

# Statistical Insights

Microsoft Professional Program

## Table of Contents

What is a Variable? .....	3
Why is this important?.....	3
Population vs. Sample .....	4
Why is this important?.....	4
Measures of Central Tendency .....	5
Measures of Variability.....	5
Why is this important?.....	6
Hypothesis Testing .....	8
Why Don't We "Accept" the Null Hypothesis? .....	9
Measures of Association: Correlation Coefficients.....	9
How to Interpret a Correlation Coefficient .....	9
How to Calculate a Correlation Coefficient .....	10
An Example.....	11
Rules of Thumb for Correlations.....	12
Comparative Measures: One Sample t-Test .....	12
How to Calculate a One Sample t-Test Statistic .....	12
Comparative Measures: Two Sample t-Test.....	13
How to Calculate a Two Sample t-Test Statistic .....	13
Comparative Measures: Paired Sample t-Test .....	14
How to Calculate a Paired Sample t-Test Statistic.....	14
Comparative Measures: Analysis of Variance (ANOVA) .....	15
Why You Shouldn't Run Multiple t-Tests .....	17
How to Calculate a One-Way ANOVA.....	18
How to Calculate a One-Way ANOVA.....	18
Predictive Measures: Linear Regression .....	19
How to Calculate a Regression.....	19
Resources .....	20



Statistics is fundamentally concerned with summarizing data so that appropriate interpretations can be made. In organizations, these interpretations are the basis of decisions that drive 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, the more of some attribute the object has. For example, temperature, sales, and number of flyers posted are quantitative variables. Quantitative variables can be:
  - **Continuous:** A value that is measured along a scale (e.g., temperature) or
  - **Discrete:** A value that is counted in fixed units (e.g., the number of flyers distributed).
- **Categorical variables** are those for which the value indicates group membership. Thus, you can't say that one person, place, or thing 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 perform meaningful mathematical calculations (e.g., means and standard deviations) on quantitative variables. While you can perform these same calculations on categorical variables, the results 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 locations of drink sales, you must assign a number, or dummy code, to each location. Although location is represented by numbers in your data, it doesn'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 spent at the park and 70% were spent at the beach).

## 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, and notation vary depending on whether you are 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 a subtle but important difference in the population and sample formulas for variance and standard deviation. In samples, you divide by  $n-1$  because the mean that we use in this calculation approximates the true mean. Because we are estimating both the mean and standard deviation from the same subset of data, we 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:

- $\sigma^2$ : Population variance
- $\sigma$ : Population standard deviation
- $s^2$ : Sample variance
- $s$ : Sample standard deviation
- $\mu$ : Population mean
- $\bar{x}$ : 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 is calculated by summing across all observations and dividing by the total number of observations.

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 in the data for a given variable; 50% of the values are above, 50% of the values 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 the data using 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 each observation's deviation from the mean. We must square these 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:

Sample:

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

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

Or more specifically:

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

One final measure of variability is the **standard error**. It 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. Thus, 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 obtained mean. Because we get closer to the real mean with larger sample sizes, the SE will decrease as we increase our sample size.

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 any statistical analysis based on samples rather than the standard deviation.

