6.4: Hypothesis Testing

Probability and Hypothesis Testing

Hypothesis testing is a process by which data are analyzed and conclusions are drawn about whether the results support or refute the hypothesis. This process allows statisticians to determine the likelihood that their results are not due to chance and, instead, likely represent truths about populations that are in keeping with their hypotheses. Note that the tenet of probability (introduced in Chapter 1) is a component of this process. When a hypothesis is tested, steps are followed and calculations are performed to assess the *probability* (likelihood) that a hypothesis is true based on sample data. Thus, the terms supported and refuted are used instead of the words proven and disproven, respectively.

Hypotheses come in a variety of forms, each of which requires different statistical methods of analysis. Some hypotheses, like the one about oatmeal and cholesterol from earlier in this chapter, state that a treatment condition (or intervention) will cause a difference in a measurable outcome variable. To review, that hypothesis written in sentence and symbol formats is:

Cholesterol levels will be lower after (post) eating oatmeal daily for six weeks compared to before (pre).

$$H_a : \mu_{\text{post}} < \mu_{\text{pre}}$$

In this example daily oatmeal consumption is the treatment condition and cholesterol level is the outcome being measured. We can see from the symbol format that data will need to be collected from a sample both before the treatment in order to compute a pretest mean and again from the same sample after the treatment in order to compute a posttest mean. The means can then be compared to see if, as is expected based on the hypothesis, the mean cholesterol level for the sample is lower at posttest than it was at pretest.

It is tempting and seems logical to simply conclude that if the posttest mean is even slightly lower than the pretest mean, the hypothesis is supported. However, before we can draw this conclusion we need to assure that our results are strong enough to conclude that the difference is unlikely to simply be due to chance and that they, instead, likely reflect a real difference in the means. This is because sample means are estimates and are expected to have some sampling error; sample means are not expected to be perfect representations of population parameters under the same conditions. Slight differences in means could simply be due to sampling error. Thus, estimates of error (such as the standard deviation or standard error) must also be considered in order to determine how likely it is that the difference in the pretest and posttest sample means represent an actual difference that would be observed in population parameters under the same conditions. For this reason, estimates of error are an important part of statistical power and determining significance.

Statistical Power and Significance

Statistical power refers to how likely it is that sample data will support the hypothesis. Think of statistical power as the ability to detect that an alternative hypothesis is true if, in fact, it is true. Power is generally increased or decreased by three factors:

1. the size of the sample,
2. the size of the change, difference, or pattern observed in the sample data, and
3. the size of the error in the relevant estimates of the changes, differences, or patterns observed.

First, if the hypothesis is true, data to support it are more likely when the sample is larger than when it is smaller. Therefore, as sample sizes increase, power also increases. Second, if the change, difference, or pattern observed in the sample is larger or clearer, it is easier to detect and is more likely to represent a difference that would occur in the population than if it is smaller. Thus, as the size of changes, differences, or patterns observed increases, power also increases. Finally, the lower the error is, the closer the observations are expected to be to the parameters of a population. Thus, as the size of error decreases, power increases. Considering these three things together, we can summarize the components that increase power as follows:

1. The larger the sample size, the more closely the sample statistics are expected to represent the population.
2. The larger or clearer the change, difference, or pattern observed in the sample, the more likely it is that it reflects a change, difference, or pattern in the population.
3. The less error there is in the sample statistics used to assess changes, differences, or patterns, the more likely it is that they reflect changes, differences, or patterns in the population.

Obtained Values

These components that impact statistical power are interconnected. Means are estimates for which variability must be considered. Some measures of variability and error (such as standard errors) include sample sizes in their calculations. Further, the greater the

sample size, the closer a sample is to being equivalent to its population size. Thus, the formulas used in inferential statistics include variations of some or all of the three components of power to yield one of several forms of obtained values. **Obtained values** are results that summarize data by using inferential formulas. Inferential formulas are those used to test hypotheses. These formulas and other analyses that accompany them take into account the components of power. Data are plugged into inferential formulas which yield obtained values. Those obtained values are compared to specific thresholds to assess whether the data supported or failed to support a hypothesis. Thus, obtained values can be thought of as summaries of how much power or evidence there is to support a hypothesis.

