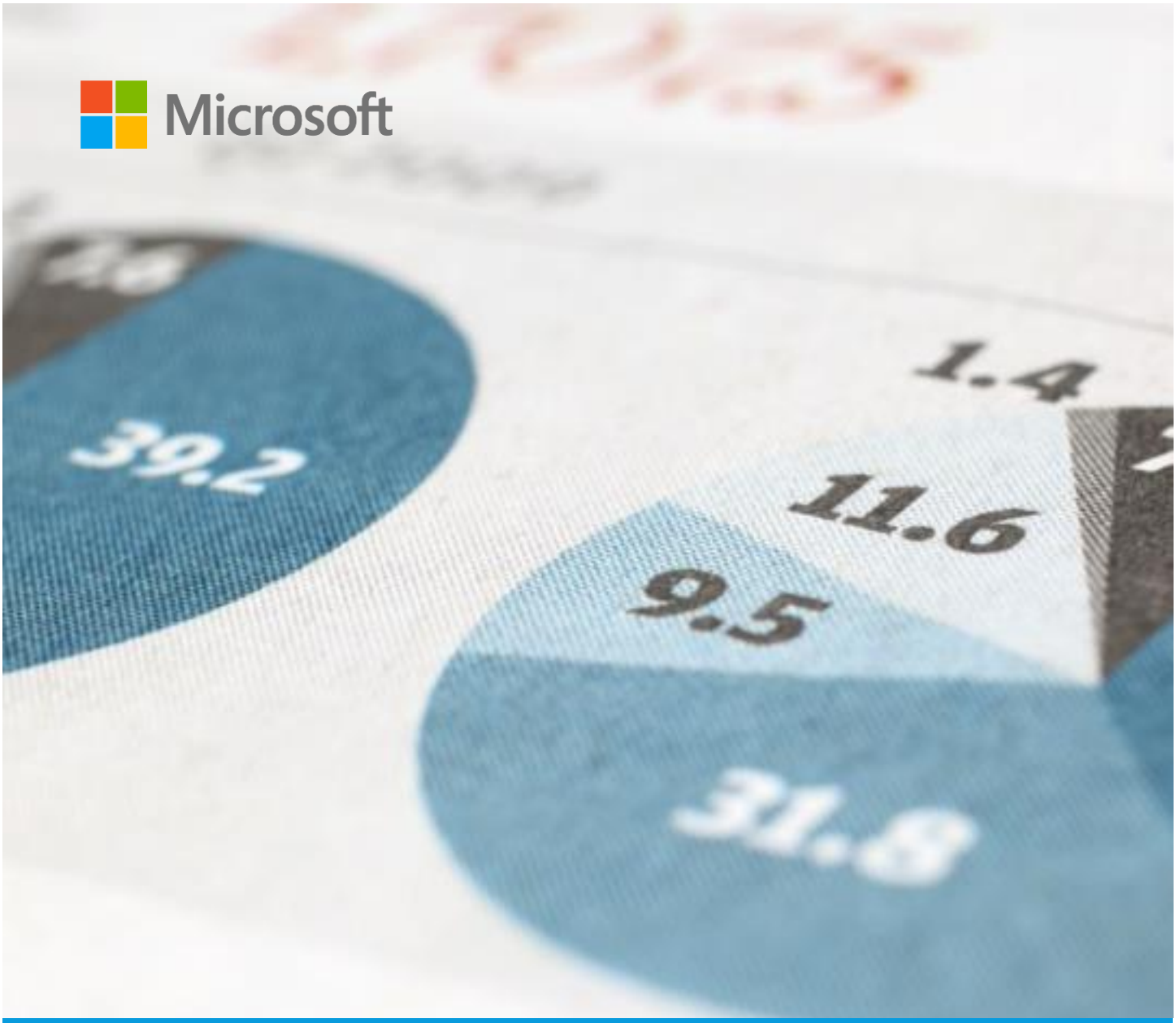


Statistics Primer

A Brief Overview of Basic Statistical and Probability Principles

Liberty J. Munson, PhD

9/19/16

Essential Statistics for Data Analysts Using Excel

Table of Contents

Statistics is fundamentally concerned with summarizing variables in such a way that allows for interpretation. In organizations, these interpretations drive decisions that lead to, ideally, changes and improvements that positively affect business outcomes.

What is a Variable?

A variable is an attribute that can be used to describe a person, place, or thing. In the case of statistics, it is any attribute that can be represented as a number. The numbers used to represent variables fall into two categories:

- **Quantitative variables** are those for which the value has numerical meaning. The value refers to a specific amount of something. The higher the number, more of some attribute the object has. For example, temperature, sales, and number of flyers posted are quantitative variables.
- **Categorical variables** are those for which the value indicates group membership. Thus, you can't say that one group has more/less of something based on the number assigned to it because it's arbitrary. In Rosie's data, location where the drinks are sold is a categorical variable. Gender is a classic example.

Why is this important?

Understanding the type of variable is extremely important because you can only do meaningful mathematical calculations (e.g., means and standard deviations) on quantitative variables. While you can perform these same calculations on categorical variables, the results are must be interpreted with care because the numbers that are used do not have meaning in the same way that quantitative variables do.

For example, to perform statistical analysis comparing genders, you must assign a number to each gender. This is called a dummy code, but it wouldn't make sense to calculate the average of those numbers. Assume you assigned a value of 1 to 'beach' and 2 to 'park' and obtained an average of 1.45. What does that mean? You only know that Rosie spent more time at the beach than at the park because the average is closer to the dummy code assigned to beach, but no other interpretations can be drawn. The best way to describe the categorical values in your data is to report frequencies of each possible group (e.g., 30% of the days were park sales and 70% were beach sales).

Population vs. Sample

When you conduct a study, you define your population of interest. This is the entire set of elements that possess the characteristics of interest. In reality, you will rarely obtain observations or measurements on all elements of interest in a particular study simply because some of them will be inaccessible for a wide variety of reasons or it will be impractical to do so.

A **population** includes all elements of interest.

A **sample** consists of a subset of observations from a population. As a result, multiple samples can be drawn from the same population.

Why is this important?

The nomenclature, statistical formulas, notation, and vary depending on whether your analyzing a population or sample.

In terms of nomenclature, for example, a measurable outcome of a population is called a **parameter**; in a sample, it is called a **statistic**.

