10

THE ONE-SAMPLE *z* TEST

Only the Lonely

Difficulty Scale ☺ ☺ ☺
(not too hard—this is the first
chapter of this kind, but you know
more than enough to master it)

WHAT YOU WILL LEARN IN THIS CHAPTER

- Deciding when the *z* test for one sample is appropriate to use
- Computing the observed *z* value
- Interpreting the *z* value
- Understanding what the *z* value means
- Understanding what effect size is and how to interpret it

INTRODUCTION TO THE ONE-SAMPLE *z* TEST

Lack of sleep can cause all kinds of problems, from grouchiness to fatigue and, in rare cases, even death. So, you can imagine health care professionals' interest in seeing that their patients get enough sleep. This is especially the case for patients who are ill and have a real need for the healing and rejuvenating qualities that sleep brings. Dr. Joseph Cappelleri and his colleagues looked at the sleep difficulties of patients with a particular illness, fibromyalgia, to evaluate the usefulness of the Medical Outcomes Study (MOS) Sleep Scale as a measure of sleep problems. Although other analyses were completed, including one that compared a treatment group and a control group with one another, the important analysis (for our discussion) was the comparison of participants' MOS scores with national MOS norms. Such a comparison between a sample's mean score (the MOS score for participants in this study) and a population's mean score (the norms) necessitates the use of a one-sample *z* test. And the researchers' findings? The treatment sample's MOS Sleep Scale scores were significantly different from normal (the population mean; $p < .05$). In other words, the null hypothesis that the sample average and the population average were equal could not be accepted.

So why use the one-sample *z* test? Cappelleri and his colleagues wanted to know whether the sample values were different from population (national) values collected using the same measure. The researchers were, in effect, comparing a sample statistic with a population parameter and seeing if they could conclude that the sample was (or was not) representative of the population.

Want to know more? Check out . . .

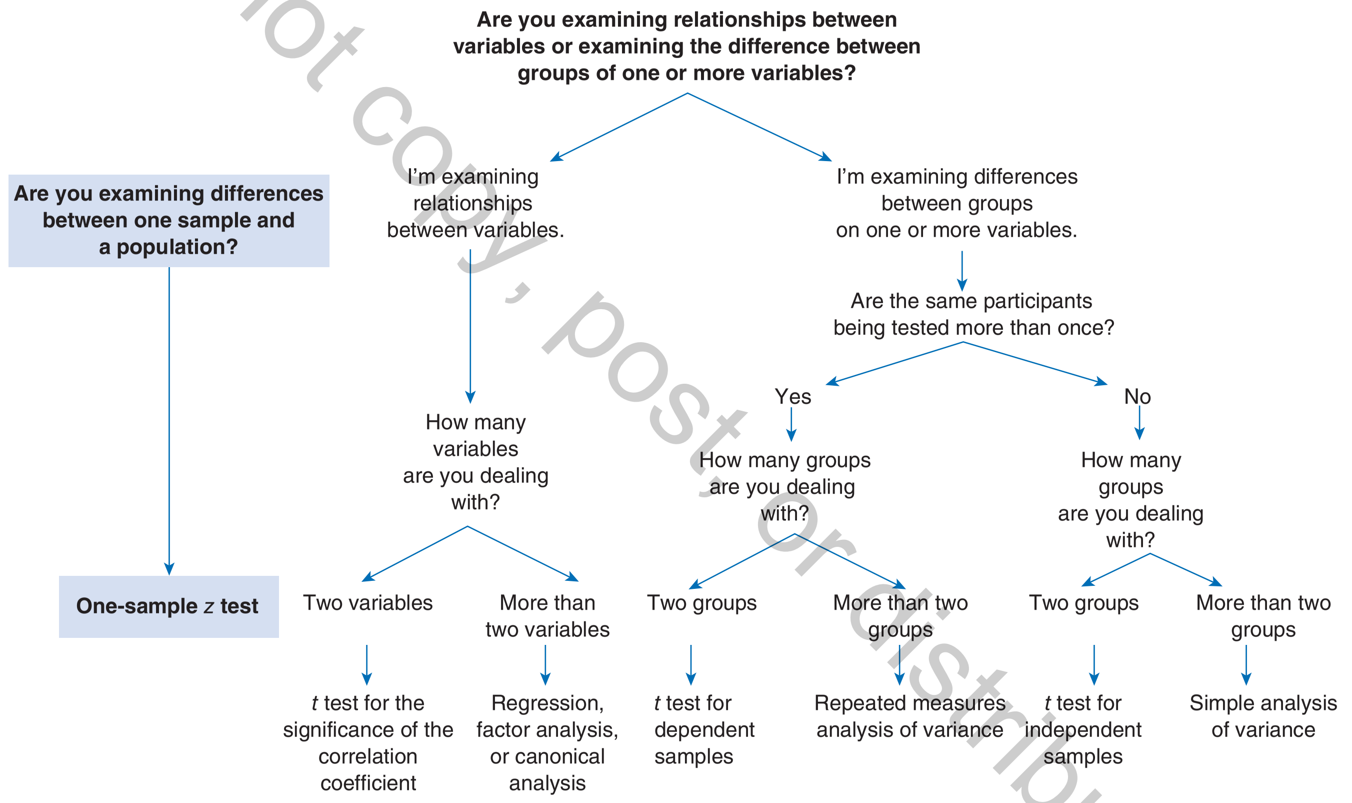
Cappelleri, J. C., Bushmakin, A. G., McDermott, A. M., Dukes, E., Sadosky, A., Petrie, C. D., & Martin, S. (2009). Measurement properties of the Medical Outcomes Study Sleep Scale in patients with fibromyalgia. *Sleep Medicine, 10,* 766–770.