Determining Significance

Statistical significance refers to the determination that a hypothesis is likely true in the population because there is sufficient evidence in the sample to support the hypothesis. Another way to say this is that a statistically significant result occurs when the hypothesized result was observed in the sample with enough power to conclude that the observed result was unlikely to be simply due to random chance. Essentially, when an obtained value is high enough for a given situation, it represents sufficient evidence to declare a hypothesis is significantly supported.

Significance is not absolute. Instead, it is a determination that a hypothesis is *likely* true but not that it is proven to be true. Recall that sampling error is assumed whenever a sample is drawn and used to represent a population. Recall also that there is no guarantee that the sample will represent the population well. Thus, there is always some chance that a hypothesis is not true in the population but that it will appear to be true in the data from a sample. The stronger the evidence is in favor of the hypothesis within the sample, the more likely it is that the hypothesis is true of the population. To say it another way, the stronger the results are, the less likely it is that they would have occurred simply due to random chance rather than because they are true. Therefore, when a result matches a hypothesis and is significant, statisticians conclude that a hypothesis is likely true and, thus, is supported by the evidence. Note that statistical significance is not necessarily indicative that a result is meaningful or useful. Instead, statistical significance simply indicates that the hypothesis is likely true based on the evidence.

Critical Values

Obtained values are compared to critical values to determine whether a hypothesis has enough evidence to be declared significant and, thus, supported. **Critical values** represent thresholds of the minimum amount of evidence that is needed to determine statistical significance and conclude that a hypothesis is supported. Thus, when the obtained value (which represents the amount of evidence) exceeds the critical value (which represents the minimum amount of evidence needed to support a hypothesis), the conclusion is that the hypothesis is supported. Conversely, when the obtained value does not exceed the critical value, the conclusion is that there is insufficient evidence to support the hypothesis. Another way to say this is that the null hypothesis is rejected when the obtained value exceeds the critical value and is retained or accepted when the obtained value does not exceed the critical value.

Obtained and critical values depend on several things which can include whether or not a hypothesis is directional, which inferential formula was used, and the relevant components of power for the hypothesis and corresponding formula used. We will review the specific differences and ways both obtained values and their critical values are found in subsequent chapters. For now, it is only necessary to know that each time a hypothesis is tested, an obtained value must be found, a critical value must be found, and the two must be compared. These are important steps in the larger processes of hypothesis testing.

Steps in Hypothesis Testing

In order to test a hypothesis, these steps should be followed, in the recommended order:

1. State the hypothesis.

This is a necessary first step. Before a study can be designed, a researcher needs to specify exactly what the hypothesis is what they intend to test. Then the process for collecting data (which is the research method) can be developed and carried out, accordingly.

It is worth noting that it is possible to develop a hypothesis after data have been collected but this is not ideal as it introduces important limitations to the research process. Though these limitations are beyond the scope of this book, they are an important topic which is generally covered in a Research Methods course. The focus of this book is best practices for statistical analysis; in keeping, we will always presume a hypothesis was developed before data were collected to test it. Thus, the first step for our analyses and reporting our results will always include stating the hypothesis.

2. Choose the inferential test (formula) that best fits the hypothesis.

There are a variety of formulas, each of which best fits only certain kinds of data and, thus, each only fits certain hypotheses. For example, one test is used to compare the means of the same group at posttest to itself at pretest, a different one is used to compare the mean of one group to the mean of a different group, another is used to compare the means of three or more distinct groups, and still others are used to assess patterns between two or more quantitative variables. The test selected should be the one that is best suited to the hypothesis under investigation. Note: A brief summary of the different kinds of inferential tests included in this book appears towards the end of this chapter.