In terms of formulas, you will notice some subtle but important differences in the population and sample formulas for variance and standard deviation. In samples, you divide by n-1 because the mean that we use is in this calculation approximates the true mean. Because we are estimating both the mean and standard deviation from the same subset of data, we tend to underestimate of the true variability in the data. This correction results in an unbiased estimate of the variance and standard deviations.

In terms of notation, population parameters are always represented by Greek characters or capital letters whereas sample statistics are always represented by lower case letters.

Some examples:

- σ^2: Population variance
- σ: Population standard deviation
- s^2: Sample variance
- s: Sample standard deviation
- μ: Population mean
- x: Sample mean
- N: Number of observations in the population
- n: Number of observations in the sample

Measures of Central Tendency

Central tendency measures simply provide information on the most typical values in the data for a given variable.

The **mean** represents the average value of the variable, and can be calculated as:

Population:

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

Sample:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where x is each value in the data set for a given variable and n is the total number of observations of that variable.

The **median** is the middle most value for a given variable; 50% of values are above, 50% of responses are below. To find the median, you must order your data from smallest to largest. The result of the formula below gives you the location of the median. Simply count down from the top of your sorted list of values until you reach that location. The value at that location is the median. If you have an odd number of observations, this will be a single value that actually appears in the data; if you have an even number of observations, the median is the average of the value above and below this location.

$$M_d = (n+1)/2$$

The **mode** is the most frequently occurring value for a given variable; if there is more than one mode, report them all. The best way to identify the mode is to plot a histogram.

Measures of Variability

To better understand the shape of a variable's distribution, we use measures of variability.

The simplest measure of variability is the **range** which is simply the minimum value subtracted from the maximum value:

$$\text{Range} = \text{Max (i)} - \text{Min(i)}$$

The **variance** measures the dispersion of the data from the mean:

Population:

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

Sample:

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

As you can see, variance is the sum of deviations from the mean. We must square the deviations because if we didn't, the sum would always be zero.

The **standard deviation** is the square root of the variance. By taking the square root, we return the measure to the same scale as the mean. It indicates how close the data is to the mean.

Population:

$$\sigma = \sqrt{\sigma^2}$$

Sample:

$$s = \sqrt{s^2}$$

Or more specifically:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

One final measure of variability is the **standard error**. indicates how close the sample mean is from the true population mean. The means obtained from samples are estimates of the population mean, and it will vary if we were to calculate the means of different samples from the same population. As a result, we should be asking, "how close is the mean obtained from our sample to the true mean?" The standard error gives us an indication of the reliability of the mean that we obtained. Because we get closer to the real mean with larger sample sizes, the SE will decrease as we increase our sample size. Thus, the standard error is calculated by dividing the standard deviation by the square root of the total number of observations. We use this calculation in many of our statistics analyses rather than the standard deviation.

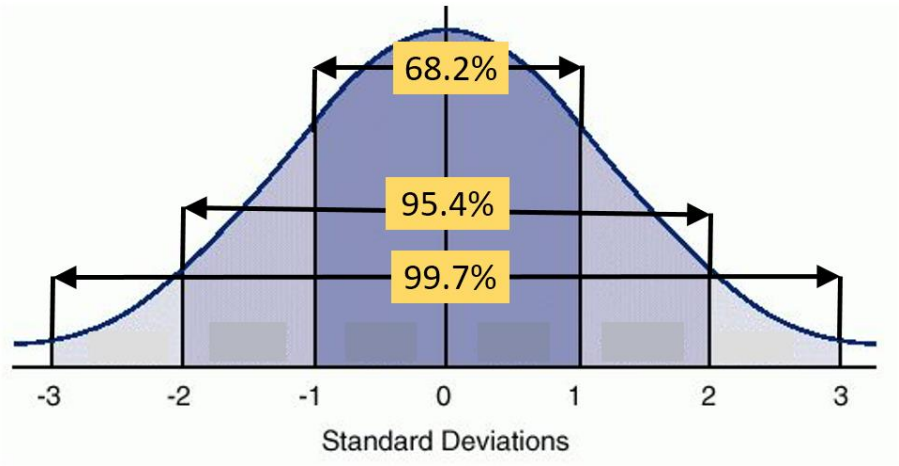
$$SE = \frac{s}{\sqrt{n}}$$

Why is this important?

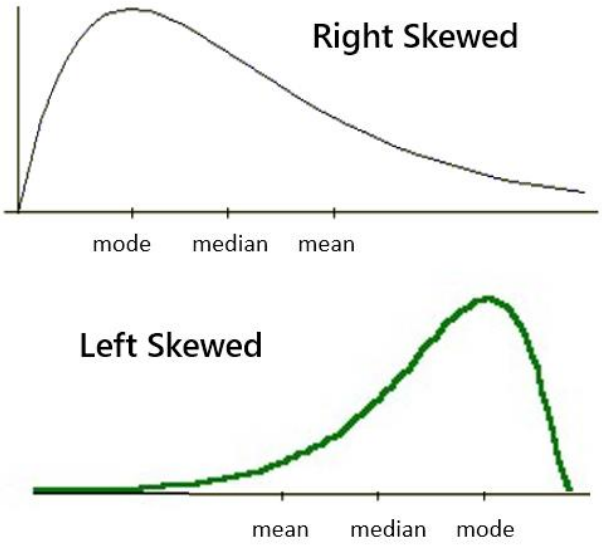
Knowing not only where the mean, median, and mode lie in the data but also the amount of variability provides you with needed information about the shape of the distribution. Most statistics assume a normal distribution, meaning that the data approximate a bell-shaped curve. In normal distributions, 68% of the data fall within +/-1 standard deviation from the mean; 95% within 2 standard deviations, and 99% within 3 standard deviations.

However, the data can take on other shapes, including right or positive skewed, where the tail of the distribution is on the right side of the curve (as indicated by a median and mode that is less than the mean) or left or negative skewed, where the tail is to the left (as indicated by a median and mode that is greater than the mean).

Normal Distribution:



Skewed Distributions:



Two other measures provide additional information about the shape of the distribution and how closely it approximates a normal distribution:

- **Kurtosis** is a measure of peakedness. Is the tall and narrow or is it short and flat?
- **Skewness** is a measure of symmetry of the data. The skewness value tells you the direction of the tail. If it is positive, the distribution is right skewed; if negative, the distribution is left skewed. A normal distribution has a skew of 0.

