppose that the answer is $d = .7$. Assuming the same desired level of power as above, the required number of subjects would be:

$$N = 2 \left(\frac{3.0}{.7} \right)^2 = (2) (18.4) = 36.7$$

The above calculation tells us that there would be no point in using fewer than 37 subjects in each of the two groups. Using fewer subjects would mean that the power would be less than desirable even for the largest effect size expected, and therefore less than desirable for the somewhat smaller effect sizes that are likely to be true.

When dealing with small samples, Table A.6 gives more accurate estimates. For instance, suppose that d for a two-group experiment is large - i.e., .8. If we take the ANOVA approach, the equivalent f is $.8 / 2 = .4$. With only 16 participants in each of the two groups, $N = f^2 m = .4^2 * 4 = 1.6$. Entering the $k = 2$ section of Table A.6 with $df_w = 30$ (i.e., $16 + 16 - 2$), we see that a N of 1.6 corresponds to a power of .59. Using Table A.5 instead, $*$ is $.8 * \sqrt{8} = 2.26$, so power can be estimated to be about .61 or .62 (which is a slight overestimate compared to the result from Table A.6).

Definitions of Key Terms

Beta (β): The probability of making a Type II error (i.e., accepting the null hypothesis, when the null hypothesis is not true).

Power: The probability of rejecting the null hypothesis, when the null hypothesis is not true. Power equals $1 - \beta$.

Alternative Hypothesis Distribution (AHD): The distribution of the test statistic when the null is not true. When dealing with one or two sample means, the AHD is usually a noncentral t distribution (i.e., it is shaped like the t distribution, but it is centered on some value other than zero).

Delta ($*$): The t value that can be expected for a particular alternative hypothesis. Delta, referred to as the "expected t" in this text, is more formally known as the noncentrality parameter, because it is the value upon which a noncentral t distribution is centered.

Effect Size or d : The separation of two population means in terms of standard deviations (assuming homogeneity of variance). The effect size determines the amount of overlap between two population distributions.

Chapter 7

To review the calculation of the two-way ANOVA, we have created a hypothetical experiment as follows. A clinical psychologist wants to know whether the number of sessions per week has an impact on the effectiveness of psychotherapy, and whether this impact depends on the type of therapy. It was decided that therapy effectiveness would be measured by an improvement rating provided by judges who are blind to the conditions of the experiment. Four levels are chosen for the first factor: one, two, three, or four sessions per week; and two types of therapy for the second factor: classical psychoanalysis, and client-centered therapy. When these two factors are completely-crossed, eight cells are formed. For this example there are five randomly-selected subjects in each cell, and the subjects in each cell are selected independently of the subjects in any other cell. The improvement rating for each subject is presented in Table 7.1, along with column, row, and cell means.

Table 7.1

	# of Sessions Per Week				
	1	2	3	4	Row Means
Psychoanalytic	4 6 3 1 4	2 4 5 3 5	6 5 8 6 5	8 6 6 8 5	
Cell Means	3.6	3.8	6.0	6.6	5.0
Client-Centered	6 3 4 4 3	4 3 6 6 4	4 7 2 4 3	5 4 4 6 4	
Cell Means	4.0	4.6	4.0	4.6	4.3
Column Means	3.8	4.2	5.0	5.6	4.65

The simplest strategy is to begin by calculating SS_{total} and $SS_{bet-cell}$ and then subtract to find SS_w . We can use Formula 7.1 for both of these. Note that N_T equals $nrc = 5 * 2 * 4 = 40$.

$$SS_{total} = N_T \sigma^2(\text{all scores}) = 40 (2.6275) = 105.1$$

$$SS_{bet-cell} = N_T \sigma^2(\text{cell means}) = 40 (1.0375) = 41.5$$

Therefore, $SS_w = 105.1 - 41.5 = 63.6$. You can also verify this by calculating the unbiased variance for each cell, averaging these 8 cell variances to find MS_w , and then multiplying by df_w . (If you didn't have the raw data, but you did have the SD's, you would have to square each one to get the cell variances.) Next, we use Formula 7.1 twice more to calculate the SS's that correspond to the sessions and type of therapy factors. In Table 7.1, type of therapy is represented by the rows, and sessions by the columns.

$$SS_{rows} = N_T \sigma^2(\text{row means}) = 40 * \sigma^2(5.0, 4.3) = 40 (.1225) = 4.9$$

$$SS_{col} = N_T \sigma^2(\text{column means}) = 40 (.4875) = 19.5$$

As usual we find the SS for interaction by subtracting SS_{rows} and $SS_{columns}$ from $SS_{bet-cell}$: $SS_{inter} = SS_{bet-cell} - SS_{therapy} - SS_{sessions} = 41.5 - 4.9 - 19.5 = 17.1$.

Having analyzed the total SS into its components, each SS is divided by the appropriate df. The total df for this example is broken down into the following components:

$$\begin{aligned} df_w &= N_T - rc = 40 - (2)(4) = 40 - 8 = 32 \\ df_{therapy} &= r - 1 = 2 - 1 = 1 \\ df_{sessions} &= c - 1 = 4 - 1 = 3 \\ df_{inter} &= (r - 1)(c - 1) = (1)(3) = 3 \end{aligned}$$

Notice that these df's add up to df_{total} which equals $nrc-1 = (5)(2)(4) = 40 - 1 = 39$. Now we can find the MS's by dividing each SS by its df.

$$\begin{aligned} MS_w &= SS_w / df_w = 63.6 / 32 = 1.99 \\ MS_{therapy} &= SS_{therapy} / df_{therapy} = 4.9 / 1 = 4.9 \\ MS_{sessions} &= SS_{sessions} / df_{sessions} = 19.5 / 3 = 6.5 \\ MS_{inter} &= SS_{inter} / df_{inter} = 17.1 / 3 = 5.7 \end{aligned}$$

Finally, we can find the three F ratios as follows:

$$\begin{aligned}F_{\text{therapy}} &= MS_{\text{therapy}}/MS_w = 4.9/1.99 = 2.46 \\F_{\text{sessions}} &= MS_{\text{sessions}}/MS_w = 6.5/1.99 = 3.27 \\F_{\text{inter}} &= MS_{\text{inter}}/MS_w = 5.7/1.99 = 2.87\end{aligned}$$

The critical F ($\alpha = .05$) for testing the therapy effect is based on 1 and 32 degrees of freedom, and equals about 4.15 (see Table A.3). The critical F's for the sessions effect, and the interaction are both based on 3 and 32 df, and therefore equal about 2.90. The F for the main effect of therapy (2.46) is not greater than its critical F (4.15), so the null hypothesis cannot be rejected in this case. On the other hand, the F for the main effect of sessions (3.27) is greater than its critical value (2.9), so for this factor the null hypothesis can be rejected. The interaction has fallen only slightly short of statistical significance; the calculated F (2.87) is just under the critical F.

