6 Statistical terminology and one-sample tests

This final chapter of Part II serves as a bridge between descriptive and inferential statistics. We begin by defining a number of statistical terms and concepts. These terms are important in describing research designs and understanding statistical analysis, so comprehension of this vocabulary will help to establish the conceptual foundation for inferential statistics. Discussion of these terms is necessarily limited; we instead aim to provide a working vocabulary and a brief synopsis of the major ideas for quick reference should the need arise. Several of the presented concepts, such as effect size and confidence intervals, are treated in greater detail in subsequent chapters.

The one-sample tests at the conclusion of this chapter provide easy-to-understand examples of how to conduct inferential statistics. The tests themselves are of modest utility compared to the more powerful and commonly-employed tests later in the book. However, we describe these basic tests in a step-by-step manner to clarify the process of hypothesis testing with these examples, in order to build the foundational knowledge that will allow for learning more complex tests.

Introduction to statistical terminology

In Chapter 4 we covered descriptive statistics, which are used to summarize sample data. In Chapters 7–12, we will explain the application of inferential statistics, which attempt to reach conclusions about the underlying population. In other words, the goal of these analyses is to generalize the findings from the observed sample data as a way to understand the larger population from which the sample was drawn. Statistics provides the tools and structure for making those inferences, and this section is devoted to a minimal set of definitions that will assist in understanding that process. Some of these terms have already been introduced and rough definitions suggested; however, this section provides more formal definitions along with some examples and discussion of the various terms.

Definition of a statistic

The formal definition of a *statistic* (as opposed to the discipline of statistics) is any function of the sample data. Therefore, a statistic is a numerical value that is calculated from measurements in the collected sample. The calculation can involve all of the sample data, in the way that the sample mean consists of adding all of the observations and dividing by the sample size. The calculation can involve fewer pieces of data, however, such as the median, which uses only one or two of the observations, depending on whether the sample size is odd or even. A statistic might also simply be one data point such as the maximum observation.

A statistic can serve three purposes. First, it can serve as a descriptive statistic of the sample. For example, the sample mean serves as a descriptive measure of center. The use of various descriptive statistics was covered in Chapter 4. The remaining two purposes are needed for inferential statistics. The second use of a statistic is as an estimator, in which the number provides an approximation of the unknown population parameter. For example, the sample mean is an estimator for the unknown population mean. The next subsection of this chapter is devoted to the topic of parameter estimation. The third application of a statistic is for hypothesis testing, in which the calculated figure is used to compare the sample data to a theoretical distribution. This comparison is made in order to determine whether to reject the null hypothesis. Hypothesis testing is described more fully below.

Parameter estimation

One purpose of inferential statistics is *parameter estimation*. A *parameter* is an unknown property of a population that is estimated by the collection of sample data and calculation of a statistic. The parameter thus serves as the target quantity when thinking in terms of validity and reliability. A parameter is considered fixed for a given population at a given time. Conversely, a statistic varies based on several sources of error. The reasons for variability include sampling error and measurement error. Theoretically, if an entire population could be measured simultaneously (a *census*), then the parameter could be known with certainty.

An example of a parameter is the average height of all the adult women in Norway. At any moment in time, this is a fixed number. If we could measure every single woman in the country simultaneously, we could calculate the true average height of the population. Of course, this data collection process is impossible; by the time we gathered and processed the data, the population would have changed (e.g., people grow or immigrate) so that the number would no longer be the true parameter. To overcome this challenge, we collect a random sample and use the average height of the sample as an estimate of the population parameter.

Just like the previous height example, we could hypothetically calculate the average hourly wage of every person in the country to obtain a single average. This number would encompass all of the variability that exists with respect to professions, years of working experience, and any other variable impacting how much a person earns. We could do the same if we were interested in the average wage of a certified court or medical interpreter in a given country. However, all of these scenarios are impractical, and we would instead calculate an estimate. Data could be collected from a random sample of interpreters by means of a survey on wage information in order to estimate this unknown parameter.