Hypothesis Testing

Hypothesis testing is the process by which we reject or fail to reject statistical hypotheses. The two types of statistical hypotheses:

- The **null hypothesis**, H_0, is the hypothesis that the result from the statistical analysis occurs purely by chance.
- The **alternative hypothesis**, H_1 or H_a, is the hypothesis that the result from the statistical analysis is meaningful and not influenced by chance.

The result of the analysis is always used to test the null hypothesis. You do not test the alternate hypothesis because it's easier to reject of fail to reject the null than it is to do this for the alternate hypothesis.

To test a statistical hypothesis, you:

1. State the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. The null hypothesis is usually the opposite of the result that you hope to find. Alternate hypothesis, what you hope to find as the result of your analysis, can take one of three forms—that there is a difference but the direction of that difference is unknown, that mean of group A is greater than the mean of group B, and that the mean of group A is less than the mean of group B. So, your hypotheses will look something like this:

   Predicts a difference but not the direction (two tailed test):

   $$H_0: \mu = \bar{x}$$

   $$H_a: \mu \neq \bar{x}$$

   Predicts population mean is larger than sample mean (one tailed test):

   $$H_0: \mu \leq \bar{x}$$

   $$H_a: \mu > \bar{x}$$

   Predicts population mean is smaller than sample mean (one tailed test):

   $$H_0: \mu \geq \bar{x}$$

   $$H_a: \mu < \bar{x}$$

2. Select the significance level. Often, researchers choose significance levels, or *p* values, of 0.01, 0.05, or 0.10. This value represents the probability of obtaining a significant result if the

null were true, meaning that you rejected the null hypothesis when you should have failed to reject it. Essentially, this value is used to determine if the resulting statistic is "significant" and reflects the amount of "risk" that you're willing to assume that you fail to reject the null when you should have. Another way to look at this is that with a p value of .05, you have 5 chances in 100 of finding a significant result when it doesn't exist. Because this value is critical in determining significance, you must establish it in advance (e.g., $p < .05$).

3. Determine which statistical analysis to conduct. This choice is based on your hypotheses.

4. Analyze the data. Obtain the test statistic and determine the probability of that statistic. Is that probability larger or smaller than the significance level you selected? If it's larger, you fail to reject the null hypothesis; if it's smaller, you reject the null hypothesis.

Why Don't We Accept the Null Hypothesis?

Acceptance implies that the null hypothesis is true. Failure to reject implies that the data are not sufficiently persuasive for us to prefer the alternative hypothesis over the null hypothesis.

Measures of Association: Correlation Coefficients

Correlations measure the strength of the relationship, or association, between two variables. One common misinterpretation is that correlations imply causation, but they do not. Correlations are not tied to causation.

The most common is the Pearson product-moment correlation coefficient, which measures the linear association between variables.

The sample's correlation coefficient is denoted by *r*, while the population correlation is denoted by ρ or *R*.

How to Interpret a Correlation Coefficient

The sign and absolute value of a correlation describe the direction and magnitude of the relationship between two variables. To interpret correlations, keep in mind that:

- The value of a correlation coefficient ranges between -1 and 1.
- The closer the absolute value is to 1, the stronger the linear relationship. The closer to 0, the weaker the relationship. Because we almost always mean a Pearson correlation, which is measure of linear association, a correlation near 0 doesn't necessarily mean that there is no relationship; it only means that the relationship is not linear (it could be curvilinear or logarithmic, for example).

- The sign of the correlation tells you the direction of the relationship. Positive correlations mean that the variables are moving in the same direction; as one increases, so does the other. Negative correlations mean that the variables are moving in opposite directions; as one increases, the other decreases.

How to Calculate a Correlation Coefficient

1. State your null and alternate hypotheses. For example,

$$H_0: r = 0, H_a: r \neq 0$$
$$H_0: r \geq 0, H_a: r < 0$$
$$H_0: r \leq 0, H_a: r > 0$$

2. Establish your significance level.

3. Run the statistical analysis. In the case of correlations, the formula is as follows:

$$r = \frac{1}{n-1} \sum \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right)$$

where n is the number of observations in the sample, x_i is the x value for observation i, x is the sample mean of x, y_i is the y value for observation i, y is the sample mean of y, s_x is the sample standard deviation of x, and s_y is the sample standard deviation of y.

Calculate **degrees of freedom**. Degrees of freedom is the number of values in your data that can be allowed to vary and still result in the same mean. Imagine you have four numbers (a, b, c and d) that must add up to a total of x. You are free to choose the first three numbers at random, but the fourth must be chosen so that it makes the total equal to x because that's the only way you can obtain the same mean. Thus, in this example, three is the number that you use for degrees of freedom.

For correlations, *df* = n-2 because two parameters (means) are estimated.

4. Compare the correlation you obtain to a table of critical values for correlations. If the correlation obtained is larger than the correlation in that table given your alternate hypothesis, significance level, and degrees of freedom, the correlation is considered significant and you reject the null hypothesis.

Tables of critical values can be found online. For example, Illinois.edu, Radford.edu, and Vassar have tables and calculators that you can use.

To use these tables, you:

1. Decide if you are doing a one or two tailed test. This is based on your alternate hypothesis. If you have predicted a direction for the correlation (e.g., that it's positive or negative), you are running a one-tailed test. If you have not predicted a direction, you are running a two-tailed test.
2. Calculate degrees of freedom.
3. Locate this *df* in the table. This row becomes the lower limit needed for the correlation to be significant.
4. Find your significance level in the top row. Follow that column down the table until it intersects with the *df* row.
5. If your correlation is equal to or greater than the correlation in that cell, it is significant, and you reject the null hypothesis.

Imagine that you have a correlation of .35 with 25 degrees of freedom ($p<.05$). To determine if that is significant, you find the row for 25 and the column for significance of .05. If you're running a two-tailed test, the minimum correlation needed is .381; thus, you fail to reject the null hypothesis. But, you are running a one-tailed, test the minimum correlation is .323, and you do reject the null hypothesis. This pattern will always be true. If you can predict the direction of the result, you put all your risk on one side of the distribution, meaning that the resulting statistic can be smaller and yield significant results. However, if you're wrong about the direction, you will never obtain a significant result!