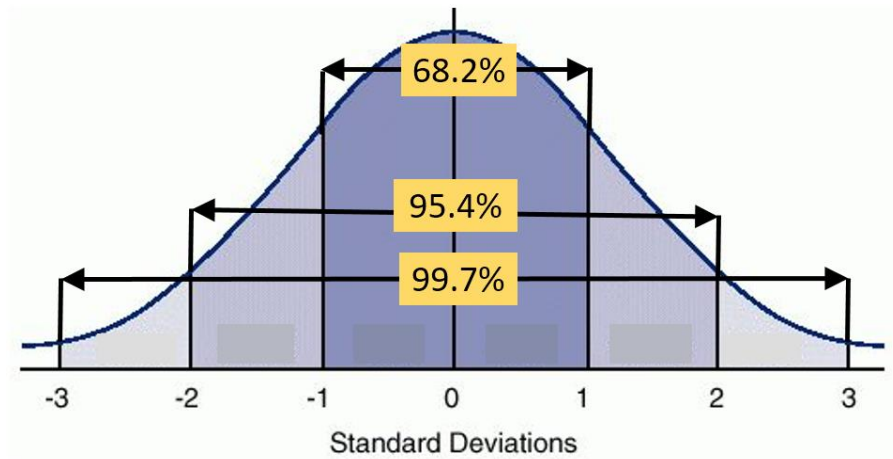
$$SE = \frac{s}{\sqrt{n}}$$

### Why is this important?

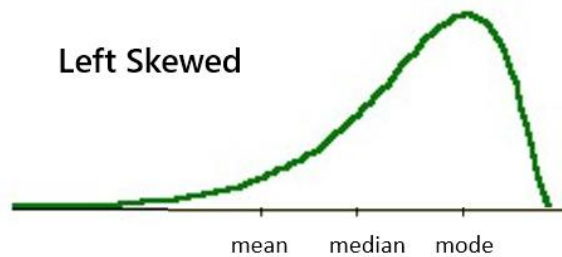
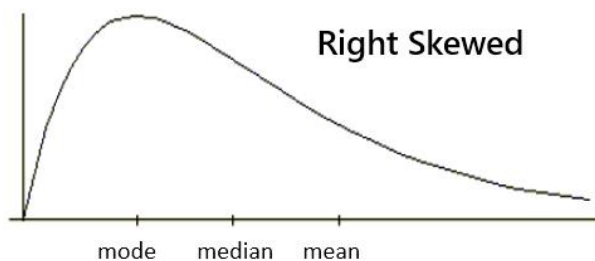
Knowing not only where the mean, median, and mode lie in the data but also the amount of variability provides you with needed information about the shape of the distribution. Most statistics assume a normal distribution, meaning that the data approximate a bell-shaped curve. In normal distributions, 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 (positive) skewed, where the tail of the distribution is on the right side of the curve (as indicated by a median and mode that are less than the mean) or left (negative) skewed, where the tail is to the left (as indicated by a median and mode that are 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 distribution tall and narrow, or is it short and flat?
- **Skewness** is a measure of the symmetry of the data. The skewness value indicates 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 **null hypothesis**,  $H_0$ , is the hypothesis that the result obtained from the statistical analysis occurred purely by chance.
- The **alternative hypothesis**,  $H_1$  or  $H_a$ , is the hypothesis that the result obtained from the statistical analysis is meaningful and not influenced by chance.

The result of the statistical analysis is always used to test the null hypothesis. Because it's easier to reject or fail to reject the null hypothesis than it is to do this for the alternate, you do not test the 'truth' of 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 the opposite of the result that you hope to find. The 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 **not** have.

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 rejected it. 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. Failure to reject recognizing the possibility that we could be wrong in this decision. Although beyond the scope of this overview, this is known as Type II error.

## 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 correlation calculation is the Pearson product-moment correlation, which measures the linear association between two variables.

The sample’s correlation coefficient is denoted by  $r$  while the population correlation is denoted by  $\rho$  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:

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

where  $n$  is the number of observations in the sample,  $x$  is a given value of the first variable,  $\bar{x}$  is the sample mean of  $x$ ,  $y$  is a given value of the second variable,  $\bar{y}$  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 the **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](http://Illinois.edu), [Radford.edu](http://Radford.edu), and [Vassar](http://Vassar) have tables and calculators that you can use, but any Bing search for 'correlation significance tables' will provide a long list of options.

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 either 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.

**An Example.** 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, if 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 of 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 size of correlation obtained, some general rules of thumb for interpreting the size of correlations have been developed, especially when you have extremely large samples where even very small correlations are likely to be statistically significant (traditional measures of statistical significance are heavily influenced by sample size with larger sample size requiring small statistics for significance).

