# Hypothesis Testing: A Review

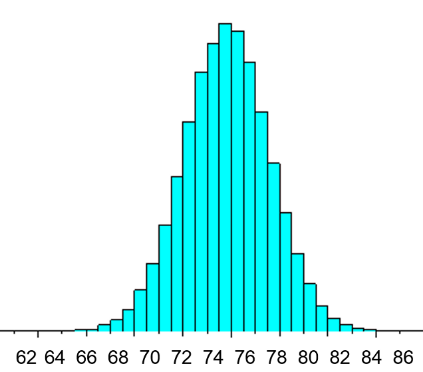
Recall that a population is a well-defined collection of objects, and a sample is a subset of the population that’s obtained in some way.

In order to understand the idea behind hypothesis testing, recall that statisticians rely on sample data to make statements about the population from which these samples came. For example, the sample mean is often used to make statements about population mean . That is, the sample mean is a *point estimator* of the population mean, and its value is a *point estimate* of the population mean.

Imagine that we have the following scenario: a real estate agent claims that the average value of houses in her town is $75,000; however, one of the buyers, concerned about her investment, disagrees, and believes that the mean value of all the houses in the town is actually lower.

Let’s say that the buyer collects a sample of *n* = 100 houses recently sold in the town, and calculates the average house price in the sample, , which turns out to be $60,000. But the buyer realizes that the value of will depend on which 100 houses end up being in her sample, meaning that values will vary from one sample to the next. In fact, if the buyer were to take, say, 500different samples of 100 houses, each of those 500 samples would likely yield a different value of , and a distribution of all values might look like the one in Figure 1. Statisticians have been able to describe distributions for the sample mean, sample proportion, sample correlation coefficient, and many other *sample statistics* – i.e., quantities that can be calculated from sample data.

Figure 1. Sample means from 500 different samples with n=100 houses in each



Of course, in practice, the buyer will only have *one* sample of 100 houses. But she will use her knowledge of the distribution of the sample means to do a hypothesis test (: vs. : ) and assess the likelihood of obtaining a sample where the house price is *at least* $15,000 different[[1]](#footnote-1) from what’s being claimed by the realtor, assuming that the realtor’s claim is true. That is, if the true population house value is indeed $75,000, the buyer wants to calculate the probability of her drawing a sample where the average house price is at most $60,000. This probability is known as the *p-value*. Statisticians typically reject[[2]](#footnote-2) the null hypothesis, , in favor of the alternative hypothesis, , if *p* < 0.05.[[3]](#footnote-3)

So let’s solve the problem! Again, our null hypothesis is that : , and our alternative hypothesis is : . Let’s say, for the sake of the example, that we know the population standard deviation, , to be $40,000.

Recall that under the Central Limit Theorem, for any *n* if *X* is normal, or for *n* > 30 if *X* is not normal, the statistic corresponding to the sample mean, , will have a normal distribution. Furthermore, the mean of is equal to the population mean of *X* (i.e., ), and the standard deviation of is equal to the standard deviation of *X* divided by the square root of the sample size (i.e., ).

In this example, *n* = 100 > 30, so the Central Limit Theorem applies regardless of whether *X* is normal. So, *if* the null hypothesis is true, then will be normal, its mean will be $75,000 (i.e., ), and its standard deviation will be .

We can then standardize *X* – i.e., re-express the distance between $75,000 and $60,000 in terms of standard deviations:

The statistic *Z* has a standard normal distribution (which is a normal distribution with mean 0 and standard deviation of 1), and by using any standard normal table (available in most statistics textbooks and software packages), *z* = -3.75 corresponds to a p-value of 0.000177. Because p < 0.05, we can reject for . That is, it’s very unlikely that a sample where the mean is $60,000 came from the population where the mean is $75,000 simply due to sampling variability.

In this example, we tested a hypothesis about the population mean, but the same idea applies to hypothesis testing about any other population parameter. In general, we want to answer the following question: if the population parameter’s value is in fact the one that’s stated under the null hypothesis (i.e., if is true), then what is the probability of finding the observed, or more extreme, results with our sample data? This probability is the *p-value[[4]](#footnote-4)*. For our intents and purposes, if the p-value is less than 0.05, we reject in favor of . That is, if the p-value < 0.05, it’s unlikely that the sample that we are observing came from the population with the parameter value stated in .

In order to test a hypothesis about any population parameter, we need to know the distribution of the statistic that is used to estimate it, and for this class, we will be given this information.

## **Reviewing Hypothesis Testing about Pearson Correlation Coefficients**

The sample Pearson correlation coefficient between variables *x* and *y* is denoted by *r*, and is calculated as follows:

In the formula above, *n* is the number of observations, and are the sample means of variables *x* and *y*, respectively, and and are the sample standard deviations of the variables *x* and *y*, respectively.

Pearson correlation is a unitless measure of how much the variables *x* and *y* change together, and its values range between -1 and 1. A correlation of *r* = -1 indicates that all points lie on a line with a negative slope, and a correlation of *r* = 1 indicates that all points lie on a line with a positive slope. A Pearson correlation of *r* = 0 indicates that there’s no linear relationship between *x* and *y*.