	Level of Significance for a One-Tailed Test										
	.05	.025	.01	.005	.0005		.05	.025	.01	.005	.0005
	Level of Significance for a Two-Tailed Test										
df=(N-2)	.10	.05	.02	.01	.001	*df*=(N-2)	.10	.05	.02	.01	.001
1	0.988	0.997	0.9995	0.9999	0.99999	21	0.352	0.413	0.482	0.526	0.640
2	0.900	0.950	0.980	0.990	0.999	22	0.344	0.404	0.472	0.515	0.629
3	0.805	0.878	0.934	0.959	0.991	23	0.337	0.396	0.462	0.505	0.618
4	0.729	0.811	0.882	0.971	0.974	24	0.330	0.388	0.453	0.496	0.607
5	0.669	0.755	0.833	0.875	0.951	25	0.323	0.381	0.445	0.487	0.597
6	0.621	0.707	0.789	0.834	0.928	26	0.317	0.374	0.437	0.479	0.588
7	0.582	0.666	0.750	0.798	0.898	27	0.311	0.367	0.430	0.471	0.579
8	0.549	0.632	0.715	0.765	0.872	28	0.306	0.361	0.423	0.463	0.570
9	0.521	0.602	0.685	0.735	0.847	29	0.301	0.355	0.416	0.456	0.562
10	0.497	0.576	0.658	0.708	0.823	30	0.296	0.349	0.409	0.449	0.554

Rules of Thumb for Correlations

Although it's highly recommended that you use some form of significance testing to evaluate the correlation obtained, some general rules of thumb for interpreting the size of correlations have been developed.

Correlation	Interpretation
.00 to .20	No correlation to “negligible” positive
.20 to .40	Weak to moderately strong positive
.40 to .60	Moderately strong positive
.60 to .80	Strong to very strong positive
.80 to 1.00	Very strong to "perfect"
.00 to -.20	No correlation to “negligible” negative
-.20 to -.40	Weak to moderately strong negative
-.40 to -.60	Moderately strong negative
-.60 to -.80	Strong to very strong negative
-.80 to -1.00	Very strong to "perfect" negative

Comparative Measures: One Sample *t*-Test

One sample *t*-tests are performed when you want to compare a sample mean to a known mean, usually obtained from a population.

How to Calculate a One Sample *t*-Test Statistic

1. State your null and alternate hypotheses. For example,

$$H_0: \mu = \bar{x}, H_a: \mu \neq \bar{x}$$

$$H_0: \mu \leq \bar{x}, H_a: \mu > \bar{x}$$

$$H_0: \mu \geq \bar{x}, H_a: \mu < \bar{x}$$

2. Establish your significance level.

3. Run the statistical analysis. In the case of one sample t-tests, the formula is as follows:

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

   where $\bar{x}$ is the sample mean, μ is the hypothesized population mean in the null hypothesis, s is the standard deviation, and n is the number of observations.

   To do this, you will need to compute the following:

   - Standard error and $SE = \frac{s}{\sqrt{n}}$
   - Degrees of freedom (*df* = n – 1).

4. Compare the t-statistic obtained to a critical t-value table. If the obtained t-statistic is larger than the value in that table given your alternate hypothesis, significance level, and

degrees of freedom, the result is considered significant, and you reject the null hypothesis.

Tables of critical values as well as calculators that can be used to determine the probability of the statistic obtained can be found online. For example, Texas A&M, Dell Statistics, and San Jose State University have tables that could be used and, of course, you can use Excel. You use these tables exactly as described for correlations in the section above.

NOTE: Most statistical packages that calculate this statistic will provide not only the t-value obtained from the formula but also its significance. These tables are provided because they clarify the hypothesis testing process and illustrate the difference between one- and two-tailed tests and how sample size can affect the outcome.

Comparative Measures: Two Sample *t*-Test

Two sample t-tests are performed when you want to compare means from two independent groups.

How to Calculate a Two Sample *t*-Test Statistic

1. State your null and alternate hypotheses. For example:

$$H_0\colon \mu_1 - \mu_2 = 0 \text{ or } H_0\colon \mu_1 - \mu_2 = d$$
$$H_a\colon \mu_1 - \mu_2 \neq 0 \text{ or } H_a\colon \mu_1 - \mu_2 \neq d$$

$$H_0\colon \mu_1 - \mu_2 \leq 0 \text{ or } H_0\colon \mu_1 - \mu_2 \leq d$$
$$H_a\colon \mu_1 - \mu_2 > 0 \text{ or } H_a\colon \mu_1 - \mu_2 > d$$

$$H_0\colon \mu_1 - \mu_2 \geq 0 \text{ or } H_0\colon \mu_1 - \mu_2 \geq d$$
$$H_a\colon \mu_1 - \mu_2 < 0 \text{ or } H_a\colon \mu_1 - \mu_2 < d$$

   d is the hypothesized difference between the means. When this difference is 0, the first variation of the hypotheses above would be used.

2. Establish your significance level.

3. Run the statistical analysis. In the case of two sample t-tests, the formula is as follows:

$$\frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\bar{x}$ are the sample means, D_0 is the hypothesized difference between the means, s^2_p is the pooled variance, and n is the number of observations.

To do this, you will need to compute the following:

- Pooled variance (which is used to calculate the standard error that is used as the denominator in the formula above) and $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
- Degrees of freedom ($df = n_1 + n_2 - 2$).

4. Compare the t-statistic obtained to a critical t-value table. If the obtained t-statistic is larger than the value in that table given your alternate hypothesis, significance level, and degrees of freedom, the result is considered significant, and you reject the null hypothesis.

The same critical t-value tables used for a one-sample test are used for a two-sample test.

Comparative Measures: Paired Sample *t*-Test

Paired sample t-tests are performed when you want to compare means for dependent groups. This dependence is usually the result of having matched pairs across groups or when the same people or things have been measured twice. Analyzing "pre" and "post" mean differences is a good example when you would use a paired sample t-test.

How to Calculate a Paired Sample *t*-Test Statistic

1. State your null and alternate hypotheses. For example:

$$H_0: \mu_d = d, H_a: \mu_d \neq d$$

$$H_0: \mu_d \leq d, H_a: \mu_d > d$$

$$H_0: \mu_d \geq d, H_a: \mu_d < d$$

μ_d *is* the true difference in population values and *d* is the hypothesized difference between paired values from two different groups. *d* is calculated as: $d = x_1 - x_2$, where x_1 is the value of variable x in the first data set, and x_2 is the value of the variable from the second data set that is paired with x_1.

2. Establish your significance level.

3. Run the statistical analysis. In the case of two sample t-tests, the formula is as follows:

$$t = \frac{\bar{d} - D_0}{s_d / \sqrt{n}}$$

where $\bar{d}$ is the mean difference between paired observations, D_0 is the hypothesized difference between these pairs, s_d is the standard deviation of the differences, and n is the number of observations.