Thinking of a parameter as a fixed goal of estimation is appropriate only after careful definition of the population. A parameter is fixed only for a well-defined population at a given moment in time. The parameter of average wage of court interpreters in one country is different from the average wage of court interpreters in another country. Moreover, parameters can and will change over time. The quality of a parameter estimate also depends on the quality of the sample, the validity and reliability of the estimator, and various other errors that can introduce variability. One of the main tasks of statistics is to provide high-quality parameter estimates.[1]

Confidence intervals

Parameter estimation results in a single number estimate, called a *point estimate*. This estimate can be thought of as a best guess approximation of the population parameter. However, different samples will result in different point estimates; there remains uncertainty about how far our sample estimate lies from the population parameter. Rather than estimating with a single number, therefore, it is possible to construct an interval estimate called a *confidence interval* around the point estimate. The goal is to define the confidence interval in such a way that the parameter lies within the interval.

If we think again about calculating the average height of women in Norway, one sample of 30 women might give an estimate of 165 cm. If we repeat the experiment with a new sample of 30 different women, we might get a sample mean of 169 cm instead. Theoretically, we could repeat the process over and over again: collecting a sample of 30 women, calculating the sample mean, and constructing a confidence interval around the mean. In every case, the confidence interval would be a particular function of the sample data. Therefore, in each case the confidence interval either contains the parameter or it does not. By repeating the process many times and by constructing the confidence intervals in a particular way based on probability theory, we can predict how often the confidence interval contains the parameter.

In almost all cases, confidence intervals are constructed for 95% confidence and symmetrically around the point estimate. Therefore, the formula for the endpoints of a confidence interval can be compactly written as the point estimate plus or minus a distance determined by the sample's variability and a number from the appropriate probability distribution. Examples that accompany each statistical test in this volume will clarify the exact procedures necessary, and most software packages provide appropriate confidence intervals automatically around point estimates. For the primary tests of a research project, a confidence interval should always be reported to improve interpretation beyond the simple point estimate.

The best way to think of a confidence interval is as an interval estimate based on the sample data. Imagine you were offered a prize if you could guess how many marbles were in a jar. The number of marbles in the jar is a parameter, a fixed but unknown trait. If you were allowed only to guess one number, say 210, you would be very unlikely to be correct. However, if you could guess an interval, perhaps guessing that the number of marbles was between 170 and 250, then you would be more likely to win the prize. The estimate would be more likely to include the parameter. Of course, your guess is either right or wrong; you are not 95% likely to win the prize. A 95% confidence interval means that if the sampling and calculation process was repeated an infinite number of times, then 95% of the confidence intervals would contain the true parameter. Over-interpretation beyond that should be avoided.

Hypothesis testing

Null and alternative hypotheses were introduced in Chapter 1. The two hypotheses represent different models or structures of the underlying population. In the process of hypothesis testing, the null hypothesis is assumed to be true unless the sample data provide convincing evidence otherwise. In that case, the sample data would suggest that the alternative hypothesis is a better description of the sample data. For instance, a very simple statistical model might have a null hypothesis claiming that the average hourly wage of a conference interpreter is $1,200. Obviously any collected data would reject this null hypothesis as a poor model of the actual hourly wage of conference interpreters (though it would be nice!).

The process of conducting a hypothesis test involves the determination of two numbers. First, a test statistic is calculated from the sample data. The test statistic could be the same as a descriptive statistic, such as the variance, or a statistic with a more complicated formula. Second, a critical value is found from a theoretical distribution, such as the F-distribution. Then, the magnitude of these two numbers is compared. The comparison leads to a decision about the null hypothesis. Test statistics that are more extreme (more negative or more positive) than the critical value lead to rejection of the null hypothesis. This section describes various terms related to hypothesis testing and closes with a discussion of some criticisms, strengths, and weaknesses of the procedure.

Errors, power, and p-values

In reaching the decision of a hypothesis test, two different errors could occur. Type I error describes the situation in which the null hypothesis is rejected, even though it is true. A Type I error corresponds with a false positive, wherein the sample suggests a relationship or a difference exists, even though no such effect is present. In contrast, a Type II error describes the opposite situation, in which the test procedure fails to reject the null hypothesis, even though it is false. The probability of a Type I error is denoted with the Greek letter alpha, α, and the probability of a Type II error is denoted with beta, β.

