

## 2.4: Hypothesis Testing in Quantitative Research

Statistical reasoning is built on the assumption that data are normally distributed, meaning that they will be distributed in the shape of a bell curve as discussed in the chapter on Univariate Analysis. While real life often—perhaps even usually—does not resemble a bell curve, basic statistical analysis assumes that if all possible random samples from a population were drawn and the mean taken from each sample, the distribution of sample means, when plotted on a graph, would be normally distributed (this assumption is called the Central Limit Theorem). Given this assumption, we can use the mathematical techniques developed for the study of probability to determine the likelihood that the relationships or patterns we observe in our data occurred due to random chance rather than due some actual real-world connection, which we call statistical significance.

*Statistical* significance is not the same as *practical* significance. The fact that we have determined that a given result is unlikely to have occurred due to random chance does not mean that this given result is important, that it matters, or that it is useful. Similarly, we might observe a relationship or result that is very important in practical terms, but that we cannot claim is statistically significant—perhaps because our sample size is too small, for instance. Such a result might have occurred by chance, but ignoring it might still be a mistake. Let's consider some examples to make this a bit clearer. Assume we were interested in the impacts of diet on health outcomes and found the statistically significant result that people who eat a lot of citrus fruit end up having pinky fingernails that are, on average, 1.5 millimeters longer than those who tend not to eat any citrus fruit. Should anyone change their diet due to this finding? Probably not, even though it is statistically significant. On the other hand, if we found that the people who ate the diets highest in processed sugar died on average five years sooner than those who ate the least processed sugar, even in the absence of a statistically significant result we might want to advise that people consider limiting sugar in their diet. This latter result has more practical significance (lifespan matters more than the length of your pinky fingernail) as well as a larger effect size or association (5 years of life as opposed to 1.5 millimeters of length), a factor that will be discussed in the chapter on association.

While people generally use the shorthand of “the likelihood that the results occurred by chance” when talking about statistical significance, it is actually a bit more complicated than that. What statistical significance is *really* telling us is the likelihood (or probability) that a result equal to or more “extreme<sup>[1]</sup>” is true in the real world, rather than our results having occurred due to random chance or sampling error. Testing for statistical significance, then, requires us to understand something about probability.

### A Brief Review of Probability

You might remember having studied probability in a math class, with questions about coin flips or drawing marbles out of a jar. Such exercises can make probability seem very abstract. But in reality, computations of probability are deeply important for a wide variety of activities, ranging from gambling and stock trading to weather forecasts and, yes, statistical significance.

Probability is represented as a proportion (or decimal number) somewhere between 0 and 1. At 0, there is absolutely no likelihood that the event or pattern of interest would occur; at 1, it is absolutely certain that the event or pattern of interest will occur. We indicate that we are talking about probability by using the symbol  $p$ . For example, if something has a 50% chance of occurring, we would write  $p = 0.5$  or  $\frac{1}{2}$ . If we want to represent the likelihood of something *not* occurring, we can write  $1 - p$ .

**Check your thinking:** Assume you were flipping coins, and you called heads. The probability of getting heads on a coin flip using a fair coin (in other words, a normal coin that has not been weighted to bias the result) is 0.5. Thus, in 50% of coin flips you should get heads. Consider the following probability questions and write down your answers so you can check them against the discussion below.

- Imagine you have flipped the coin 29 times and you have gotten heads each time. What is the probability you will get heads on flip 30?
- What is the probability that you will get heads on *all* of the first five coin flips?
- What is the probability that you will get heads on *at least one* of the first five coin flips?

There are a few basic concepts from the mathematical study of probability that are important for beginner data analysts to know, and we will review them here.

**Probability over Repeated Trials:** The probability of the outcome of interest is the same in each trial or test, regardless of the results of the prior test. So, if we flip a coin 29 times and get heads each time, what happens when we flip it the 29th time? The probability of heads is still 0.5! The belief that “this time it must be tails because it has been heads so many times” or “this coin just wants to come up heads” is simply superstition, and—assuming a fair coin—the results of prior trials do not influence the results of this one.