Interpreting the results

First, we check the significance of the interaction as an indication of whether the main effects will be interpretable, and of which approach should be taken to follow up the ANOVA with additional comparisons. Although the interaction is not significant at the .05 level, it is so close to significance ($p = .052$) that caution is advised in the interpretation of the main effects. If you graph the cell means with sessions along the X axis, you will see that the lines for the two types of therapy cross each other. Whereas client-centered therapy produces slightly more improvement than psychoanalysis when therapy is given one or two times per week, this trend reverses direction when there are three or four sessions. The significant main effect of sessions is clearly due to the psychoanalytic conditions, which show a strong linear trend as the number of sessions increases. If it is surprising that the interaction falls short of significance, bear in mind that the sample sizes are quite small, providing little power unless the amount of interaction is rather large.

Following up the results

If you were to ignore the nearly significant interaction, you would look for significance among the main effects. Had the main effect of therapy type been significant, there would still be no follow up for this effect, simply because it has only two levels. However, the main effect of sessions was significant, and could be followed up with Tukey's HSD (recommended because sessions has more than three levels). First we look up the critical q in Table A.4 (number of groups = 4, and df for the error term is df_w from the two-way ANOVA, which is 32). Then, we apply Formula 5.12, but be careful to note that "n", when we are comparing levels of a main effect, is not the cell size -- rather it is the number of cases at each level of the main effect being tested (n for the sessions effect is 10, because there are 5 participants in each therapy condition for any one level of the sessions effect).

$$HSD = q_{crit} \sqrt{\frac{MS_w}{n}} = 3.84 \sqrt{\frac{1.99}{10}} = 3.84 (.446) = 1.713$$

By Tukey's HSD test, the four-session level differs significantly from the one-session level, but no other pairs of levels differ significantly.

If you were to treat the interaction as significant (not a bad idea when $p = .052$), a reasonable follow-up plan would involve testing simple main effects: conducting a one-way ANOVA over sessions, separately for each type of therapy (but using MS_w from the two-way ANOVA in each case), and/or comparing the two types of therapy at each number of sessions. The simple main effect of sessions is significant for psychoanalytic therapy ($F = 5.6$), but nowhere near significance for client-centered therapy ($F = .32$). The final step in this plan would be to use HSD to compare pairs of session levels, but only for the psychoanalytic condition.

Another reasonable way to follow-up a significant interaction is with 2 X 2 interaction contrasts. There are six possible 2 X 2 contrasts to test corresponding to the six possible pairs for the session levels -- always crossed with the two therapy types. An obvious 2 X 2 to try is one versus four sessions per week for the

two therapies. Using the appropriate cell means from Table 7.1, we can see that L for this contrast is: $3.6 - 4.0 - (6.6 - 4.6) = -.4 - 2.0 = -2.4$. Noting that n in Formula 5.13 is the cell size (5), and that each of the four c's equals +1 after being squared, we find that the contrast yields:

$$SS_{contrast} = \frac{nL^2}{\sum C_i^2} = \frac{5 (-2.4)^2}{4} = \frac{28.8}{4} = 7.2$$

Because this contrast (like all such contrasts) has only one df, we can divide 7.2 by 1.99 (according to Formula 5.14) to obtain the F ratio, which therefore equals 3.62. This is less than the critical value (about 4.15), so this contrast is not significant at the .05 level. If instead we had planned to test two versus three sessions per week, L would be: $3.8 - 4.6 - (6.0 - 4.0) = -.8 - 2.0 = -2.8$, and $SS_{contrast}$ would equal $5 (-2.8)^2 / 4 = 39.2 / 4 = 9.8$. When divided by 1.99, the resulting F ratio, 4.92, is larger than 4.15, so this particular contrast would be statistically significant. But only if it had been planned. If the contrast were discovered after looking through the data to find something promising, the accepted procedure would be to conduct Scheffé's test. For this example, you would take the critical F for testing the entire two-way interaction (about 2.9), and multiply it by df_{inter} (which equals 3) to get Scheffé's critical F (F_s), which equals $2.9 * 3 = 8.7$. So even the two versus three session contrast would not be significant as a post hoc complex comparison. This is not surprising. Given that the interaction from the omnibus two-way ANOVA was not significant, we know that none of the 2 X 2 contrasts will be significant by Scheffé's test.

Finally, although the main effect of sessions was significant, it was only barely significant. We could have attained a much higher F ratio if we took advantage of the obvious upward trend over sessions. The appropriate coefficients for testing an upward linear trend with four levels are: -3, -1, +1, +3 (note: -2, -1, +1, +2 won't work because they are not equally spaced). Applied to the column means of Table 7.1, the L for the linear trend is: $(-3)(3.8) + (-1)(4.2) + (+1)(5) + (+3)(5.6) = -11.4 - 4.2 + 5 + 16.8 = 21.8 - 15.6 = 6.2$. The SS for the linear contrast is found by using Formula 5.13 again, noting that this time n equals 10 (because there are ten cases per column), and the sum of the squared coefficients is 20 ($9+1+1+9$).

$$SS_{linear} = \frac{10 (6.2)^2}{20} = \frac{384.4}{20} = 19.22$$

The F ratio for the linear trend is $19.22 / 1.99 = 9.66$, which is much larger than the F found for the sessions main effect (3.27). It is even significant by Scheffé's test ($9.66 > 8.7$; F_s is the same for both the interaction and the main effect of sessions, because the df's are the same).

The Advantages of the Two-way ANOVA

1. Exploration of the interaction of two independent variables.
2. Economy -- if interaction not significant.
3. Greater generalization of results -- if interaction not significant.
4. Reduction in MS_w (i.e., error term), when the second factor is a grouping variable that differs on the DV.

Definitions of Key Terms

Two-way (independent-groups) ANOVA: A statistical procedure in which independent groups of observations (usually on individual subjects) differ on two independent variables. The results of only one dependent variable can be analyzed at a time (to analyze several dependent variables simultaneously, multivariate procedures, such as MANOVA, are required).

Factor: An independent variable with at least two different levels; it may involve conditions created by the experimenter or it may involve pre-existing differences among subjects.