The simile of a Type I error as a false positive is often clear in a medical context. If a person tests positive for a disease they do not have, they might be subjected to unnecessary treatments. If a pill is incorrectly proclaimed to be able to cure a disease, many people might waste time and money (not to mention risking side effects) with no benefit. In translation and interpreting (T&I) studies, a simple corresponding example might be pedagogical studies. In that setting a Type I error would be an announcement that a new method of instruction or studying would increase the speed of language acquisition, when in fact, there is no measurable benefit when applied to the whole population. A Type I error provides false information that can lead to bad results, wasted resources, and misdirected research efforts based on incorrect decisions.

The level of Type I error is set in advance by the researcher by choosing a significance level for the statistical test. In every statistical test in this volume, the significance level is set at 5%, which is a nearly universal standard in applied social science research. The statistical testing procedures are designed to maintain this chance of rejecting the null hypothesis when it is true (committing a Type I error). Keep in mind that the level is arbitrary and that interpretation of statistics needs to include more than one piece of information.

A Type II error is the failure to reject an incorrect null hypothesis. This type of mistake would rarely result in any announcement of the findings or action taken, so a Type II error is considered less severe than a Type I error. When the level of Type II error is subtracted from one, the resulting number is referred to as *statistical power*. So a 20% probability of Type II error implies 80% statistical power. Power is the probability of correctly rejecting the null hypothesis when it is false. The level of power is not directly controlled by the researcher, but larger sample sizes always lead to increases in power. Balanced designs that have the same number of participants in each group also generally increase power. Finally, several different statistical procedures will exist for testing any given research design; selection of the best test is often driven by power considerations.

Power considerations should always occur before any statistical procedures are conducted. Estimating a minimally meaningful effect, selecting the best test statistic, and determining the necessary sample size are all related to statistical power and should be part of the planning stage of any research project (Lenth 2001). Post hoc power considerations are not useful and do not convey any additional information beyond the *p*-value and confidence interval (Hoenig and Heisey 2001; Colegrave and Ruxton 2003).

When a test statistic is calculated, the result can be compared to a theoretical probability distribution. The distribution implies the likelihood of observing that particular result. An important probability in this context is called the *p*-value. In calculating the *p*-value, it is assumed that the null hypothesis is true, and the probability distribution of the test statistic is developed under that assumption. Then the computed test statistic from the sample is compared to the distribution. The probability of getting a more extreme test statistic (larger in absolute value) is the formal definition of the *p*-value. Therefore, the *p*-value represents the probability of an even more unusual result than the particular sample. If that probability is low, meaning less than 5%, then the sample would appear to violate the assumption that the null hypothesis is true. Thus, the null hypothesis would be rejected.

For example, we might compare the means of two groups with a null hypothesis that they are the same. When a difference is observed, the appropriate question is whether the difference is large enough to constitute a significant difference or whether the difference is small and due simply to random experimental error. Thanks to modern statistical computing software, a *p*-value provides a fast way to make a decision regarding a statistical test. If the *p*-value is less than .05, the null hypothesis is rejected. Otherwise, there is not enough evidence to reject the null hypothesis at the 5% level of significance.

A result should be declared either statistically significant or not statistically significant only at the pre-determined level. The significance level is almost always 5% in social science, but this level is traditional and arbitrary. The determination that a result is meaningful should be provided by a measure of effect size,[2] reporting of the descriptive statistics, and a discussion of how the results fit into the larger body of scholarship on the particular topic. Strictly describing *p*-values in terms of "statistical significance" rather than just "significance" can help prevent this misunderstanding.

The precise numeric value should not be over-interpreted.[3] An all too common, although incorrect, practice is attempting to use a *p*-value to demonstrate the importance of a statistical result. Asterisks sometimes appear in tables for tests that result in different levels of *p*-values with an implication that smaller p-values are somehow more significant. Rasch, Kubinger, Schmidtke, and Häusler (2004) argue strongly against this practice, despite its wide adoption by statistical software and its use in many publications. An equally erroneous practice is the suggestion that a *p*-value larger than .05 somehow "approaches" significance or is "nearly significant." A final caution regarding *p*-values is that simple comparisons between them are not appropriate (Gelman and Stern 2006). This issue is discussed further in the final chapter of the volume on the topic of reporting.