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,  $\mu$  is the hypothesized population mean,  $s$  is the standard deviation, and  $n$  is the number of observations.

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

- Standard error (which is the denominator in the formula above) and

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

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

4. Compare the t-statistic obtained to a critical t-score 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. Again, any Bing search for '*critical t-score table*' will yield a wide variety of options.

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 mentioned above because this statistic is very easy to calculate without using any statistical software. In addition, understanding how to use these tables can help you better understand the hypothesis testing process and illustrate the difference between one- and two-tailed tests as well as 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:

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

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

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

$d$  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,  $s_p^2$  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-score 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-score 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 of 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:

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

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

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

$\mu_d$  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

$$s_{x_1 - x_2} = \sqrt{\left( \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

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

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

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

4. Compare the t-statistic obtained to a critical t-score 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: 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



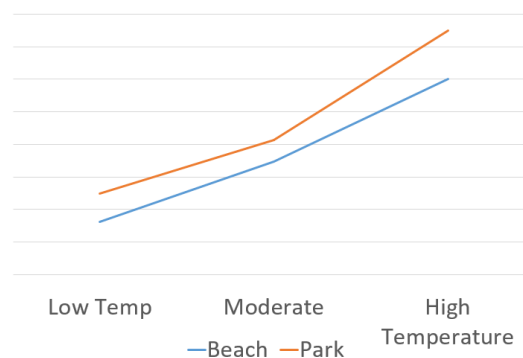
be used with two although it should be noted that this is technically a  $t$ -test). In the case of ANOVA, group membership is the “treatment” condition (also known as the “independent variable”) and is a categorical variable. The outcome of interest (known as the “dependent variable” because the result ‘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 reported for that day, 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.

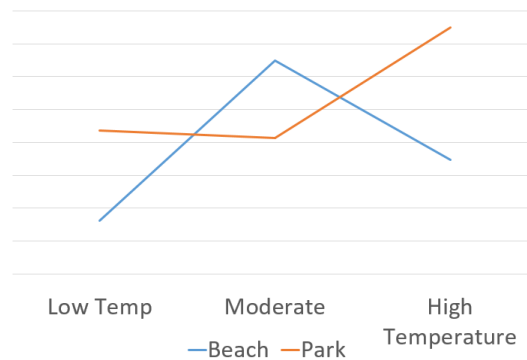
A two-way ANOVA allows for the comparison of multiple treatment conditions on an outcome of interest. For example, Rosie may wonder if she sells more lemonade at the beach or the park based on temperature. Two-way ANOVA analysis allows you to test not only for main effects but also for interactions between 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 that a treatment has on the outcome differs at some level(s) 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. There are several different types of post hoc analyses that can be performed. The most common are Bonferroni, Fisher's Least Squared Differences, and Tukey. Regardless of the method selected, they all control for the possibility of capitalizing on chance when multiple  $t$ -tests are performed as described in the next section.

### Why You Shouldn't Run Multiple $t$ -Tests

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

If you run multiple, independent statistical tests 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 probability of each independent event:  $.95 \times .95 = .90$ , meaning that you have nearly a 10% chance of failing to reject the null when you should.

Imagine if you were making 3 comparisons ( $.95 \times .95 \times .95 = .857$ ), 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 that you select (e.g.,  $p < .05$ ).

## 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:

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

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

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

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 (hence, the name analysis of *variance*). 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*-score 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, however, the *f*-statistic distribution is 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>. Again, a Bing search of 'critical *f*-score tables' will yield a wide variety of results.

## How to Calculate a One-Way ANOVA

The process for calculating a two-way ANOVA is essentially the same as 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 a hypothesis about the existence of an interaction.

Is there a main effect for variable 1?

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

Is there a main effect for variable 2?

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

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

$$H_a: \beta_1 \neq 0$$

$$H_a: \beta_2 \neq 0$$

$$H_a: \beta_3 \neq 0$$

...

$$H_a: \beta_n \neq 0$$

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  $f$ -score 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>