Completely-crossed factorial design: Each and every level of the first factor is combined with each and every level of the second factor to form a cell (this definition can be generalized to include any number of factors), and there is at least one observation in every cell. More often, this design is just called a "factorial" design.

Balanced design: All of the cells in the factorial design contain the same number of observations.

Marginal mean: The mean for any one level of one factor, averaging across all the levels of the other factor. If a two-way design is looked at as a matrix, the marginal means are the means of each row and each column.

Interaction: The variability among cell means not accounted for by variability among the row means or the column means. If the effects of one factor change at different levels of the other factor, some amount of interaction is present (however, some interaction is expected just from chance factors). If a graph of the cell means produces parallel lines, demonstrating that the effects of the two factors are simply additive, then there is no interaction in your data at all.

Ordinal Interaction: The effects of one factor change in amount, but not direction, with different levels of the other factor. In a graph of cell means, the lines would differ in angles, but slant in the same direction, and not cross. Despite the presence of this type of interaction, the main effects may still be interpretable.

Disordinal Interaction: The direction (or order) of the effect for one factor changes for different levels of the other factor. In a graph of cell means, the lines would either cross, or slant in different directions. This type of interaction usually renders the main effects meaningless.

Degrees of Freedom Tree: A diagram which shows the number of degrees of freedom associated with each component of variation in an ANOVA design.

Simple effects: A simple effect is the effect of one factor at only one level of the other factor. In a two-way design, any one row or column represents a simple effect.

Interaction contrasts: This is a 2 x 2 comparison that is a subset of a larger design. When the interaction is significant for the full design, it may be appropriate to test the interaction in the various 2 x 2 subsets to localize the effect.

Chapter 8

One-way RM ANOVA

The same RM ANOVA procedures that are applied when each subject is measured several times are used to analyze the results of a randomized-blocks (RB) design, as well. In the example below we will analyze the results of an RB design using Formula 7.1. Imagine that a consumer organization wants to answer the question: Are all weight-loss programs equally effective? The four best-known programs are chosen for comparison. Subjects are matched into blocks of four based on having the same initial weight, similar frame, and similar dieting history. Within each block, the four subjects are randomly assigned to the different weight-loss programs. In this hypothetical situation, the researchers may actually believe or wish the null hypothesis to be true. For this example, there are five blocks of four subjects each, and the subjects within each block are assigned at random to the four weight-loss programs. After three months of dieting the number of pounds lost by each subject is recorded, as appears in Table 8.1.

Table 8.1

Block #	Weight Loss Program				Row Means
	I	II	III	IV	
1	2	4	5	1	3.0
2	6	3	7	5	5.25
3	7	9	6	7	7.25
4	5	8	8	6	6.75
5	10	10	13	8	10.25
Col Means	6.0	6.8	7.8	5.4	6.5

The first step is to calculate SS_{total} . Note that the total number of values (N_T) = 20. After entering all 20 values, you can use Formula 7.1.

$$SS_{total} = N_T \sigma^2(\text{all scores}) = 20 (7.85) = 157$$

When you find the biased variance of all the scores, check that the grand mean is equal to 6.5. SS_{RM} is found by inserting the column means in Table 8.1 into Formula 7.1:

$$SS_{RM} = 20 * \sigma^2(6.0, 6.8, 7.8, 5.4) = 20 (.81) = 16.2$$

Check again that the mean of the column means equals the grand mean, 6.5. Formula 7.1 is used one more time to find $SS_{subject}$.

$$SS_{subject} = 20 * \sigma^2(3.0, 5.25, 7.25, 6.75, 10.25) = 20 (5.7) = 114$$

The last SS component is found by subtraction:

$$SS_{sub \times RM} = SS_{total} - SS_{subject} - SS_{RM} = 157 - 114 - 16.2 = 26.8$$

$df_{RM} = c - 1 = 4 - 1 = 3$; $df_{sub \times RM} = (n - 1)(c - 1) = 4 * 3 = 12$ (note that in the RB design, n equals the number of blocks, rather than the actual number of subjects, which is N_T in the one-way case).

The MS's that we need for the F ratio are as follows:

$$MS_{RM} = \frac{SS_{RM}}{df_{RM}} = \frac{16.2}{3} = 5.4, \text{ and } MS_{sub \times RM} = \frac{SS_{sub \times RM}}{df_{sub \times RM}} = \frac{26.8}{12} = 2.23$$

Finally, the F ratio equals $MS_{RM} / MS_{sub \times RM} = 5.4 / 2.23 = 2.42$.

The critical F for (3, 12) df and $\alpha = .05$ is 3.49. Because our calculated F (2.42) is less than the critical F, we cannot reject the null hypothesis. We cannot conclude that there are any differences in weight-loss effectiveness among the four programs tested. Even with reasonably good matching, our sample was too small to give us much power, unless the weight-loss programs actually differ a great deal. In a real experiment, a much larger sample would be used; otherwise the experimenters would have to be very cautious in making any conclusions from a lack of significant results. If the F ratio had been significant, we would have checked it against a conservative critical value derived from multiplying the df's by the lower-bound of epsilon, which is $1 / (c - 1) = 1 / 3 = .333$, in this case. The conservative value is $F(1, 4) = 7.71$, so if the F ratio had landed between 3.49 and 7.71, we would

have needed a computer to calculate a more exact epsilon to determine significance.

Summary Table

The components of the analysis performed above can be presented in the form of a summary table, as shown below:

Table 8.2

Source	SS	df	MS	F	p
Between-Subjects	114	4			
Within-Subjects					
Between-Treatments	16.2	3	5.4	2.42	> .05
Interaction (Residual)	26.8	12	2.23		
Total	157	19			

Assumptions of the RM ANOVA

1. Random sampling. In an RM design, subjects should all be selected independently. In an RB design, subjects in different blocks should be independent, but subjects within a block should be matched.
2. Normal population distributions.
3. Sphericity (occasionally called circularity). The amount of interaction (or the variability of the difference scores) between any two levels should be the same as for any other pair of levels.

When to Use the RM ANOVA

Repeated Measures. If feasible this design is the most desirable, as it generally has more power than independent groups or randomized blocks. For obvious reasons, it can only be used with an IV that involves experimental manipulations, and not a grouping variable. The two types are simultaneous and successive.

1. Simultaneous: The different conditions are presented as part of the same stimulus (e.g., different types of pictures in a collage), are as randomly-mixed trials (e.g., trials of different difficulty levels mixed together).
2. Successive: Usually requires counterbalancing to avoid simple order effects, and is not valid if there are differential carry-over effects.