Degrees of freedom

For many people who have survived an introductory statistics class, *degrees of freedom* (often abbreviated *df* in reporting) is a poorly understood concept that amounts to rote formulas needed for certain statistical tests and distributions. We cannot hope to provide a complete formal understanding in a brief treatment, but a definition in simple language and some examples will hopefully clarify the main intent. Degrees of freedom in general can be thought of as how many additional facts or pieces of data are required so that everything about the relevant variable is known. For example, imagine that you are told that a sample contains 50 people, who all speak either Mandarin or Korean as their primary language. If you are then also told that 35 people in the sample speak Mandarin, then it is automatically known that the remaining 15 people speak Korean. This example has only one degree of freedom because learning one piece of information (i.e., that 35 people speak Mandarin) allowed for complete knowledge about the distribution of language in the sample.

We can extend this example to multiple languages. A sample of 100 citizens of Switzerland could contain people who have as their L1 one of its four official languages: German, French, Italian, or Romansh. If it is known that 38 people have German as their first language, 27 French, and 23 Italian, then it automatically follows that 12 people speak Romansh as their L1. Knowing three pieces of information provide complete information about the variable in this case. Consequently, there are three degrees of freedom. This pattern that the degrees of freedom are one less than the number of categories holds true in many cases, particularly for analysis of variance (ANOVA; see Chapter 8). In many situations the degrees of freedom are one less than the sample size, based on a similar argument.

A bit more formally, we mentioned in Chapter 4 that the denominator in the sample variance calculation (n–1) is the number of degrees of freedom for the estimate. The reason is that the normal distribution has only two parameters: the mean and the variance. In order to calculate the variance, the mean must first be estimated, and this estimation creates a restriction in the relationships among the sample data. Therefore, another way of thinking about degrees of freedom is that you start with the sample size and subtract one degree of freedom for every estimated parameter. This way of thinking aligns well with regression design and more complicated ANOVA models.

The reason degrees of freedom matter when conducting inferential statistics is that they determine the shape of the theoretical distribution that is used for finding the critical value. In particular, the t-distribution and the χ^2-distribution have different shapes based on the related degrees of freedom, and the F-distribution requires two different measures of degrees of freedom, called the numerator and denominator degrees of freedom. For the non-mathematically inclined, the degrees of freedom are usually easy to find in computer output and should always be reported in parentheses or brackets behind the name of the distribution in any research report. For instance, if a t-statistic was 1.85 with 24 degrees of freedom, it would be reported as $t(24) = 1.85$.

Residual errors

Hypothesis testing always involves the consideration of statistical models to describe data. Once a model has been estimated, the predictions of the model can be compared to the observed data. The difference between the predicted values and the actual values are referred to as *residuals* or *residual errors*.

The sample mean is an example of one of the simplest statistical models. If the average height of American men is approximately 177 cm, then the best guess for any randomly selected man is that he will be 177 cm tall. If a randomly selected man is 180 cm tall, then the residual error for that observation is 3 cm. More complicated models can make better predictions using more information. For instance, a regression model could use a person's gender, weight, shoe size, and age to predict his or her height. The model's equation would make a prediction that could be compared to sample data and the residual would be the difference between the prediction and the observed measurement.

Residuals are useful for checking the adequacy and accuracy of statistical models. For most models, the residuals should be random and approximately follow the normal distribution. Obvious patterns or outliers in the residuals can suggest problems with the chosen statistical model. Therefore, we will discuss residual errors often in Chapters 7 through 12.

Adjustments to hypothesis tests

The validity of null hypothesis significance testing (NHST) always rests on a set of assumptions about the sample data and the underlying population. In practice, these assumptions are never fully met, only approximated. Furthermore, a research project often involves conducting multiple tests with the same set of data. Statistical adjustments can be made in an attempt to correct for these issues.