THE PATH TO WISDOM AND KNOWLEDGE

Here's how you can use Figure 10.1, the flowchart introduced in Chapter 9, to select the appropriate test statistic, the **one-sample *z* test**. Follow along the highlighted sequence of steps in Figure 10.1. Now this is pretty easy (and they are not all this easy) because this is the only inferential *comparison* procedure in all of Part IV of this book where we have only one group. (We compare the mean of that one group to a theoretical invisible population.) Plus, there's lots of stuff here that will take you back to Chapter 8 and standard scores, and because you're an expert on those. . . .

1. We are examining differences between a sample and a population.
2. There is only one group being tested.
3. The appropriate test statistic is a one-sample *z* test.

FIGURE 10.1 ● Determining that a one-sample *z* test is the correct statistic


Are you examining differences between one sample and a population?
One-sample z test
Are you examining relationships between variables or examining the difference between groups of one or more variables?
I'm examining relationships between variables.
How many variables are you dealing with?
Two variables
t test for the significance of the correlation coefficient
More than two variables
Regression, factor analysis, or canonical analysis
I'm examining differences between groups on one or more variables.
Are the same participants being tested more than once?
Yes
How many groups are you dealing with?
Two groups
t test for dependent samples
More than two groups
Repeated measures analysis of variance
No
How many groups are you dealing with?
Two groups
t test for independent samples
More than two groups
Simple analysis of variance

COMPUTING THE *z* TEST STATISTIC

The formula used for computing the value for the one-sample *z* test is shown in Formula 10.1. Remember that we are testing whether a sample mean belongs to or is a fair estimate of a population. The difference between the sample mean ($\bar{X}$) and the population mean (μ) makes up the numerator (the value on top) for the *z* test value. The denominator (the value on the bottom that you divide by), an error term, is called the *standard error of the mean* and is the value we would expect by chance, given all the variability that surrounds the selection of all possible sample means from a population. Using this standard error of the mean (and the key term here is *standard*) allows us once again (as we showed in Chapter 9) to use the table of *z* scores to determine the probability of an outcome. It turns out that sample means drawn randomly from populations are normally distributed, so we can use a *z* table because it assumes a normal curve.

$$z = \frac{\bar{X} - \mu}{SEM}, \tag{10.1}$$

where

- $\bar{X}$ is the mean of the sample,
- μ is the population average, and
- *SEM* is the standard error of the mean.