**Probability of Multiple Events:** The probability that the outcome of interest will occur repeatedly across multiple trials is the product<sup>[2]</sup> of the probability of the outcome on each individual trial. This is called the multiplication theorem. Thinking about the multiplication theorem requires that we keep in mind the fact that when we multiply decimal numbers together, those numbers get *smaller*—thus, the probability that a series of outcomes will occur is *smaller than* the probability of any one of those outcomes occurring on its own. So, what is the probability that we will get heads on all five of our coin flips? Well, to figure that out, we need to multiply the probability of getting heads on each of our coin flips together. The math looks like this (and produces a very small probability indeed):

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.03125 \quad (2.4.1)$$

**Probability of One of Many Events:** Determining the probability that the outcome of interest will occur on at least one out of a series of events or repeated trials is a little bit more complicated. Mathematicians use the addition theorem to refer to this, because the basic way to calculate it is to calculate the probability of each sequence of events (say, heads-heads-heads, heads-heads-tails, heads-tails-heads, and so on) and add them together. But the greater the number of repeated trials, the more complicated that gets, so there is a simpler way to do it. Consider that the probability of getting *no* heads is the same as the probability of getting *all tails* (which would be the same as the probability of getting all heads that we calculated above). And the only circumstance in which we would not have at least one flip resulting in heads would be a circumstance in which *all* flips had resulted in tails. Therefore, what we need to do in order to calculate the probability that we get at least one heads is to subtract the probability that we get *no* heads from 1—and as you can imagine, this procedure shows us that the probability of the outcome of interest occurring at least once over repeated trials is *higher* than the probability of the occurrence on any given trial. The math would look like this:

$$1 - \left(\frac{1}{2}\right)^5 = 0.9688 \quad (2.4.2)$$

So why is this digression into the math of probability important? Well, when we test for statistical significance, what we are really doing is determining the probability that the outcome we observed—or one that is more extreme than that which we observed—occurred by chance. We perform this analysis via a procedure called Null Hypothesis Significance Testing.

## Null Hypothesis Significance Testing

Null hypothesis significance testing, or NHST, is a method of testing for statistical significance by comparing observed data to the data we would expect to see if there were no relationship between the variables or phenomena in question. NHST can take a little while to wrap one's head around, especially because it relies on a logic of double negatives: first, we state a hypothesis we believe *not* to be true (there is no relationship between the variables in question) and then, we look for evidence that disconfirms this hypothesis. In other words, we are assuming that there is no relationship between the variables—even though our research hypothesis states that we think there *is* a relationship—and then looking to see if there is any evidence to suggest there is *not* no relationship. Confusing, right?

So why do we use the null hypothesis significance testing approach?

- The null hypothesis—that there is no relationship between the variables we are exploring—would be what we would generally accept as true in the absence of other information,
- It means we are assuming that differences or patterns occur due to chance unless there is strong evidence to suggest otherwise,
- It provides a benchmark for comparing observed outcomes, and
- It means we are searching for evidence that *disconfirms* our hypothesis, making it less likely that we will accept a conclusion that turns out to be untrue.

Thus, NHST helps us avoid making errors in our interpretation of the result. In particular, it helps us avoid Type 2 error, as discussed in the chapter on Bivariate Analyses. As a reminder, Type 2 error is error where you accept a hypothesis as true when in fact it was false (while Type 1 error is error where you reject the hypothesis when in fact it was true). For example, you are making a Type 1 error if you decide not to study for a test because you assume you are so bad at the subject that studying simply cannot help you, when in fact we know from research that studying does lead to higher grades. And you are making a Type 2 error if your boss tells you that she is going to promote you if you do enough overtime and you then work lots of overtime in response, when actually your boss is just trying to make you work more hours and already had someone else in mind to promote.

We can never remove all sources of error from our analyses, though larger sample sizes help reduce error. Looking at the formula for computing standard error, we can see that the standard error ( $SE$ ) would get smaller as the sample size ( $N$ ) gets larger. Note:  $\sigma$  is

the symbol we use to represent standard deviation.

$$SE = \frac{\sigma}{\sqrt{N}} \quad (2.4.3)$$

Besides making our samples larger, another thing that we can do is that we can choose whether we are more willing to accept Type 1 error or Type 2 error and adjust our strategies accordingly. In most research, we would prefer to accept more Type 1 error, because we are more willing to miss out on a finding than we are to make a finding that turns out later to be inaccurate (though, of course, lots of research does eventually turn out to be inaccurate).