Any adjustments made to NHST procedures are typically made in the service of two goals: controlling the probability of a Type I error at a specified level (usually 5%) and maximizing the power of the test (which is equivalent to minimizing the probability of a Type II error). Adjustments can be made in three ways. First, the computation of the statistic itself can be changed; taken to the extreme, a different test statistic can be used. Instead of the mean, we can use the trimmed mean or the median, for example (Wilcox 1995). Second, the degrees of freedom can be computed differently. Lowering the degrees of freedom is one way to acknowledge a greater degree of uncertainty, thereby creating a more conservative test in order to reduce the chance of a Type I error. Third, the *p*-value cutoff can be reduced below a nominal 5% level. The second method corresponds to Welch's adjustment to the *t*-test and the third method to the Bonferroni correction, both of which are discussed in Chapter 7.

Effect size

NHST can provide information on the probability of observing a given result, but *effect sizes* provide an understanding of the strength and practical impact of a difference or relationship. Ellis (2010: 5) describes this difference between two types of significance, by noting that "[p]ractical significance is inferred from the size of the effect while statistical significance is inferred from the precision of the estimate." Although the shortcomings of NHST have been debated for decades, only in recent years has the reporting of effect sizes come to the forefront as a primary supplement to statistical testing. The *Publication Manual* of the APA (2009: 33–34) now emphasizes the importance and need for reporting of effect sizes to convey the full results of a study. However, the published literature still lags in providing effect sizes (Ferguson 2009; Fritz, Morris, and Richler 2012). Even rarer is the reporting of a confidence interval for the effect size (Algina and Keselman 2003). In our description of test statistics, we focus on raw effect sizes with only occasional references to confidence intervals surrounding them. Applying the lessons of confidence intervals, however, would improve interpretation, especially for smaller effect sizes.

The reporting of effect sizes communicates the practical impact or meaningfulness of a study's results. Therefore, an effect size should be included for every statistical test, whether or not it is significant. Knowing the relative importance of variables assists in the accumulation of knowledge and theory development (Lakens 2013).

There are three primary ways to report effect sizes. The simplest is to use the same units of measurement as the study and describe the actual difference. For instance, a study could report that post-editing MT output was, on average, 5 minutes faster (95% CI [2.5, 7.5 minutes]) than human translation of the same passage. The advantage of this approach is that it communicates the relevant difference in meaningful terms. However, the original units do not allow for comparison to other studies, so the two alternative approaches involve standardized effect sizes.

Standardized effect sizes come in two primary types, known as the *d* family and the *r* family. Although almost any test can be described by an effect size from either group, the *d* family measures the standardized difference between groups, so it is appropriate for the tests of difference that we report in Part III. Meanwhile, effect size in the *r* family focuses on the strength of relationships and are most appropriate for the tests in Part IV. For each of the tests we discuss, we include a section on the relevant effect size. For a more thorough but approachable treatment, see Ellis (2010).

Parametric and nonparametric tests

The distinction between *parametric* and *nonparametric* tests lies principally in the assumptions that are made about the underlying population. To begin with, a parametric test typically assumes that the variable of interest follows a normal distribution in the population. Nonparametric tests make fewer assumptions (though, it should be noted, they are not assumption-free). In particular, nonparametric tests do not assume normality of the sample data.

The two reasons for preferring nonparametric tests in certain cases is to maintain control of the probability of a Type I error and to increase statistical power. Whenever the assumptions of a parametric test cannot be met, there is generally a nonparametric procedure available that will meet these two criteria. In this volume, we present nonparametric tests side-by-side with their parametric counterparts, rather than relegating them to their own chapter, as is common in previous books.[4] The smaller sample sizes and unknown population distributions that characterize much of T&I research suggest that nonparametric procedures deserve a more prominent placement and wider adoption.