3. Determine the critical value.

The critical value refers to the number you must surpass in order to conclude that your results are unlikely to be due to chance and, thus, likely reflects a truth about the population. The critical value is a concept we will discuss in more detail in subsequent chapters. For now, know that we weigh the implications of an inaccurate conclusion (e.g. what are the risks of concluding our medication worked when it actually did not) and then set the statistical risk we are willing to take that we might be wrong (which is used to determine the critical value). In the behavioral sciences, we very often decide that we are willing to accept less than a 5% chance that we will conclude a hypothesis is true when it is not; this means we want less than a 5% chance that our result is simply a false positive. Thus, critical values are usually computed to represent the amount (or strength) of evidence that is needed to be at least 95% confident that the hypothesis is true.

4. Calculate the test statistic.

This is the step of the scientific method (and, thus, also in the process of hypothesis testing) in which data are analyzed. In this step, the statistician uses the inferential test that was chosen in step 2 to analyze the data and yield a result. The result is represented by the obtained value (which is also known as a test statistic or result). This is the most math-intensive step of testing a hypothesis.

5. Apply a decision rule and determine whether the result is significant.

In this step, we assess whether our result (i.e. our obtained value or test statistic) exceeds the critical value. When it does, we can conclude that there is a strong probability that the hypothesis is true in the population based on the evidence observed in the sample. In so doing, the result is concluded to be significant. Conversely, when the test statistics does not exceed the critical value, we conclude that the evidence is not strong enough to conclude that the hypothesis is likely true in the population and, thus, that the hypothesis is not supported. In so doing, the result is concluded to be non-significant.

> **Note**
>
> When a result is close to exceeding the critical value but does not, it may be prudent for researchers to retest the hypothesis or similar hypotheses with new samples in the future. A result which is close to, but does not surpass, the critical value may be referred to as “trending”; however, trending results should not be referred to as significant.

When it is determined that the result is significant, proceed through each of the remaining steps. When it is determined that the result is not significant, skip to step 7 to complete the process of hypothesis testing.

6. Calculate the effect size and other relevant secondary analyses.

An effect size can be reported alongside a significant result. Essentially, an **effect size** is an estimate of the magnitude of an effect, change, or pattern observed in the sample data. Effect size can help statisticians and audiences deduce practical significance. **Practical significance** refers to whether there is a large enough magnitude of effect to be meaningful or useful. This is important because it is possible to have a result that is statistically significant without being practically significant. Thus, it is often recommended that practical significance be reported as a secondary analysis when a result is statistically significant.

Some tests have additional secondary analyses which are necessary to adequately test a hypothesis. In each chapter for which these are recommended, they will be included in the section for step 6 of hypothesis testing.

7. Report the results in American Psychological Associate (APA) format.

Results for inferential tests are often best summarized using a paragraph that states the following:

a. The hypothesis and specific inferential test used,
b. The main results of the test and whether they were significant,
c. Any additional results that clarify or add details about the results, and
d. Whether the results support or refute the hypothesis.

It is recommended that effect sizes be reported with the additional results, when possible and/or common in the field into which the researcher is disseminating results. Dissemination refers to the formal sharing of results which is often done through publishing peer-reviewed, empirical articles in research or academic journals, giving conference presentations, and/ors reporting results in books that focus on summarizing several empirical studies. APA format specifies the level of rounding and types of symbols which should be used when reporting results for each of the various descriptive and inferential tests.

We will employ these steps when we learn how to select and properly use inferential statistics to test hypotheses in the subsequent chapters of this book. In each of those chapters, the details of the formulas, the calculations they require, and the symbols and rounding rules will be covered in detail. It will likely be helpful to refer back to this section with each of those chapters to remind yourself of the order and purpose of each of these steps to testing a hypothesis and reporting the results.

Reading Review 6.3

1. In which step of hypothesis testing are data analyzed?
2. What does statistical significance mean?
3. Which two values are compared to determine whether a result is statistically significant?
4. What is used to estimate practical significance?