Now, to compute the standard error of the mean, which you need in Formula 10.1, use Formula 10.2:

$$SEM = \frac{\sigma}{\sqrt{n}}, \tag{10.2}$$

where

- σ is the standard deviation for the population and
- *n* is the size of the sample.

The standard error of the mean is the standard deviation of all the possible means selected from the population. It's the best *estimate* of a population mean that we can come up with, given that it is impossible to compute *all* the possible means. If our sample selection were perfect, and the sample fairly represents the population, the difference between the sample and the population averages would be zero, right? Right. If the sampling from a population were not done correctly (randomly and representatively), however, then the standard deviation of all the means of all these samples could be huge, right? Right. So we try to select the perfect sample, but no matter how diligent we are in our efforts, there's always some error. The standard error of the mean gives a range (remember that confidence interval from Chapter 9?) of where the mean for the entire population probably lies. There can be (and are) standard errors for other measures as well.

Time for an example.

Dr. McDonald thinks that his group of earth science students is particularly special (in a good way), and he is interested in knowing whether their class average falls within the boundaries of the average score for the larger group of students who have taken earth science over the past 20 years. Because he's kept good records, he knows the means and standard deviations for his current group of 36 students and the larger population of 1,000 past enrollees. Here are the data.

	Size	Mean	Standard Deviation
Sample	36	100	5.0
Population	1,000	99	2.5

Here are the famous eight steps and the computation of the *z* test statistic.

1. State the null and research hypotheses.
   The null hypothesis states that the sample average is equal to the population average. If the null is not rejected, it suggests that the sample is representative of the population. If the null is rejected in favor of the research hypothesis, it means that the sample average is probably different from the population average.
   The null hypothesis is

$$H_0 : \bar{X} = \mu \tag{10.3}$$

The research hypothesis in this example is

$$H_1 : \bar{X} \neq \mu \tag{10.4}$$

2. Set the level of risk (or the level of significance or Type I error) associated with the null hypothesis.
   The level of risk or Type I error or level of significance (any other names?) here is .05, but this is totally at the discretion of the researcher.

3. Select the appropriate test statistic.
   Using the flowchart shown in Figure 10.1, we determine that the appropriate test is a one-sample *z* test.

4. Compute the test statistic value (called the obtained value).
   Now's your chance to plug in values and do some computation. The formula for the *z* value was shown in Formula 10.1. The specific values

are plugged in (first for *SEM* in Formula 10.5 and then for *z* in Formula 10.6). With the values plugged in, we get the following results:

$$SEM = \frac{2.5}{\sqrt{36}} = 0.42. \tag{10.5}$$

$$z = \frac{100 - 99}{0.42} = 2.38. \tag{10.6}$$

The *z* value for a comparison of the sample mean to this population mean, given Dr. McDonald's data, is 2.38.

5. Determine the value needed for rejection of the null hypothesis using the appropriate table of critical values for the particular statistic.

   Here's where we go to Table B.1 in Appendix B, which lists the probabilities associated with specific *z* values, which are the critical values for the rejection of the null hypothesis. This is exactly the same thing we did with several examples in Chapter 9.

   We can use the values in Table B.1 to see if two means "belong" to one another by comparing what we would expect by chance (the tabled or critical value) with what we observe (the obtained value).

   From our work in Chapter 9, we know that a *z* value of +1.96 has associated with it a probability of .025, and if we consider that the sample mean could be bigger, or smaller, than the population mean, we need to consider both ends of the distribution (and a range of ±1.96) and a total Type I error rate of .05.

6. Compare the obtained value and the critical value.

   The obtained *z* value is 2.38. So, for a test of this null hypothesis at the .05 level with 36 participants, the critical value is ±1.96. This value represents the value at which chance is the most attractive explanation of why the sample mean and the population mean differ. A result beyond that critical value in either direction (remember that the research hypothesis is nondirectional and this is a two-tailed test) means that we need to provide an explanation as to why the sample and the population means differ.