With large enough sample sizes, a claim is often made that the Central Limit Theorem (CLT) provides assurance of approximate normality. Therefore, nonparametric tests are used primarily for experiments with smaller sample sizes. Nonparametric tests are valid for any sample size, but their advantage in terms of power and Type I error are generally slight with larger sample sizes. We stress that the decision between parametric and nonparametric methods should be based on an assessment of the assumptions. Both can be valid for small or large sample sizes, but generally nonparametric tests are more common for sample sizes smaller than 40. This number serves as a guideline only, and the final decision should be multi-faceted.

Reporting results

The treatment of the terminology covered in this chapter serves two purposes. First, knowing the definitions will help in understanding the presentation of the statistical tests in later chapters. Second, the headings should serve as a checklist for reporting the results of statistical analysis. Reporting the results of the main test procedures of any research paper should include the following:

1. An estimate of the population parameter of interest, including a confidence interval;
2. A description of the statistical test procedure, including whether it was parametric or nonparametric and the reason that the test's assumptions are adequately met;
3. The results of the statistical test, including the value of the test statistic, its degrees of freedom (if appropriate), and its exact *p*-value;
4. A measure of effect size.

Descriptive statistics, graphs, and tables should also be included when they improve interpretation. All of this information can be concisely reported in a few paragraphs at most, as we will illustrate in later chapters of this book. The responsibility of the researcher is to interpret these numbers in a meaningful way in terms of the research hypothesis.

One-sample test for the mean

This section will demonstrate the procedure for the statistical test of the value of the mean. In doing so, we will illustrate many of the terms described so far in this chapter and transition from preparing and describing data to testing and making inferences, topics that dominate Parts III and IV. To test whether the mean value equals a numerical constant we employ a one-sample *t*-test.[5] The null hypothesis is that the mean of the underlying population equals a specified number, and the two-sided alternative is that the mean does not equal that number.

The one-sample *t*-test examines a simple research question. For instance, we might investigate whether the average time needed for to translate a 500-word passage was 120 minutes for a sample of 100 translators. The null and alternative hypotheses would be the following: $H_0\colon \mu = 120$ and $H_1\colon \mu \neq 120$. We generated fictional data for this situation and calculated the following descriptive statistics: $M = 122.6$ minutes, $SD = 10.9$ minutes, with observations ranging from 101.6 minutes to 156.8 minutes.

The one-sample *t*-test assumes that the sample data are continuous and drawn from a population that is normally distributed. In this case, given a fictitious sample size of 100, we can argue that the CLT promises approximate normality. Additionally, the data must be independently collected. Other tests have more restrictive assumptions, but this particular test is used in this chapter precisely because it is a simple test for introducing the general procedure.

The appropriate test statistic is built on the sampling distribution of the sample mean (see Chapter 5) and is calculated by dividing the difference in the estimated and hypothesized mean by the standard error:

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}} = \frac{122.6 - 120}{10.9 / \sqrt{100}} \cong 2.37$$

This test statistic follows a *t*-distribution with 99 degrees of freedom (one less than the sample size). The appropriate critical value for this situation is 1.98; therefore, because the test statistic exceeds the critical value, the null hypothesis can be rejected.

When the test is conducted with statistical software, a *p*-value will also be provided. In this case, the *p*-value is .0197, which is less than the 5% cutoff. Notice that the decisions based on the critical value or on the *p*-value will always be identical, but rejection occurs when the test statistic exceeds the critical value and, equivalently, when the *p*-value is below the 5% level.

The point estimate of the population mean is simply the sample mean, and a confidence interval can be built around the sample mean with the following formula:

$$\bar{X} \pm t_{crit} * \left(\frac{s}{\sqrt{n}} \right)$$

The critical value from the *t*-distribution appears in this formula. For the example data, the 95% confidence interval would result in the following:

$$122.6 \pm 1.98 * \left(\frac{10.9}{\sqrt{100}} \right) \cong [120.42, 124.76]$$

The final step is to calculate the effect size. Because we are making a comparison of the degree that a sample potentially differs from a hypothesized mean, we use the formula for Cohen's *d*:

$$d = \frac{\bar{X} - \mu}{s} = \frac{122.6 - 120}{10.9} = .239$$

A more substantial discussion of Cohen's *d* appears in Chapter 7. For now, it is sufficient to understand that .239 is a rather small effect size.