Randomized Blocks. This design has much of the power of the RM design, and avoids the possible carry-over effects of a successive RM design. The two types are experimental and naturally-occurring.

1. Experimental: In the simplest RB design, the number of subjects per block is the same as the number of experimental conditions; the subjects in each block are assigned to the conditions at random.
2. Naturally-Occurring: The drawback to this design is that you cannot conclude that your independent variable caused the differences in your dependent variable.

Mixed-Design ANOVA

In this section we will present a mixed design example in which both variables involve experimental manipulations. Our example is based on a hypothetical school that is comparing three methods for teaching the sixth grade: traditional, computer (in the classroom), and home (using computers). At the end of the school year, each pupil is given final exams in four subject areas: math, English, science, and social studies. To simplify the analysis three equal-sized samples will be drawn and randomly assigned to the three teaching methods. The highest possible score on each of the final exams is 30; because each pupil takes four final exams, there are a total of $9 \times 4 = 36$ scores. The data are shown in the Table 8.3.

Table 8.3

Method	Math	English	Science	Social St.	Row Means
	15	26	20	17	19.5
Traditional	10	23	16	12	15.25
	5	18	4	1	7
(Cell Means)	10	22.33	13.33	10	13.917
	27	25	26	28	26.5
Computer	18	20	20	23	20.25
	16	17	12	10	13.75
(Cell Means)	20.33	20.67	19.33	20.33	20.167
	25	20	23	27	23.75
Home	22	15	19	25	20.25
	17	9	10	14	12.5
(Cell Means)	21.33	14.67	17.33	22	18.833
Column Means	17.22	19.22	16.67	17.44	17.64

We will begin by finding the number of degrees of freedom for each component in the analysis (see Rapid Reference 8.3). For this example, k (number of groups - in this example, methods) = 3, c (number of repeated measures - in this example, subject areas) = 4, and N_s , or just n (number of subjects per group) = 3. Therefore,

$$\begin{aligned}
 df_{total} &= nkc - 1 = 3*3*4 - 1 = 36 - 1 = 35 \\
 df_{subject} &= nk - 1 = 3*3 - 1 = 9 - 1 = 8 \\
 df_{method} &= k - 1 = 3 - 1 = 2 \\
 df_w &= k(n - 1) = 3*2 = 6 \\
 df_{area} &= c - 1 = 4 - 1 = 3 \\
 df_{inter} &= (k - 1)(c - 1) = 2*3 = 6 \\
 df_{sub \times RM} &= k(c - 1)(n - 1) = 3*3*2 = 18
 \end{aligned}$$

The total SS is found by entering all 36 scores from Table 8.3 into a calculator, and then using Formula 7.1.

$$SS_{total} = N_T \sigma^2(\text{all scores}) = 36 (46.786) = 1684.3$$

Proceeding as with any two-way ANOVA, we find $SS_{\text{between-cells}}$, the SS's for the two main effects, and then subtract to find the SS for the interaction of the two factors. We will use Formula 7.1 three times.

$$SS_{\text{between-cells}} = N_T \sigma^2(\text{cell means}) = 36 (18.78) = 676.1$$

$$SS_{\text{Method}} = 36 * \sigma^2(13.917, 20.167, 18.833) = 36 (7.224) = 260$$

$$SS_{\text{Area}} = 36 * \sigma^2(17.22, 19.22, 16.67, 17.44) = 36 (.916) = 33$$

$$SS_{\text{inter}} = SS_{\text{bet-cell}} - SS_{\text{method}} - SS_{\text{area}} = 676.1 - 260 - 33 = 383.1.$$

The corresponding MS's are as follows: $MS_{\text{method}} = 260 / 2 = 130$; $MS_{\text{area}} = 33 / 3 = 11$; $MS_{\text{inter}} = 383.1 / 6 = 63.85$. To find the two different error terms that we need to complete the analysis, we begin by calculating the subject-to-subject variability, based on the subject means (averaging across the four subject areas).

$$SS_{\text{subject}} = 36 * \sigma^2(\text{the nine subject means}) = 36 (32.6) = 1173.6$$

Subtracting SS_{method} from SS_{subject} leaves SS_{w} , the error term for the between-subjects part of the analysis. $SS_{\text{w}} = SS_{\text{subject}} - SS_{\text{method}} = 1173.6 - 260 = 913.6$. Therefore, $MS_{\text{w}} = 913.6 / 6 = 152.3$, and $F_{\text{method}} = 130 / 152.3 = .85$.

Next, we find the error term for the within-subjects part of the analysis. $SS_{\text{Sub} \times \text{RM}}$ is found by subtracting $SS_{\text{between-cells}}$ from SS_{total} to obtain $SS_{\text{within-cells}}$, and then subtracting SS_{w} from that: $SS_{\text{Sub} \times \text{RM}} = 1684.3 - 676.1 - 913.6 = 94.6$. Therefore, $MS_{\text{Sub} \times \text{RM}} = 94.6 / 18 = 5.26$.

Finally, we can form the F ratios to test each of the within-subjects effects: the main effect of subject area, and the interaction of the two factors:

$$F_{\text{area}} = \frac{MS_{\text{area}}}{MS_{\text{Sub} \times \text{RM}}} = \frac{11}{5.26} = 2.09; F_{\text{inter}} = \frac{MS_{\text{inter}}}{MS_{\text{Sub} \times \text{RM}}} = \frac{63.85}{5.26} = 12.14$$

To test the main effect of method, we are looking for a critical F based on df_{method} and df_{w} : $F(2, 6) = 5.14$. For the main effect of subject area we need F ($df_{\text{area}}, df_{\text{sub} \times \text{RM}}$) = $F(3, 18) = 3.16$. For the interaction, we are looking for F ($df_{\text{inter}}, df_{\text{sub} \times \text{RM}}$) = $F(6, 18) = 2.66$.

Because the F for the main effect of method is less than 1.0, we do not even need to compare it to its critical value to know that it is not significant at the .05 level. The F for the main effect of the subject area is much better, but also falls short of significance ($2.09 < 3.16$) also falls short of its critical value, so this null hypothesis cannot be rejected, either. The F observed for the interaction of the two factors (12.2), however, is greater than its critical value (2.66), so this effect is statistically significant. Sphericity is not an issue for the main effect of subject area, because this effect is not significant. However, a lack of sphericity in the population could have inflated the F for the interaction, so to be cautious, we can check to see whether this F is still significant when we assume the worst case for lack of sphericity. The value for the lower bound of epsilon is 1/3, so the conservative critical F is $F(2, 6) = 5.14$. Because $12.14 > 5.14$, we don't have to worry about the sphericity assumption in declaring the interaction to be significant at the .05 level.