### Performing NHST

Performing NHST requires that our data meet several assumptions:

1. Our sample must be a random sample—statistical significance testing and other inferential and explanatory statistical methods are generally not appropriate for non-random samples<sup>[3]</sup>—as well as representative and of a sufficient size (see the Central Limit Theorem above).
2. Observations must be independent of other observations, or else additional statistical manipulation must be performed. For instance, a dataset of data about siblings would need to be handled differently due to the fact that siblings affect one another, so data on each person in the dataset is not truly independent.
3. You must determine the rules for your significance test, including the level of uncertainty you are willing to accept (significance level) and whether or not you are interested in the direction of the result (one-tailed versus two-tailed tests, to be discussed below), in advance of performing any analysis.
4. The number of significance tests you run should be limited, because the more tests you run, the greater the likelihood that one of your tests will result in an error. To make this more clear, if you are willing to accept a 5% probability that you will make the error of accepting a hypothesis as true when it is really false, and you run 20 tests, one of those tests (5% of them!) is pretty likely to have produced an incorrect result.

If our data has met these assumptions, we can move forward with the process of conducting an NHST. This requires us to make three decisions: determining our null hypothesis, our confidence level (or acceptable significance level), and whether we will conduct a one-tailed or a two-tailed test. In keeping with Assumption 3 above, we must make these decisions before performing our analysis. The null hypothesis is the hypothesis that there is no relationship between the variables in question. So, for example, if our research hypothesis was that people who spend more time with their friends are happier, our null hypothesis would be that there is no relationship between how much time people spend with their friends and their happiness.

Our confidence level is the level of risk we are willing to accept that our results could have occurred by chance. Typically, in social science research, researchers use  $p < 0.05$  (we are willing to accept up to a 5% risk that our results occurred by chance),  $p < 0.01$  (we are willing to accept up to a 1% risk that our results occurred by chance), and/or  $p < 0.001$  (we are willing to accept up to a 0.1% risk that our results occurred by chance).  $P$ , as was noted above, is the mathematical notation for probability, and that's why we use a  $p$ -value to indicate the probability that our results may have occurred by chance. A higher  $p$ -value increases the likelihood that we will accept as accurate a result that really occurred by chance; a lower  $p$ -value increases the likelihood that we will assume a result occurred by chance when actually it was real. Remember, what the  $p$ -value tells us is not the probability that our own research hypothesis is true, but rather this: assuming that the null hypothesis is correct, what is the probability that the data we observed—or data more extreme than the data we observed—would have occurred by chance.

Whether we choose a one-tailed or a two-tailed test tells us what we mean when we say “data more extreme than.” Remember that normal curve? A two-tailed test is agnostic as to the direction of our results—and many of the most common tests for statistical significance that we perform, like the Chi square, are two-tailed by default. However, if you are only interested in a result that occurs in a particular direction, you might choose a one-tailed test. For instance, if you were testing a new blood pressure medication, you might only care if the blood pressure of those taking the medication is significantly *lower* than those not taking the medication—having blood pressure significantly *higher* would not be a good or helpful result, so you might not want to test for that.

Having determined the parameters for our analysis, we then compute our test of statistical significance. There are different tests of statistical significance for different variables (for example, the Chi square discussed in the chapter on bivariate analyses), as you will see in other chapters of this text, but all of them produce results in a similar format. We then compare this result to the  $p$  value we already selected. If the  $p$  value produced by our analysis is lower than the confidence level we selected, we can reject the null hypothesis, as the probability that our result occurred by chance is very low. If, on the other hand, the  $p$  value produced by our

analysis is higher than the confidence level we selected, we fail to reject the null hypothesis, as the probability that our result occurred by chance is too high to accept. Keep in mind this is what we do even when the p value produced by our analysis is quite close to the threshold we have selected. So, for instance, if we have selected the confidence level of  $p < 0.05$  and the p value produced by our analysis is  $p = 0.0501$ , we still fail to reject the null hypothesis and proceed as if there is not any support for our research hypothesis.

I actually like to think of the null hypothesis as ‘innocent until proven guilty’: the null hypothesis (innocence) is assumed to be true as long as there isn’t enough evidence to reject it. —Patrick Altmeyer @paltmey via twitter, 09/13/2022, 3:55 pm.

Thus, the process of null hypothesis significance testing proceeds according to the following steps:

1. Determine the null hypothesis
2. Set the confidence level and whether this will be a one-tailed or two-tailed test
3. Compute the test value for the appropriate significance test
4. Compare the test value to the critical value of that test statistic for the confidence level you selected
5. Determine whether or not to reject the null hypothesis

Your statistical analysis software will perform steps 3 and 4 for you (before there was computer software to do this, researchers had to do the calculations by hand and compare their results to figures on published tables of critical values). But you as the researcher must perform steps 1, 2, and 5 yourself.