We have now completed all of the necessary calculations to report the results of this experiment. In practice, most of them would be conducted by statistical software, but effect sizes in particular often require some hand calculation. Furthermore, knowing the process allows for better interpretation of computer output. Complete reporting would look something like that shown in Figure 6.1.

To test whether the average time-on-task was 120 minutes, a one-sample *t*-test was conducted. Results imply that the average time is greater than the hypothesized 120 minutes (*M* = 122.6, *SD* = 10.9, 95% CI [120.42, 124.76], *t*[99] = 2.37, *p* = .02). However, the effect size was small (Cohen's *d* = .239). The observed mean time exceeded the hypothesized time by only 2.6 minutes.

Figure 6.1 Reported statistics for mean time on task example

This reporting includes all of the required information listed above. The purpose is to allow the reader to understand not only the statistical significance but the practical impact of the study. The particular example here is statistically significant. However, with a small effect size and a 95% confidence interval that nearly includes the hypothesized mean of 120 minutes, the meaningfulness of the results would be rather low. Further discussion of the results and how they relate to the literature would appear in the discussion and conclusion of the research report.

One-sample test for the median

Our first example of a nonparametric test is the *Wilcoxon signed-ranks test* (1945) for the median, also called simply the *signed-ranks test*. The procedure is relatively simple and introduces some of the most common issues related to nonparametric statistics. Assume that a professor assigns a translation exam to 100 students but only takes 10 of the completed exam papers home to grade and forgetfully leaves the rest in her office. She grades the 10 exams and observes the following scores:

55, 64, 69, 71, 74, 82, 85, 86, 89, 98

She now wants to determine whether the class median is 70. The sample median is 78 (the average of 74 and 82). The statistical test can determine if this difference is unusual or extreme enough to be statistically significant or if the difference between 70 (the hypothesized median) and 78 (the sample median) is due to random error.

As with all statistical tests, the first step is to check that the data meet the assumptions. The signed-ranks test has relatively few assumptions, and they are easily met:

1 The sample is randomly drawn from the population of interest;
2 The population is symmetric around the median for the variable of interest;
3 The variable of interest is continuous and measured at least at the interval scale;
4 The observations are independent.

The null hypothesis of the test is that the median is equal to some specified value, and the most common alternative hypothesis is the two-tailed version that the median is not equal to the specified value. In mathematical notation, $H_0 : M = c$ and H_1: $M \neq c$ where c represents some number. In the example, $c = 70$, the hypothesized median score.

The test statistic is completed in four steps. First, the hypothesized median is subtracted from every observation. Second, the absolute values of the differences are ranked. The smallest value is given a rank of one with any differences of zero ignored. If ties occur, all of the tied observations receive the same rank, which is the mean of the rank positions they would have occupied. This rank transformation procedure is very common in nonparametric statistical tests. The ranking eliminates the outliers and effectively moves the data from a higher level of measurement to the ordinal level of measurement.

Third, each rank is assigned the same sign as the associated difference score. Fourth, the ranks of the positive and negative scores are summed separately and labeled T^+ and T^-. The smaller of these two sums is the final test statistic, denoted T.

The procedure sounds more complex than it is in practice. Like many nonparametric procedures, the work can be completed quite easily by hand or with Excel. Of course, most statistical software packages will also do these calculations automatically. Table 6.1 illustrates these steps for the sample exam scores.

The first column displays the raw scores. The "Difference" column is the raw scores minus the hypothesized median value of 70. Notice that the data have been ordered according to the absolute value of this difference. The "Ranks" column provides the ranks of the differences, with ties recorded as the average of the associated ranks. The "Signed ranks" column contains the same number as the "Ranks" column with the same sign (positive or negative) as the "Difference" column. Finally, the two sums are calculated. The smaller sum is 12, and this value is the final T-statistic for the test.[6]

The final decision regarding the null hypothesis requires comparing this value to a table of critical values to obtain a critical value and/or p-value. Published tables are available for sample sizes up to 30. For a sample size of 10, the critical value of the two-tailed test is 8. Test statistics that are less than 8 result in rejection of the null hypothesis. Since the calculated test statistic in the example is 12, we