Different fields have different standards for what is considered a weak, moderate, and strong correlations. For instance, Devore suggests the following rules of thumb:

* 0 ≤ | *r* | ≤ 0.5: weak correlation
* 0.5 < | *r* | < 0.8: moderate correlation
* 0.8 ≤ | *r* | < 1: strong correlation

However, the way that statisticians assess whether two variables *x* and *y* are related is by doing a hypothesis test about the population correlation coefficient, . In those tests, the idea is to see whether the correlation coefficient is *statistically significantly different* from 0. So, we formulate two hypotheses:

: , vs. :

We saw in the realtor example above that under certain conditions, the distribution of the sample mean will be normal, with mean and standard deviation . When we are dealing with hypothesis tests about , we need to have certain information about the distribution of *R*, the statistic that corresponds to the sample correlation coefficient.

And we have this information: if is true, then , and statisticians have estimated[[5]](#footnote-5) to be .

Furthermore, we are told that the quantity has a *t* distribution[[6]](#footnote-6) that has *n* – 2 degrees of freedom.

As mentioned earlier, in a typical problem, we will be given sample data from which we will calculate the sample correlation coefficient *r*. Then, we will want to see whether the value of *r* is significantly different from 0. The idea here, once again, is that the values of *r* will vary from one sample to the next – i.e., if you draw many samples of *x* and *y* and calculate *r* from each of those samples, it will change. So even if the true *population* correlation between *x* and *y* is 0, the *sample* correlation coefficients will not be exactly 0, but they will be close to 0, and values of *r* obtained from different samples will be centered around 0. Once again, in practice, we will be dealing with only one sample, and we want to see how likely it is to have a value of *r* that’s at least as different from 0 as it is in the sample we have at hand simply due to sampling variability.

Let’s do an example. Let’s say that we have a data set (i.e., sample) of 100 observations for which we compute the correlation *r* between values of *x* and *y*. Let’s imagine that we compute the sample correlation coefficient to be *r* = 0.7. The next question is how likely we are to get the sample correlation coefficient that’s at least 0.7 *if* the true population correlation coefficient is 0 (i.e., if is true).

We can calculate the quantity *T* as follows:

We know that this quantity has a *t*-distribution with *n* – 2 = 100 – 2 = 98 degrees of freedom.

Using any *t* table (available in most statistics textbooks and software packages), *t =* 9.703 corresponds to a p-value < 0.0001. Since p-value < 0.05, we can reject the null hypothesis for the alternative. Said differently, the sample correlation coefficient *r* = 0.7 is 9.703 standard deviations higher than the value of 0, and it’s extremely unlikely for us to observe *r* that is at least 0.7 if the sample came from a population where the true population correlation coefficient was (value stated in ). Said in yet another way, if the true population correlation coefficient is , it would be highly unlikely for us to observe *r* ≥ 0.7 simply due to the sampling variability in the values of *r*.

1. $15,000 = $75,000 - $60,000 [↑](#footnote-ref-1)
2. There are only two options in hypothesis testing: you reject for , or you fail to reject for . You never accept, or prove, a hypothesis. [↑](#footnote-ref-2)
3. Because statistics is a field of uncertainty, you can commit two types of error: Type I error, where you incorrectly reject the null hypothesis when it’s actually true, and Type II error, where you incorrectly *fail to reject* the null hypothesis when it’s actually false. Ideally, you would like to minimize both types of error, but for any given sample size, they are inversely related, and the only way to minimize both simultaneously is to increase the sample size. Because it’s generally not easy (and often not possible) to increase the sample size, statisticians typically focus on Type I error, which is usually the more serious of the two. The probability of making a Type I error, or incorrectly rejecting the null hypothesis when it’s true, is denoted by *α* (alpha) and is referred to as the *significance level* of the test. This value is set *a priori*, most often to 0.05, although some social scientists will be more liberal and set it to 0.10, and some other scientists, especially in the medical field, will be more conservative and set it to 0.01. Setting *α* to 0.05 (or 0.10, or 0.01) means that you are willing to live with the fact that you will incorrectly reject your null hypothesis 5% (or 10%, or 1%) of the time. We will be using *α* of 0.05 for this class, and will be comparing the p-value to 0.05. [↑](#footnote-ref-3)
4. Furthermore, the p-value can be thought of as the *observed* significance level – that is, the *smallest*level of significance at which the null hypothesis would be rejected when a specified test procedure is used on a given data set. For instance, if our p-value is 0.23, then it means that our could be rejected only if *α* were set to a level of at least 0.23 (which isn't something we'd do in practice, because incorrectly rejecting the null hypothesis 23% of the time is a bit too much). Keep in mind that the p-value is *not* the probability that the null hypothesis is true or the probability that the alternative hypothesis is false. [↑](#footnote-ref-4)
5. Here, and in a lot of tests, we use the sample standard deviation *S* as an estimator of the population standard deviation σ, and the value of the sample standard deviation, *s*, as a *point estimate* of the population standard deviation σ. This is done because the value of σ is generally unknown in practice. [↑](#footnote-ref-5)
6. Recall that the *t*-distribution with ν degrees of freedom looks like the standard normal distribution *Z*, but has more area under the tails for smaller values of ν. As ν approaches infinity, the *t*-distribution approaches the standard normal distribution. [↑](#footnote-ref-6)