### Confidence Intervals & Margins of Error

When talking about statistical significance, some researchers also use the terms confidence intervals and margins of error. Confidence intervals are ranges of probabilities within which we can assume the true population parameter lies. Most typically, analysts aim for 95% confidence intervals, meaning that in 95 out of 100 cases, the population parameter will lie within the upper and lower levels specified by your confidence interval. These are calculated by your statistics software as well. The margin of error, then, is the range of values within the confidence interval. So, for instance, a [2021 survey](#) of Americans conducted by the Robert Wood Johnson Foundation and the Harvard T.H. Chan School of Public Health found that 71% of respondents favor substantially increasing federal spending on public health programs. This poll had a 95% confidence interval with a  $\pm 3.6$  margin of error. What this tells us is that there is a 95% probability (19 in 20) that between 67.4% ( $71 - 3.6$ ) and 74.6% ( $71 + 3.6$ ) of Americans favored increasing federal public health spending at the time the poll was conducted. When a figure reflects an overwhelming majority, such as this one, the margin of error may seem of little relevance. But consider a similar poll with the same margin of error that sought to predict support for a political candidate and found that 51.5% of people said they would vote for that candidate. In that case, we would have found that there was a 95% probability that between 47.9% and 55.1% of people intended to vote for the candidate—which means the race is total tossup and we really would have no idea what to expect. For some people, thinking in terms of confidence intervals and margins of error is easier to understand than thinking in terms of p values; confidence intervals and margins of error are more frequently used in analyses of polls while p values are found more often in academic research. But basically, both approaches are doing the same fundamental analysis—they are determining the likelihood that the results we observed or a similarly-meaningful result would have occurred by chance.

### What Does Significance Testing Tell Us?

One of the most important things to remember about significance testing is that, while the word “significance” is used in ordinary speech to mean importance, significance testing does *not* tell us whether our results are important—or even whether they are interesting. A full understanding of the relationship between a given set of variables requires looking at statistical significance *as well as* association and the theoretical importance of the findings. Table 1 provides a perspective on using the combination of significance and association to determine how important the results of statistical analysis are—but even using Table 1 as a guide, evaluating findings based on theoretical importance remains key. So: make sure that when you are conducting analyses, you avoid being misled into assuming that significant results are sufficient for making broad claims about the importance and meaning of results. And remember as well that significance only tells us the likelihood that the pattern of relationships we observe occurred by chance—not whether that pattern is causal. For, after all, quantitative research can never eliminate all plausible alternative explanations for the phenomenon in question (one of the three elements of causation, along with association and temporal order).

Table 1. Significance and Association

	Significance	
	Significant	Not Significant

Strength of Association	<b>Strong</b>	Something's happening here!	Could be interesting, but might have occurred by chance
	<b>Weak</b>	Probably did not occur by chance, but not interesting	Nothing's happening here

## Exercises

1. Using the approach described in this chapter, calculate the probability of the following coin flip scenarios:

- Getting 7 heads on 7 coin flips
- Getting 5 heads on 7 coin flips
- Getting 1 head on 10 coin flips

Then check your work using the [Coin Flip Probability Calculator](#).

2. Write the null hypothesis for each of the following research hypotheses:

- As the advertised hourly pay for a job goes up, the number of job applicants increases.
- Teenagers who watch more hours of makeup tutorial videos on TikTok have, on average, lower self-esteem.
- Couples who share hobbies in common are less likely to get divorced.

3. Assume a research conducted a study that found that people wearing green socks type on average one word per minute faster than people who are not wearing green socks, and that this study found a p value of  $p < 0.01$ . Is this result *statistically* significant? Is this result *practically* significant? Explain your answers.

4. If we conduct a political poll and have a 95% confidence interval and a margin of error of  $\pm 2.3\%$ , what can we conclude about support for Candidate X if 49.3% of respondents tell us they will vote for Candidate X? If 24.7% do? If 52.1% do? If 83.7% do?

1. One way to think about this is to imagine that your result has been plotted on a bell curve. Statistical significance tells us the probability that the "real" result—the thing that is true in the real world and not due to random chance—is at the same point as or further along the skinny tails of the bell curve than the result we have plotted. ↩
2. In other words, what you get when you multiply. ↩
3. They also are not appropriate for censuses—but you do not need inferential statistics in a census because you are looking at the entire population rather than a sample, so you can simply describe the relationships that do exist. ↩

This page titled [2.4: Hypothesis Testing in Quantitative Research](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by [Mikaila Mariel Lemonik Arthur](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.