Interpreting the Results

The significant interaction suggests that the different teaching methods do make some difference in how much students learn, but that this difference is not uniform among the different subject areas. That is, some subject areas benefit more than others from a particular teaching method, and which subject area is benefitted more depends on which teaching method the pupils receive. For instance, by inspecting Table 8.3 you can see that English benefits the most from the traditional method, but benefits the least from learning on a home computer. A graph of the cell means makes it easy to see that subject area makes little difference when the computer is used, but a good deal of difference for the traditional method. An appropriate way to follow up the significant interaction in this case is with the analysis of simple effects. A one-way RM ANOVA could be conducted separately for each method. However, it would be more meaningful in this case, to test the effect of method separately for each subject area, followed by pairwise comparisons of the methods, for whichever subject areas yield a significant independent-groups ANOVA. The results of a mixed two-way ANOVA are often displayed as in Table 8.4, with the between-subjects components separated from the within-subjects components (SPSS presents the within-subjects effects first).

Summary Table for the Mixed ANOVA

Table 8.4

Source	SS	df	MS	F	p
Between-Subjects	1173.6	8			
Methods	260	2	130	.85	>.05
Within-groups	913.6	6	152.3		
Within-Subjects	510.7	27			
Area	33	3	11	2.09	>.05
Method X Area	383.1	6	63.85	12.14	<.05
Residual (S X Area)	94.6	18	5.26		
Total	1684.3	35			

Assumptions of the Mixed Design ANOVA

Because the mixed design involves performing an independent samples ANOVA and a repeated measures ANOVA, the assumptions underlying both of these statistical procedures are required, as appropriate.

1. Independent Random Sampling.

2. Normal Distributions.

3. Sphericity within RM factor. Same as the sphericity assumption for the one-way RM ANOVA.

4. Homogeneity between groups. The amount of subject by treatment interaction for any pair of levels in one group is the same as the amount of interaction for that pair of levels in any other group.

Definitions of Key Terms

One-way repeated measures ANOVA: ANOVA with one independent variable, of which all the levels are presented to each subject (or a block of matched subjects).

Repeated measures (RM) design: An experiment in which each subject receives all levels of the independent variable.

Randomized blocks (RB) design: An experiment in which subjects are matched in groups (called blocks), and then randomly assigned to the different levels of the independent variable. If the number of subjects per block equals the number of levels of the IV, then the RB design is analyzed in exactly the same way as an RM design.

Order effects: In the simplest case, treatment levels receive an advantage or disadvantage depending only on the ordinal position of the treatment, and not depending on which treatment level or levels were presented earlier. Such order effects are called simple order effects; they can be caused by practice or fatigue, and can be averaged out by counterbalancing.

Latin-Square design: A system for counterbalancing in which the number of different sequences of treatment levels equals the number of treatment levels. Each treatment level appears once in each ordinal position; in addition, it can be arranged that each treatment level is preceded by each other treatment level only once.

Carry-over effects: The effect of a treatment level can depend, to some extent, on the particular treatment levels that preceded it. When these effects are not symmetric (e.g., the effect of level B may be very different when preceded by level A, while the effect of level A may be the same whether or not it is preceded by level B), these effects are often called differential carry-over effects. Differential carry-over effects cannot be averaged out by counterbalancing, and if

they cannot be eliminated by changing the experimental procedure, an RB or independent groups design should be used.

Sphericity: This condition, which is also called circularity, applies to the population when the amount of interaction (or the variance of the difference scores) for any pair of levels is the same as for any other pair of levels in the study. If this condition does not apply, the usual RM ANOVA procedure will lead to a higher rate of type I errors than the alpha level that has been set.

Conservative correction: A very conservative approach to the RM ANOVA is to assume that there is the maximum amount of heterogeneity of covariance possible in the population, and to adjust the degrees of freedom accordingly before looking up the critical F. If the observed F ratio falls between the usual critical F and the maximally adjusted critical F, a more precise estimate of the degree of homogeneity of covariance is called for.

Mixed (or Split-Plot) Design: An experimental design which contains one or more between-subjects factors along with one or more within-subjects factors.

Between-Subjects factor: An independent variable for which each subject participates at only one level (also called a between-groups factor or variable).

Within-Subjects factor: An independent variable for which each subject participates at every level, or subjects are matched across levels (also called a repeated measures factor or variable).

Analysis of Covariance (ANCOVA): A form of ANOVA, in which a concomitant variable (i.e., covariate) that is linearly related to the dependent variable, but not the independent variable is used to remove unwanted variance from the dependent variable and the group means.

Chapter 9

The Sign Test

The Sign test can be used in place of the matched t-test when the amount of difference between the members of a matched pair cannot be determined, but the direction of that difference can be. Because the direction of each difference must fall into one of only two categories ("+" or "-"), the binomial distribution can be used to determine whether a given imbalance between those categories is likely to occur by chance (generally, an equal number of pluses and minuses are expected under the null hypothesis).

Imagine an example in which there are 14 pairs of matched subjects, and 11 of the differences are in the negative direction, whereas only 3 differences are positive. A table of the binomial distribution (for $P = .5$, $N = 14$) would show that the probability for $X = 11$ is .0222. In addition, the probabilities for $X = 12$, 13, and 14 would be needed, and they are .0056, .0009, and .0001, respectively. Summing these probabilities, we obtain: $.0222 + .0056 + .0009 + .0001 = .0288$. Doubling this sum, we get: $.0288 * 2 = .0576$; this is our exact two-tailed p level. In this case, we could not reject the null hypothesis at the .05 level, two-tailed.

The Correction for Continuity

Although we would not ordinarily calculate the normal approximation with such a small N, we will do so below to review the procedure. We will use Formula 9.1 (with the continuity correction):

$$z = \frac{|X - NP| - .5}{\sqrt{NPQ}} = \frac{|11 - 7| - .5}{\sqrt{3.5}} = \frac{4 - .5}{1.87} = \frac{3.5}{1.87} = 1.87$$

The above z-score leads to the same conclusion we reached with the test based directly on the binomial distribution; the null hypothesis cannot be rejected because the calculated z (1.87) is less than the critical z (1.96). If you use Table A.1 to look up the p level (area beyond z) that corresponds to the calculated z-score, you will see that $p = .0307$, which is quite close to the one-tailed p level

(.0288) found from the binomial distribution. Even with a sample size as small as 14, the correction for continuity provides a fairly good normal approximation.