To do this, you will need to compute the following:

- Standard deviation of the paired differences (which is used to calculate the standard error that is used as the denominator in the formula above) and

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n2}\right)}$$

$$s_1^2 = \frac{\sum (X - \bar{X}_1)^2}{n_1 - 1}$$

$$s_2^2 = \frac{\sum (X - \bar{X}_2)^2}{n_2 - 1}$$

Degrees of freedom (*df* = n-1).

4. Compare the t-statistic obtained to a critical t-value table. If the obtained t-statistic is larger than the value in that table given your alternate hypothesis, significance level, and degrees of freedom, the result is considered significant, and you reject the null hypothesis.

The same critical t-value tables used for a one-sample test are used for a two-sample test.

Probability Basics

A probability is a number that reflects the chance or likelihood that a given event will occur. **Probabilities** can be expressed as proportions that range from 0 to 1 or as percentages, ranging from 0% to 100%.

Probability of a Single Event

A roll of a six-sided die can result in one of six possible outcomes, and each of these outcomes is equally likely. You are just as likely to roll a five as you are a two. What is the probability that you will roll a six? Because there are six possible outcomes, the probability is 1/6.

What is the probability that roll a five or six? The two outcomes about which we are concerned (a one or a six coming up) are called **events of interest** or **favorable outcomes** (favorable,

meaning that the event will occur…this is not necessarily favorable to you, especially if you were betting on the toss of the die being a one). Given that all outcomes are equally likely, we can compute the probability of a five or a six by summing the total number of events of interest (2) and dividing by the number of possible outcomes (6). In this case, 2/6 or 1/3, meaning that 1/3 of the time, you will roll a five or a six.

You can apply this logic to many different probability calculations. For example, what is the probability that a card drawn at random from a standard deck of 52 cards (i.e., 4 aces, 4 twos, 4 threes, …, 4 tens, 4 jacks, 4 queens, and 4 kings) that contains 4 suits (hearts, clubs, spades, and diamonds) will be an ace? Since the deck has four aces, there are four outcomes that would result in you drawing an ace. Because the deck has 52 cards, there are 52 possible outcomes. The probability is, therefore, 4/52 = 1/13. What about the probability that the card will be a heart? Since there are 13 hearts, the probability is 13/52 = 1/4.

Let's say you have a bag with 50 candy bars: 34 are Snickers®, 10 are Reese's® Peanut Butter Cups, and 6 are Butterfingers®. If you pick a candy bar at random, what is the probability that it will be a peanut butter cup? Given that there are 50 candy bars and 10 of them are peanut butter cups, the probability is 10/50 or 1/5, assuming that the probability of picking any of the candy bars is the same. This formula couldn't be applied if you were more likely to pick a peanut butter cup because it was a different shape or bigger than the other types of candy bars. The assumption of equal probability for all outcomes is critical if this formula is used to determine the probability of a given event.

A More Complex Example

Let's imagine you toss 2 dice. What is the probability that the sum of the two dice will be 10? To solve this problem, we need to know how many total possible combinations can result from tossing two dice. It turns out that there are 36 possible outcomes from tossing two dice as shown below:

Die 1	Die 2	Total		Die 1	Die 2	Total		Die 1	Die 2	Total
1	1	2		3	1	4		5	1	6
1	2	3		3	2	5		5	2	7
1	3	4		3	3	6		5	3	8
1	4	5		3	4	7		5	4	9
1	5	6		3	5	8		5	5	10
1	6	7		3	6	9		5	6	11
2	1	3		4	1	5		6	1	7
2	2	4		4	2	6		6	2	8
2	3	5		4	3	7		6	3	9
2	4	6		4	4	8		6	4	10

2	5	7		4	5	9		6	5	11
2	6	8		4	6	10		6	6	12

You can see that 3 of the 36 possibilities total 10. Therefore, the probability is 3/36.

Probability Axioms (also known as "Rules")

Axiom One

The first axiom of probability is that the probability of an event is at least zero. This means that the smallest that a probability can ever be is zero. While this axiom says nothing about how large the probability of an event can be, it does eliminate the possibility of negative probabilities. It reflects the notion that smallest probability, reserved for impossible events, is zero.

Axiom Two

The second axiom of probability is that the probability of the entire sample space is one. The chance that something in the outcome space occurs is 100% because the outcome space contains every possible outcome. By itself, this axiom does not set an upper limit on the probabilities of events unless, of course, there is only one event that makes up the set of all possible outcomes (known as the sample space), but it does say with certainty that something in the sample space will occur.

Axiom Three

The third axiom of probability deals with mutually exclusive events. If A and B are mutually exclusive or disjoint, meaning that they share no outcomes, the probability that either of the events happens is the sum of the probabilities that each happens.

Axiom Four or the Law of Complements

The complement of an event is all outcomes that are NOT the event of interest. If P(A) is the probability of Event A, then 1 - P(A) is the probability that the event does not occur. If we toss two dice and want to know the probability that the sum will NOT equal 10, we simply subtract the probability that it will be 10 from 1. If the probability that the dice will sum to 10 is 3/36, then the complement is 1 – 3/36 or 33/36.

Probability of Two (or More) Independent Events

Events A and B are independent if the outcome of one doesn't affect the outcome of the other. In other words, if the occurrence of Event A does not change the probability of Event B and vice versa, then Events A and B are independent.

For example, a fair coin is tossed twice. The probability that you will toss a head on the second toss is ½ regardless of if you tossed a head on the first toss (and it doesn't matter how many

times you toss the coin, or how many heads you toss, the probability is the same with each toss… don't let the gambler's fallacy tell you differently!).

If events A and B are independent, then the **probability of both A and B** occurring is:

P(A and B) = P(A) x P(B)

where P(A) is the probability of event A occurring, and P(B) is the probability of event B occurring.

If you flip a coin twice, what is the probability that it will be heads both times?

- Event A = toss a head (P = ½)
- Event B = toss a head (P = ½)

Thus, the probability that both events occur is: ½ x ½ = ¼.

Let's imagine that you flip a coin and roll a six-sided die. What is the probability that you toss a tail, and roll a six?

- Event A = toss a tail (P = ½)
- Event B = roll a 6 (P = 1/6)

Thus, the probability of both events occurring is ½ x 1/6 = 1/12.