Table 6.1 One-sample test for the median example data

Scores	*Difference*	*Ranks*	*Signed ranks*
69	−1	1.5	−1.5
71	1	1.5	1.5
74	4	3	3
64	−6	4	−4
82	12	5	5
55	−15	6.5	−6.5
85	15	6.5	6.5
86	16	8	8
89	19	9	9
98	28	10	10
	Sum of positive ranks: T^+		43
	Sum of negative ranks: T^-		12

cannot reject the null hypothesis. There is not enough evidence that the population's median is different from 70.

The remaining statistics for the signed-ranks test demand more complicated formulas. Therefore, we will omit their details and refer the reader to Daniel (1990) for general details. Of course, most applied researchers will rely on statistical software. Output from the program R provides a 95% confidence interval of [67.5, 87.0] and a *p*-value of .131.

The effect size will often need to be calculated by hand until statistical software begins to more regularly incorporate such estimates into its procedures. Kerby (2014) provides a method for calculating an effect size measure in the *r* family. First, calculate the total sum of the ranks for the given sample size: $S = \frac{n\left(n+1\right)}{2}$ and then use that figure in the final calculation:

$$r = \frac{\left(S-T\right)}{S} - \frac{T}{S}$$

For our sample, $S = \frac{10*11}{2} = 55$ and $r = \frac{55-12}{55} - \frac{12}{55} = .56$. As we mentioned previously, the interpretation of effect sizes is discussed further in later chapters, but an *r* value of .56 would be considered a large effect. Notice that the null hypothesis could not be rejected but the effect size was large. The implication is that further research is needed. Replication with a larger sample size or better experimental controls will likely lead to statistical significance. However, the practical impact of the difference can also be interpreted directly without the need for statistical significance. The relevant question is whether a class median of 78 instead of the hypothesized value of 70 is meaningful in this case.

The nonparametric test described here exhibits many common features, beginning with the rank transformation of the data. Complications in confidence intervals and *p*-values are also common. In some cases, nonparametric tests employ large-sample approximations and correction factors for tied observations. These corrections complicate the numerical procedures but not the final interpretation. It always remains the case that *p*-values less than 5% imply rejection of the null hypothesis. However, confidence intervals and effect sizes are also necessary and arguably more important than the simple decision of whether to reject the null hypothesis. Finally, reporting of the results (see Figure 6.2) would be similar to the parametric case:

To test whether the median test score was 70, a one-sample Wilcoxon signed-ranks test was conducted. The observed median (78) exceeded the hypothesized value. However, the results were not statistically significant (*T* = 12, *p* = .131, 95% CI [67.5, 87.0]). The effect size was large (*r* = .56), which suggests that the difference may have practical meaning for this application.

Figure 6.2 Reported statistics for median test score example

Notes

1 In theoretical statistics, estimators are evaluated based on a set of criteria that includes *unbiasedness*, *efficiency*, and *consistency*. For reasons of space, we must omit formal definitions of these and other properties. However, a significant portion of theoretical statistics has historically been devoted to the creation of statistics that possess desirable properties. Communicating some of the results of ongoing development in the field is one motivation for this volume.

2 Effect sizes and *p*-values are mathematically related. However, the sample size plays a role in the calculation of each, so a small effect can be significant and a large effect can be nonsignificant. For this reason, researchers should always report both an exact *p*-value and a measure of effect size.

3 See Nickerson (2000) for a comprehensive discussion of interpretation fallacies related to *p*-values.

4 Space considerations allow for the selection of a limited number of nonparametric tests and relatively brief treatments. For a more comprehensive overview see Daniel (1990).

5 Some introductory statistics books and classes also teach the similar *z*-test, which can be used when the standard deviation is known. Since this will almost never be the case, we omit a description of this test.

6 There are a number of different ways to combine the final ranks. Depending on the procedure, the test statistic is sometimes denoted by *W* or *V*. Also, take careful notice that this *T*-statistic does not follow Student's *t*-distribution.