Assumptions of the Sign Test

1. Dichotomous Events. Each simple event or trial can fall into only one or the other of two categories -- not both simultaneously, or some third category. The probabilities of the two categories, P and Q, must sum to 1.0.
2. Independent Events. The outcome of one trial does not influence the outcome of any other trial.
3. Stationary Process. The probabilities of each category (i.e., P & Q) remain the same for all trials in the experiment.

When to Use the Binomial Distribution for Null Hypothesis Testing

1. The Sign Test. The binomial distribution can be used as an alternative to the matched or RM t-test, when it is possible to determine the direction of the difference between paired observations, but not the amount of that difference. The Sign test can be planned (i.e., no attempt is made to measure the amount of difference, only its direction is assessed), or unplanned (e.g., a matched t-test had been planned but the sample size is fairly small, and the difference scores are very far from following a normal distribution).
2. Correlational Research. The binomial distribution applies when this kind of research involves counting the number of individuals in each of two categories within a specified group (e.g., counting the number of smokers and nonsmokers in an intensive cardiac care unit). The values of P and Q are based on estimates of the proportion of each category in the general population (e.g., if it is estimated that 30% of the population smoke, then P = .3 and Q = .7).
3. Experimental Research. The binomial distribution is appropriate when the dependent variable is not quantifiable, but can be categorized as one of two alternatives (e.g., given a choice between two rooms in which to take a test -- both identical except that one is painted red and the other blue -- do equal numbers of subjects choose each one?).

Two-way Chi-Square test

In this example, we deal with a design in which one of the variables is actually manipulated by the experimenter. Imagine that a psychiatrist has been frustrated in her attempts to help chronic schizophrenics. She designs an experiment to test four therapeutic approaches to see if any one treatment is better than the others for improving the lives of her patients. The four treatments are: intense, individual psychodynamic therapy; constant unconditional positive regard and Rogerian therapy; extensive group therapy and social skills training; a token economy system. The dependent variable is the patient's improvement over a six-month period, measured in terms of three categories: became less schizophrenic; became more schizophrenic; or, no discernible change. Eighty schizophrenics meeting certain criteria (not responsive to previous treatment, more than a certain number of years on the ward, etc.) are selected, and then assigned at random to the four treatments, with the constraint that 20 are assigned to each group. After six months of treatment, each patient is rated as having improved, having gotten worse, or having remained the same. The data can be displayed in a 3 X 4 contingency table, as in Table 9.1.

Table 9.1

Observed Frequencies	Psycho-dynamic	Rogerian	Group	Token	Row Sums
Improved	6	4	8	12	30
No Change	6	14	3	5	28
Got Worse	8	2	9	3	22
Column Sums	20	20	20	20	N = 80

First, the expected frequencies must be found for each of the 12 cells in the contingency table. Each f_e can be calculated by multiplying its row sum by its column sum, and then dividing by the total N , which is 80 in this problem. However, sum of the expected frequencies can be found by subtraction. Finding the number of degrees of freedom associated with the above table, tells us how many f_e 's must be calculated.

$$df = (R - 1)(C - 1) = (3 - 1)(4 - 1) = (2)(3) = 6$$

Because $df = 6$, we know that only six of the f_e 's are free to vary; the remaining f_e 's can be found by subtraction (within each row and column, the f_e 's must add up to the same number as the f_o 's). However, if you want to save yourself some calculation effort in this way, you will have to choose the right six cells to calculate, as shown in Table 9.2.

Table 9.2

Expected Frequencies	Psycho-dynamic	Rogierian	Group	Token	Row Sums
Improved	7.5	7.5	7.5		30
No Change	7	7	7		28
Got Worse					22
Column Sums	20	20	20	20	80

To illustrate how the f_e 's were calculated in the above table, we show how we found the f_e for the "Rogierian-No Change" cell.

$$f_e = \frac{(\text{Row Sum})(\text{Column Sum})}{N} = \frac{(28)(20)}{80} = \frac{560}{80} = 7$$

The f_e 's not shown in Table 9.2 can now be found by subtraction. First, find the f_e 's for the "Got Worse" cells in the first three columns (subtract the other two f_e 's from 20). Then, each f_e in the "Token" column can be found by subtracting the other three treatments from each row sum. The 3 X 4 contingency table with the observed frequencies and all of the expected frequencies (in parentheses) is shown in Table 9.3.

Table 9.3

	Psycho-dynamic	Rogierian	Group	Token	Row Sums
Improved	6 (7.5)	4 (7.5)	8 (7.5)	12 (7.5)	30
No Change	8 (7)	2 (7)	9 (7)	3 (7)	22
Got Worse	6 (5.5)	14 (5.5)	3 (5.5)	5 (5.5)	28
Column Sums	20	20	20	20	80

We are now ready to apply Formula 9.2 to the data in Table 9.3.

$$\begin{aligned} \chi^2 &= \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(6-7.5)^2}{7.5} + \frac{(4-7.5)^2}{7.5} + \frac{(8-7.5)^2}{7.5} + \frac{(12-7.5)^2}{7.5} + \frac{(6-7)^2}{7} + \\ &+ \frac{(14-7)^2}{7} + \frac{(3-7)^2}{7} + \frac{(5-7)^2}{7} + \frac{(8-5.5)^2}{5.5} + \frac{(2-5.5)^2}{5.5} + \frac{(9-5.5)^2}{5.5} + \frac{(3-5.5)^2}{5.5} \\ &= .3 + 1.63 + .03 + 2.7 + .143 + 7 + 2.29 + .57 + 1.14 + 2.23 + 2.23 + 1.14 = 21.4 \end{aligned}$$

From Table A.7, we see that the critical value of P^2 for $df = 6$, and $\alpha = .05$, is 12.59. Because the calculated P^2 (21.4) is greater than the critical P^2 , the null hypothesis is rejected. We can conclude that the tendency towards improvement is not independent of the type of treatment; that is, the various treatments differ in their population proportions of improvement, no change, or worsening.