Now, imagine that you draw a card from a standard deck of 52 cards, replace it, and then draw another card. What is the probability that your first card will be a heart, and the second card will be black?

- Because 13 of the cards are hearts, the probability that the first card is a heart is 13/52 = ¼.
- Because 26 of the cards are black, the probability that the second card is black is 26/52 = ½.

Thus, the probability of both events occurring is ¼ x ½ = 1/8.

What if it didn't matter if you drew the heart on the first card or as the second? Or, if the black card came first or came second? Then, you would sum the probability of a heart first and a black card second with the probability of the black card first and the heart second, following the logic unions, or the probability that A or B will occur (in this example, event A is drawing the heart first followed by the black card and event B is the black card first followed by the heart—you just need to think about events A and B a bit differently).

Mutually Exclusive Events

Two events are **mutually exclusive** if they cannot occur at the same time. For example, when tossing a fair coin, you cannot get a head and a tail on the same toss.

For two mutually exclusive events, A and B, the probability of either one occurring, P(A or B), is the sum of the probability of each event.

- Imagine you are rolling a die, and you want to know the probability of rolling an odd number (1, 3, or 5). To calculate this, you would sum the probabilities of each event:
- Probability of rolling a 1: 1/6
- Probability of rolling a 3: 1/6
- Probability of rolling a 5: 1/6
- Probability of rolling an odd number = 1/6 + 1/6 + 1/6 = ½.

Let's take another look at our bag of candy bars. Recall that it has 50 candy bars: 34 are Snickers®, 10 are Reese's® Peanut Butter Cups, and 6 are Butterfingers®. What is the probability that it will be a peanut butter cup or a Butterfinger? Given that there are 50 candy bars, 10 of them are peanut butter cups, and 6 are Butterfingers, the probability is 16/50 if the probability of picking any of the candy bars is the same.

It's a bit weird to think of it this way, but oddly, mutually exclusive events are dependent on each other. If one occurs, the other cannot.

Unions (or the Probability of A or B with Independent Events)

If Events A and B are independent, the probability that **either** Event A or Event B occurs is:

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

Sometimes, you will see this represented as:

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

When we say that "A or B occurs," we mean one of three possibilities:

- A occurs, but B does not.
- B occurs, but A does not.
- Both A and B occur

In this case, "or" is technically called *inclusive or* because it includes the case in which both A and B occur. If we did not include this possibility, then we would be using an *exclusive or*—the latter is what we mean by "mutually exclusive" events, which we will discuss in the next section.

Here's how this works. If you flip a coin two times, what is the probability that you will get a head on the first flip, a head on the second flip, or a head on both flips? Let Event A be a head on the first flip and Event B be a head on the second flip.

- Probability of a head on the first flip: P(A) = ½
- Probability of a head on the second flip: P(B) = ½
- From the previous section, we know that P(A and B) = P(A) x P(B) = ½ x ½ = ¼.

Thus, P(A or B) = ½ + ½ - ¼ = ¾.

If you toss a six-sided die, and then flip a coin, what is the probability that you will get either a 4 on the die, a head on the coin flip, or both?

P(4 or head) = P(4) + P(head) - P(4 and head)
= (1/6) + (½) - (1/6) x (½)
= 7/12

If you toss a die five times, what is the probability that at least one of your tosses results in a 2? That is, what is the probability of getting a 2 on the first toss OR a 2 on the second toss OR a 2 on the third toss OR a 2 on the fourth toss OR a 2 on the fifth toss? The easiest way to approach this problem is to apply the law of complements. In other words, what is the probability of NOT getting 2 on any toss. Do this, you would compute the probability of:

NOT getting a 2 on the first throw (5/6)
AND not getting a 2 on the second throw (5/6)
AND not getting a 2 on the third throw (5/6)
AND not getting a 2 on the fourth throw (5/6)
AND not getting a 2 on the fifth throw (5/6)

The answer will be 1 minus this probability (remember that's the law of complements). The probability of not getting a 2 on any of the three throws is 5/6 x 5/6 x 5/6 x 5/6 x 5/6 = 3125/7776. Therefore, the probability of getting a 2 on at least one of the throws is 1 - 3125/7776 = 4651/7776 or 59.8%.

Probability of Dependent Events

What is the probability that two cards drawn at random from a deck of playing cards will both be aces? It might seem that you could use the formula for the probability of two independent events, and simply multiply 4/52 x 4/52 = 1/169. However, the two events are not independent. If the first card drawn is an ace, then the probability that the second card is also an ace would be lower because there are only three aces left in the deck.

Once the first card is chosen, the probability of drawing a similar card (e.g., one of the same value, suit, or color) changes. Thus, we state the probability of the drawing two aces like this:

P(ace on second draw | an ace on the first draw)

The vertical bar "|" is read as "given," so this literally means the probability of an ace on the second draw given an ace on the first draw. If we draw an ace on the first card, there are 3 aces out of 51 total cards left. This means that the probability that one of these remaining aces will be drawn is 3/51 = 1/17.

Two events are dependent if the outcome or occurrence of the first affects the outcome or occurrence of the second so that the probability is changed. When two events, A and B, are dependent, the probability of both occurring is:

P(A and B) = P(A) x P(B|A)

P(B|A) means that the probability of B given that A has occurred, and it's known as a conditional probability.

Imagine that a card is chosen at random from a standard deck of 52 playing cards. Without replacing it, a second card is chosen. What is the probability that the first card chosen is an ace and the second card chosen is a queen?

P(ace on first pick) = 4/52

P(queen on 2nd pick given that an ace was pulled first) = 4/51

P(ace and queen) = 4/52 x 4/51 = 16/2652

So, if Events A and B are not independent, then P(A and B) = P(A) x P(B|A) applies. Applying this to the problem of two aces, the probability of drawing two aces from a deck is 4/52 x 3/51 = 1/221.

If you draw two cards from a deck, what is the probability that you will get the ace of hearts and a black card? There are two ways you can satisfy this condition:

(a) You can get the ace of hearts first and then a black card or

(b) you can get a black card first and then the ace of hearts.

Let's calculate the first. The probability that the first card is the ace of hearts is 1/52. The probability that the second card is black given that the first card is the ace of hearts is 26/51 because 26 of the remaining 51 cards are black. The probability is 1/52 x 26/51 = 1/102.

Now for the second possibility. The probability that the first card is black is 26/52 = 1/2. The probability that the second card is the ace of hearts given that the first card is black is 1/51. The probability is, therefore, 1/2 x 1/51 = 1/102, the same as the probability of first possibility.

Recall that the probability of A or B is P(A) + P(B) - P(A and B). In this problem, P(A and B) = 0 since a card cannot be the ace of hearts and be a black card. Therefore, the probability of the first or second possibility is 1/102 + 1/102 = 2/102 = 1/51.

Now, imagine that you want to know the probability of drawing the ace of spades and black card. As before, there are two ways you can satisfy this condition:

(a) You can get the ace of spades first and then a black card or

(b) you can get a black card first and then the ace of spades.

In this case the probability of the second card changes. For example, the probability of the first is 1/52 x 25/51, rather than 26 because the first card is also black, = 25/2652. The probability of the second is 26/52 x 1/51

Conditional Probabilities

As mentioned above, the probability that Event A occurs, given that Event B has occurred, is called a conditional probability. The conditional probability of Event B given Event A, is denoted by the symbol P(B|A).

There is a formula for conditional probability is:

P(B|A) = P(A and B) / P(A).

Alternately, you may see this formula represented like this:

P(B|A) = P(A ∩ B) / P(A)

So, how does this work? Well, we know that when two events, A and B, are dependent, the probability of both occurring is P(A and B) = P(A) * P(B|A).

The formula for the conditional probability of an event can be derived from that formula. Divide both sides of the equation by P(A), so you have P(A and B) / P(A) = P(A) * P(B|A) / P(A). As you'll recall from your math courses, P(A) as a divisor cancels P(A) out in the numerator, leaving P(A and B) / P(A) = P(B|A)—the equation for conditional probability.

Imagine that probability that it rains 10% of the time. The probability that it's raining and that you're late is 5%. What is the probability that you're late given that it's raining?

P(A) = Probability that it's raining (10%)

P(A and B) = Probability that it's raining and that you're late (5%)

P(B|A) = Probability that you're late given that it's raining = 5%/10% = 50%. In other words, you are late 50% of the time when it's raining.

This is not the same probability that it's raining given that your late (P(A|B) = P(A and B)/P(B).

P(B) = Probability that you're late (5%)

P(A and B) = Probability that it's raining and that you're late (5%)

P(A|B) = Probability that it's raining given that you're late = 5%/5% = 100%.

In other words, if you are late, it's raining, but just because you're late doesn't mean it's raining. You are late for other reasons as reflected by the 50% probability that it's raining when you are late, but every time it rains, you will be late.

Let's look at one more example. Imagine that the probability that it is Friday and that you are absent from work is 0.07. The probability that it's Friday is .2 (you work five days each week). What is the probability that you are absent given that today is Friday?

P(Absent|Friday) = P(Friday and Absent) / P(Friday) = 0.07/0.20 = 0.35.

Birthday Problem

If there are 25 people in a room, what is the probability that at least two of them share the same birthday? It is not 25/365 = 0.068. In this case, calculating the complement (e.g., 1 minus the probability that no two people have the same birthday) is more straightforward approach to solving this problem.

If we choose two people at random, what is the probability that they do not share a birthday? Of the 365 days on which the second person could have a birthday, 364 of them are different from the first person's birthday. Therefore, the probability is 364/365.

Now, let's find the probability that the third person does not share a birthday with the other two. This is a conditional probability (i.e., the probability that third person does not have the same birthday as the other two given that the previous two did not share a birthday). If the previous two did not share a birthday, then two of the 365 days in a year have been used, leaving 363 non-matching days. Therefore, the probability that the third person does not share a birthday is 363/365. You would use the same logic for the fourth person (362/365), the fifth (361/365), and so on up to the 25^{th} person (341/365).

Essentially for there to be no shared birthdays, the birthday of the second person must not be the same as the first person, **and** the third person must not share a birthday with any previous

person, **and** the fourth person must not share a birthday with any previous person, and so on. Since P(A and B) = P(A)P(B), we multiply the probabilities for each successive person (person 2 through person 25). The result is 0.431. Therefore, the probability of at least one matching birthday in a group of 25 is 0.569.

Appendix: A Few More Statistics of Interest

Comparative Measures: Analysis of Variance (ANOVA)

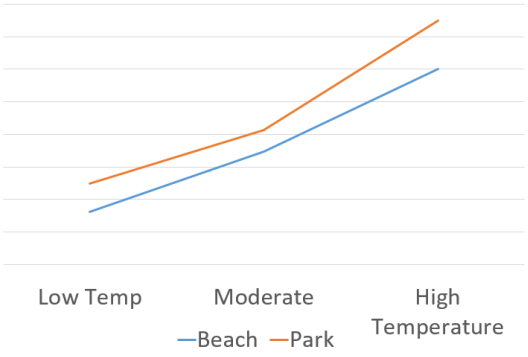
Analysis of variance (ANOVA) is used to simultaneously test for differences between the means of two or more independent groups (many sources state that the minimum is three, but it can be used with two although this is technically a t-test). In the case of ANOVA, group membership is the "treatment" condition (also known as the independent variable) and is a categorical variable. The outcome of interest (the dependent variable because the result 'depends' on group membership) must be a continuous variable.

In Rosie's lemonade example, Rosie might wonder if she sells more lemonade on hot days. To test this, she would define a temperature range for cool, moderate, and hot days. Based on the temperature, she would classify each day into one of these three groups. She would then run an ANOVA to determine if sales are significantly different for any of these groups. This is an example of a one-way ANOVA; it compares the effects of different levels of one treatment variable on an outcome of interest.

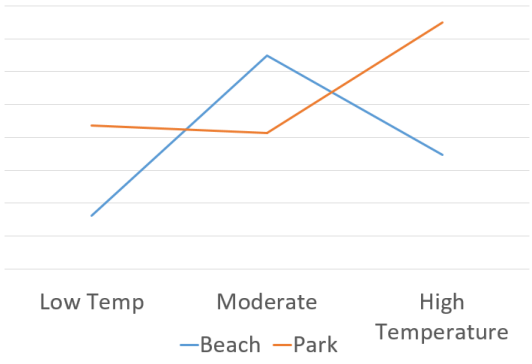
A two-way ANOVA allows for the comparison of multiple treatment conditions on an outcome of interest. For example, Rosie may wonder if she sells more lemonade at the beach or the park based on temperature. Two-way ANOVA analysis allows for the researcher to test not only for main effects but also for interactions between levels of the various treatment conditions.