Assumptions of the Chi-Square Test

1. Mutually exclusive and exhaustive categories. Each observation falls into one, and only one, category.
2. Independence of observations. Usually, this assumption is satisfied by having each frequency count represent a different subject (i.e., each subject contributes only one observation).
3. Size of expected frequencies. The rule of thumb is that no f_e should be less than 5. This rule is relaxed somewhat if there are many categories (as long as f_e is never less than 1 and no more than 20% of the f_e 's are less than 5), but becomes more stringent if $df = 1$ (when $df = 1$, f_e should be at least 10).

When to Use the Chi-Square Test for Independence

The uses of the two-variable chi-square tests generally fall into one or another of three types:

1. Two grouping variables -- e.g., proportions of left-handed and right-handed persons in various professions.
2. One grouping variable -- e.g., proportions of babies from different cultures expressing one of several possible emotional reactions to an experimental condition, such as removing a toy.
3. One experimentally-manipulated variable -- e.g., exposing children to violent, neutral, or peaceful cartoons, and categorizing their subsequent behavior as aggressive, neutral, or cooperative.

Tests for Ordinal Data

The Mann-Whitney Test

In this example, we deal with a dependent variable that is measured on a ratio scale (time in seconds), but for which the distribution is not consistent with the use of parametric statistics. In our hypothetical experiment, subjects are asked to find a "hidden" figure (e.g., a drawing of a dog) embedded in a very complex visual stimulus. The amount of time is recorded until they can trace the hidden figure. One group of subjects sees a quick flash of the hidden figure (the flash is so fast it's subliminal) before beginning the task, while the other group sees an equally quick flash of an irrelevant drawing. The amount of time spent by the "primed" group (correct hidden drawing was flashed) can be compared to the "unprimed" group (irrelevant flash) using procedures for ordinal statistics. The amount of time (in seconds) spent finding the figure is shown for each of seven subjects in Table 9.4.

Table 9.4

Primed	Unprimed
12	10
42	38
8	20
160	189
220	225
105	189
22	45

First, we rank the data for both groups combined, giving the average ranks for tied measurements, as shown in Table 9.5. Each rank can be marked with an "A" or a "B" according to which group is associated with that rank. Because the two samples are the same size, it is arbitrary which group is labelled the smaller group ($n_s = n_t = 7$); we will label the primed group (A) as the "smaller" group.

Table 9.5

Time	Rank	Group
8	1	A
10	2	B
12	3	A
20	4	B
22	5	A
38	6	B
42	7	A
45	8	B
105	9	A
160	10	A
189	11.5	B
189	11.5	B
220	13	A
225	14	B

Now we sum the ranks for the primed (A) group. $ER_A = S_s = 1 + 3 + 5 + 7 + 9 + 10 + 13 = 48$. As a check, we find $ER_B = 57$, and note that $ER_A + ER_B = 48 + 57 = 105$, which is the same as the sum of ranks 1 to 14 as found by Formula 9.5:

$$S_R = \frac{N(N+1)}{2} = \frac{14(15)}{2} = \frac{210}{2} = 105$$

The Normal Approximation to the Mann-Whitney Test

The ER_A , which we are labelling S_s , can be converted to a z-score by means of

Formula 9.6. The z-score for the present example is:

$$z = \frac{S_s - .5(n_s)(N+1)}{\sqrt{\frac{n_s n_L (N+1)}{12}}} = \frac{48 - .5(7)(15)}{\sqrt{\frac{(7)(7)(15)}{12}}} = \frac{48 - 52.5}{\sqrt{61.25}} = \frac{-4.5}{7.83} = -.57$$

Although our sample sizes are too small to justify using the normal approximation, it is not surprising that the z-score is very far from statistical significance, given that the sums of ranks for the two groups are not very different.

The Wilcoxon Test

In order to create a matched-pairs design, let us imagine that the subjects in Table 9.4 were actually paired together based on previous scores on a "field dependency" test. Thus, each row of Table 9.4 represents a pair of subjects that had been matched before being randomly assigned to either the primed or unprimed group. The differences between the two groups can be calculated as shown in Table 9.6:

Table 9.6

Pair #	Primed	Unprimed	Difference
1	12	10	-2
2	42	38	-4
3	8	20	+12
4	160	189	+29
5	220	225	+5
6	105	189	+84
7	22	45	+23

The next step is to rank order the magnitude of the difference scores without regard to their direction (i.e., sign), but next to each rank we indicate the sign of the difference score with which it is associated (see Table 9.7).

Table 9.7

Difference Score	Rank
-2	(-)1
-4	(-)2
+5	(+)3
+12	(+)4
+23	(+)5
+29	(+)6
+84	(+)7

The final step is to sum the negatively and positively associated ranks, separately. From Table 9.7, $ER_{\text{minus}} = 1 + 2 = 3$, and $ER_{\text{plus}} = 3 + 4 + 5 + 6 + 7 = 25$.

The value of T is the smaller of the two sums, so T = 3.

The Normal Approximation to the Wilcoxon Test

Wilcoxon's T can be converted to a z-score by using Formula 9.8, as shown below for the present example.

$$z = \frac{T - .25N(N+1)}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} = \frac{3 - .25(7)(8)}{\sqrt{\frac{7(8)(15)}{24}}} = \frac{3 - 14}{\sqrt{35}} = \frac{-11}{5.92} = -1.86$$

Although the normal approximation would not be used for such a small sample size, the approximation in this case is actually a reasonably good one (as compared to a more exact test). The z-score calculated above is significant at the .05 level for a one-tailed, but not a two-tailed test. If we assume that the two population distributions are similar in form, and we can justify a one-tailed test, we can assert that the median solution time for the primed population is less than for the unprimed population. Note that the Mann-Whitney test failed to even approach a significant difference between the groups, but the extra power gained by matching the subjects led to the (one-tailed) significance of the Wilcoxon test.

The Wilcoxon matched-pairs test has an advantage over the Mann-Whitney test only if the scores are fairly well matched. The degree of matching can be measured by means of the Spearman correlation coefficient, r_s .

Spearman Correlation for Ranked Data

Correlation coefficients are often used to describe the reliability of a test, or the matching between sets of scores in a study, without testing any hypothesis. We will use correlation in this way to test the matching in Table 9.4. Pearson r can be calculated directly for the data in Table 9.4, but we will assume that the distributions of the variables are not compatible with the assumptions underlying the use of Pearson's r. To circumvent these assumptions, we assign ranks to the data separately for each variable, and then apply Pearson's formula to these ranks. That is, we rank the "primed" scores 1 to 7 (giving average ranks to ties), and then rank the unprimed scores 1 to 7. The original data is then replaced by these ranks. In Table 9.8, we have included the original data from Table 9.4, along with the corresponding ranks.

Table 9.8

Primed	Rank	Unprimed	Rank	D	D ²
12	2	10	1	1	1
42	4	38	3	1	1
8	1	20	2	-1	1
160	6	189	5.5	.5	.25
220	7	225	7	0	0
105	5	189	5.5	-.5	.25
22	3	45	4	-1	1
				ED ² =	4.5

Any of the formulas for Pearson's r can be applied directly to the ranks in Table 9.4, and the result would be r_s . However, if calculating by hand, it is easier to compute the difference score for each pair of ranks, square each difference, find the sum of the squared differences (ED²), and insert this sum into

a shortcut formula for Spearman correlation. That is why these differences (D) and squared differences (D^2) have been included in Table 9.8. To find r_s we can take $\sum D^2$ from Table 9.8, and plug it into the following short-cut formula, in which N equals the number of pairs of scores.

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6(4.5)}{7(48)} = 1 - \frac{27}{336} = 1 - .080 = .920$$

As you can see, r_s is very high, which indicates that the pairs of scores were very well matched. This high correlation explains the large discrepancy between the results of the Mann-Whitney test (which ignores the matching), and the Wilcoxon test (which uses the matching).

Assumptions of Tests on Ordinal Data

All three of the tests described in this chapter are "distribution-free" in that none makes any assumption about the shape of the distribution of the dependent variable. Only the following two assumptions are required.

1. Independent random sampling. This is the same assumption that is made for parametric tests.
2. The distribution of the dependent variable is continuous. This implies that tied ranks will be rare. If there are more than a few ties, correction factors may be needed.

When to Use Ordinal Tests

There are two major situations that call for the use of ordinal tests:

1. The dependent variable has been measured on an ordinal scale. In some cases, it is not feasible to measure the DV precisely (e.g., the DV is charisma, or creativity), but it is possible to place participants in order of magnitude on the DV, and assign ranks.
2. The dependent variable has been measured on an interval or ratio scale, but the distribution of the DV does not fit the assumptions for a parametric test. The smaller the samples, the less accurate parametric tests become in this situation. The interval/ratio measurements are assigned ranks before applying ordinal tests.

Definitions of Key Terms

Dichotomous. Consisting of two distinct, mutually exclusive categories (e.g., heads or tails; on or off).

Binomial distribution. When N independent, dichotomous simple events occur, this distribution gives the probability for each value of X, where X is the number of simple events falling into one of the two categories (X ranges from 0 to N).

Sign test. An alternative to the matched t-test in which only the sign of each difference score is used. The probability of a true null hypothesis producing the observed results is derived from the binomial distribution.

Probability. A number that tells you the likelihood of some event occurring, usually expressed as a proportion that goes from zero (no chance of the event occurring) to one (certainty that the event will occur). In terms of a distribution, the probability that an event will be from a particular range of values is given by the area under the curve corresponding to that range of values.

Mutually exclusive. Two events are mutually exclusive if the occurrence of one event is incompatible with the occurrence of the other event. Areas in a distribution that do not overlap represent mutually exclusive events.

Exhaustive. A set of events is exhaustive if all possibilities are represented, such that one of the events must occur. When a coin is flipped, the events "heads" and "tails" are exhaustive (unless you think there is a chance the coin will land and remain on its edge). Heads and tails are also mutually exclusive.

Independent. Two events are independent if the occurrence of one event does not change the probability of the occurrence of the other event. Under ordinary circumstances, two flips of a coin will be independent.

Classical approach to probability. The probability of a complex event is determined by counting the number of outcomes included in that event, and dividing by the total number of possible outcomes (assuming all outcomes are equally likely). This approach is especially appropriate when dealing with games of chance.

Empirical approach to probability. Probabilities are estimated by using sampling, rather than exhaustive counting.

Distribution-free test. An hypothesis test, in which no assumption needs to be made about the distribution of the observed variables. In fact, the dependent variable can be dichotomous, as in the Sign test. Non-parametric tests generally fall in this category.

Chi-square distribution. A continuous, mathematical distribution that is usually positively skewed, but whose exact shape depends on the number of degrees of freedom.

Pearson's chi-square statistic. A measure of the discrepancy between expected and obtained frequencies in a sample. It follows one of the chi-square distributions approximately, when the null hypothesis is true, and certain assumptions are met.

Pearson's chi-square test of association (also test for independence). The expected frequencies are found for a two-way contingency table, based on the marginal sums and the assumption that the two variables are independent. Then Pearson's chi-square statistic is applied to measure the discrepancy between the expected and observed frequencies.

Goodness-of-Fit Test. This term refers to a one-variable chi-square test; the fit is measured between the frequencies of different categories in a sample, and the frequencies of those categories in a real or hypothetical population.

Contingency table. Displays the frequencies of joint occurrence for the categories of two or more variables; in the two-variable case, each category of one variable is divided into all the categories of the other variable, and the joint frequency is shown for each cell of the two-way matrix.

Yates' Correction for Continuity. A half unit (.5) is subtracted from the absolute value of the difference between an expected and an observed frequency to reduce the discrepancy between the distribution of the chi-square statistic (discrete), and the corresponding chi-square distribution (continuous), when there is only one degree of freedom.

Phi Coefficient. A Pearson correlation coefficient that is calculated for two variables that consist of only two values each. Phi (Φ) can be found from P^2 and N .
Cramér's phi. A statistic closely related to phi (e.g., it ranges from 0 to 1) that can be applied to contingency tables larger than 2 X 2.

Mann-Whitney Test. Compares two independent groups when the dependent variable has been measured on an ordinal scale. Also called the Mann-Whitney U test (if the U statistic is computed), or the Wilcoxon-Mann-Whitney test.

Wilcoxon matched-pairs test. Replaces the matched t-test when differences scores are not considered precise, but can be rank-ordered. Also called the Wilcoxon T test, because the test statistic is referred to as T.

Spearman correlation, r_s . This is the result of applying the Pearson correlation formula to two variables when both are in the form of ranks.

Kruskal-Wallis test. This test is a direct extension of the Mann-Whitney test that can accommodate any number of independent groups. Because it replaces the one-way ANOVA when the data are in the form of ranks, it is sometimes called the Kruskal-Wallis one-way analysis of variance by ranks. It is also called the Kruskal-Wallis H test, because the test statistic is referred to as H.

Friedman test. This test replaces the one-way repeated-measures or randomized blocks ANOVA when the data are in the form of ranks.