A **main effect** occurs when the effect of one treatment on the outcome variable is the same across all levels of the other treatments. An **interaction** occurs when the effect of one treatment on the outcome differs at some levels of the other independent variable.

A main effect would look something like this when plotted. As you can see, temperature is clearly influencing sales, but the effect is uniform across temperatures regardless of location.



In the figure below, you can see an interaction between the independent variables. Rosie is clearly selling more lemonade at the beach on days when the temperatures are moderate and more lemonades at the park when temperatures are high.



ANOVA is an **omnibus** test statistic, meaning that the result will tell if there is at least one significant mean difference, but it will not tell you where it is. To identify which groups are statistically different, you will need to run a post hoc analysis.

Why You Shouldn't Run Multiple *t*-Tests

For any statistical test, the significance level selected informs the probability of NOT obtaining a significant result when the null hypothesis is true; in other words, it's the probability that you'll fail to reject the null hypothesis when you should fail to reject it. If you have a *p* value = 0.05, then this probability is 0.95 (=1-.05).

If you run multiple, independent statistical test using the same sample of data, the odds of not obtaining a significant result when the null is true start to decline, making it more likely you will fail to reject the null when you should reject it. This is based on probability math for two independent outcomes. If you run two independent tests of the same data, the actual probability of failing to reject when the null is true is calculated by multiplying the probably of each independent event: .95*.95 = .90, meaning that you have nearly a 10% chance of failing to reject the null when you should.

Imagine if you were making 3 comparisons (.95*.95*.95=.857), you would have increased the chance of making an error to nearly 15%!

ANOVA compares all means simultaneously and maintains error probability at the level selected by the researcher.

How to Calculate a One-Way ANOVA

1. State your null and alternate hypotheses. In this case, the null hypothesis is that the means of all treatment levels are the same:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_p$$

The alternative is that at least one level of the treatment has a different effect:

$$H_a: \text{at least one of } \mu_1, \mu_2, \ldots, \mu_p \text{ differs}$$

2. Establish your significance level.

3. Run the statistical analysis. The formulas for ANOVA are extremely complicated, but they are fundamentally comparing the between-group variability to the within-group variability. Between-group variability is the variability of the sample means obtained for each level of the treatment condition. Within-group variability is the variability of the individual observations (that is, the values) within each treatment level. When between group variability is large relative to within group, the resulting statistic will be significant; there is a difference between the groups based on the treatment.

4. The outcome of this analysis is a f-statistic, and it is compared to the values obtained from a critical f-table in the same manner as you would do with a t-test. You are simply using a different set of tables. Unlike the t-statistic distribution, which is normally distributed, the f-statistic distribution right skewed. Thus, you will use a different table depending on your desired significance level. Examples can be found: http://homepages.wmich.edu/~hillenbr/619/AnovaTable.pdf and http://documents.software.dell.com/Statistics/Textbook/Distribution-Tables.

How to Calculate a One-Way ANOVA

The process for calculating a two-way ANOVA is essentially the same as what you do for a one-way ANOVA. However, the hypothesis testing process is different because you will have hypotheses for the main effects for all your treatment variables as well as hypothesis about the existence of an interaction.

Is there a main effect for variable 1?

- $H_0: \mu_1 = \mu_2 = \ldots = \mu_p$ (Means of all levels of treatment 1 are the same)
- H_a: at least one of $\mu_1, \mu_2, \ldots, \neq\mu_p$ differ

Is there a main effect for variable 2?

- $H_0: \mu_1 = \mu_2 = \ldots = \mu_p$ (Means of all levels of treatment 2 are the same)
- H_a: at least two of $\mu_1, \mu_2, \ldots, \neq\mu_p$ differ

Is there an interaction between the variables?

- H_0: There is no interaction
- H_a: There is an interaction

Predictive Measures: Linear Regression

Regression is used to evaluate cause and effect relationships. In these relationships, you are predicting an outcome, or dependent, variable (the "effect") from one or more predictor, or independent, variables ("the cause"). Regression can take many forms, but the most common is least squares linear regression. In least squares regression, the line minimizes the sum of squared differences between the outcome values in the data and the outcome values that are predicted based on the regression equation that is estimated based on the data.

Simply put, regression is a function that estimates the best fitting line through a set of data. In its simplest form, it looks like this: $y = b_0 + b_1x$, which is the same equation for estimating a line. Thus, we know that b_0 is the y-intercept of the line, and b_1 is the slope, or the average change in the dependent variable for a unit change in the independent variable. In regression terminology, the b values are called "beta weights."

The population regression line is:

$$Y = B_0 + B_1X,$$

where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

The sample regression line is: $\hat{y} = b_0 + b_1x$.

How to Calculate a Regression

1. State your null and alternate hypotheses. In this case, the null hypothesis is stated for regression coefficients for each independent variable included in the analysis. The alternate hypothesis is the mutually exclusive opposite of the null.

$$H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \ldots \beta_n = 0,$$

$$H_a: \beta_1 \neq 0$$

$$H_a: \beta_2 \neq 0$$

$$H_a: \beta_3 \neq 0$$

.....

$$H_a: \beta_n \neq 0$$

2. Establish your significance level.

3. Run the statistical analysis. The formulas for regression, like ANOVA, are extremely complicated, but the key output for this analysis is R^2, or the coefficient of determination. If this is significant, then at least one of the beta weights obtained from the analysis is significant; if this is not significant, no need to look further as none of the beta weights are significant.

   The coefficient of determination is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable(s). Because it is a proportion, it ranges from 0 to 1. When R^2 is near 0, the dependent variable cannot be predicted from the independent variable(s). When it's near 1, the dependent variable can be predicted *without error* from the independent variable(s). Thus, this value indicates the extent to which the dependent variable can be predicted from the independent variable(s) and is essentially the percentage of variance in the dependent variable that can be explained by the independent variables. For example, $R^2 = 0.15$ means that 15 percent of the variance in dependent variable is predictable from independent variable(s); an $R^2 = 0.68$ means that 68% can be explained.

   Assuming a significant R^2, the next step is to determine which beta weights and the corresponding independent variables are significant predictors in the regression equation.

4. The outcome of this analysis is a f-statistic, and it is compared to the values obtained from a critical f-table in the same manner as you would do with a t-test. You are simply using a different table.

Resources

A great overview/introduction to statistics: http://stattrek.com/

Distribution tables used to determine probabilities of a wide number of statistics: http://documents.software.dell.com/Statistics/Textbook/Distribution-Tables