7. and 8. Decision time!

   If the obtained value is more extreme than the critical value (remember Figure 9.2), the null hypothesis should not be accepted. If the obtained value does not exceed the critical value, the null hypothesis is the most attractive explanation. In this case, the obtained value (2.38) does exceed the critical value (1.96), and it is absolutely extreme enough for us to say that the sample of 36 students in Dr. McDonald's class is different from the previous 1,000 students who have also taken the course. If the obtained value were less than 1.96, it would mean that there is no difference between the test performance of the sample and that of the 1,000 students who have taken the test over the past 20 years. In this case, the 36 students would have performed basically at the same level as the previous 1,000.

   And the final step? Why, of course. We wonder why this group of students differs? Perhaps McDonald is right in that they are smarter, but they may also be better users of technology or more motivated. Perhaps they just studied harder. All these are questions to be tested some other time.

So How Do I Interpret $z = 2.38$, $p < .05$?

- z represents the test statistic that was used.
- 2.38 is the obtained value, calculated using the formulas we showed you earlier in the chapter.
- $p < .05$ (the really important part of this little phrase) indicates that (if the null hypothesis is true) the probability is less than 5% that on any one test of that hypothesis, the sample and the population averages will differ by that much or more.

USING SPSS TO PERFORM A *z* TEST

We're going to take a bit of a new direction here in that SPSS does not offer a one-sample z test but does offer a one-sample t test. The results are almost the same when the sample size is a decent size (greater than 30), and looking at the one-sample t test will illustrate how SPSS can be useful—our purpose here. The main difference between this and the z test is that SPSS uses a distribution of t scores to evaluate the result.

> The real difference between a z and a t test is that for a t test, the population's standard deviation is not known, while for a z test, it is known. Another difference is that the tests use different distributions of critical values to evaluate the outcomes (which makes sense given that they're using different test statistics).

In the following example, we are going to use the SPSS one-sample t test to evaluate whether one score (13) on a test is characteristic of the entire sample. Here's the entire sample:

12

9

7

10

11

15

16

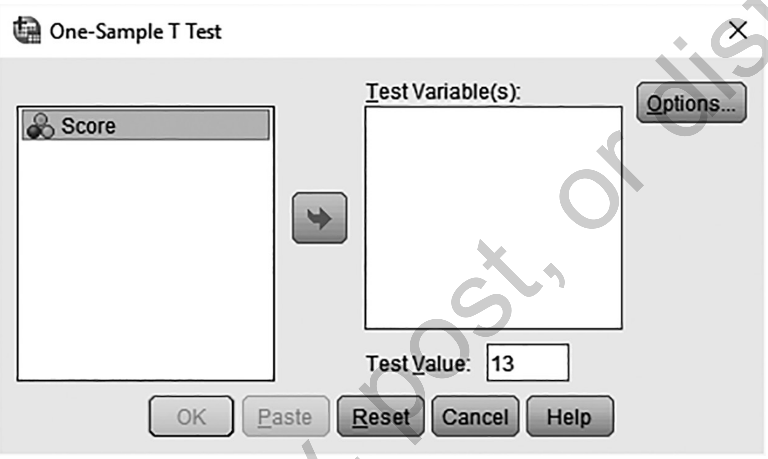
8

9

12

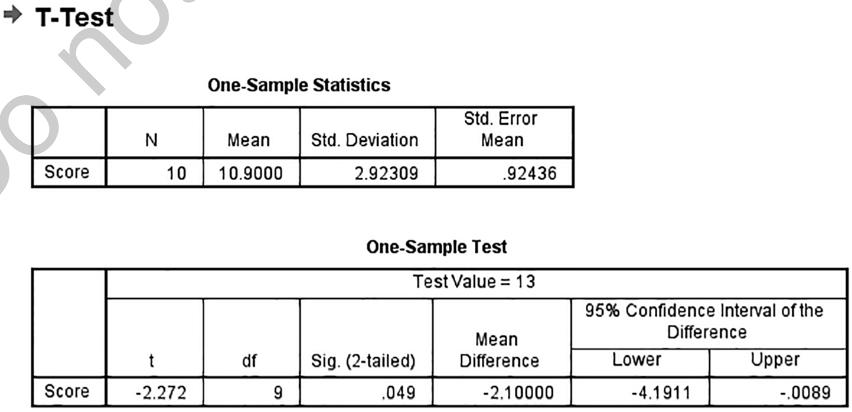
1. After the data are entered, click Analyze → Compare Means → One-Sample T test and you will see the One Sample T test dialog box as shown in Figure 10.2.
2. Double-click on the Score variable to move it to the Test Variable(s): box.
3. Enter a Test Value of 13.
4. Click OK and you will see the output in Figure 10.3.

FIGURE 10.2 ● The One-Sample T test dialog box



Test Value: 13

FIGURE 10.3 ● The output from a one-sample *t* test



T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Score	10	10.9000	2.92309	.92436

One-Sample Test

	Test Value = 13					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Score	-2.272	9	.049	-2.10000	-4.1911	-.0089

Understanding the SPSS Output

Figure 10.3 shows you the following:

1. For a sample of size 10, the mean score is 10.9 and the standard deviation is 2.92.
2. The resultant *t* value of –2.72 is significant at the .05 level (barely but it made it!).
3. The results indicate that that a test value of 13 is significantly different from the values in the sample.

SPECIAL EFFECTS: ARE THOSE DIFFERENCES FOR REAL?

Okay, time for us to revisit that big idea of effect size and learn how to use it to make your analysis of any inferential test all that much more interesting and valuable.

In general, using various inferential tools, you may find differences between samples and populations, two samples or more, and so on, but the $64,000 question is not only whether that difference is (statistically) significant but also whether it is *meaningful.* That is, does enough of a separation exist between the distributions that represent each sample or group you test that the difference is really a difference worth discussing?

Hmm. . . . Welcome to the world of effect size.

Effect size is the strength of a relationship between variables. It can be a correlation coefficient, as we talked about in Chapter 5, but the relationship between variables can also be apparent in the size of a difference between groups. It could be an indication of how effective a pill or intervention is, right?— a measure of the magnitude of the treatment. So the difference between a group that got a treatment and the group that did not shows the relationship between the independent variable (the treatment) and the dependent variable.

So effect sizes can be correlational values or values that estimate difference. What's especially interesting about computing effect size is that sample size is not taken into account. Calculating effect size, and making a judgment about it, adds a whole new dimension to understanding outcomes that does not require significance. Another interesting note about effect size is that many different inferential tests use different formulas to compute the effect size (as you saw in Chapter 5 and will see through the next few chapters), but one common metric (called Cohen's *d,* and we'll get to that shortly) tends to be used when group differences are examined.

As an example, let's take the data from Dr. McDonald and the earth science test. Here are the means and standard deviations again.

	Size	Mean	Standard Deviation
Sample	36	100	5.0
Population	1,000	99	2.5

Here's the formula for computing Cohen's *d* for the effect size for a one-sample *z* test:

$$d = \frac{\bar{X} - \mu}{\sigma}, \tag{10.7}$$

where

- $\bar{X}$ is the sample mean,
- μ is the population mean, and
- σ is the population standard deviation.

If we substitute Dr. McDonald's values in Formula 10.7, we get this:

$$d = \frac{100 - 99}{2.5} = .4$$

We know from our previous calculations that the obtained *z* score of 2.38 is significant, meaning that, indeed, the performance of Dr. McDonald's class is different from that of the population. Now we have figured out the effect size (.4), so let's turn our attention to what this statistically significant outcome might mean regarding the size of the effects.

Understanding Effect Size

The great pooh-bah of effect size was Jacob Cohen, who wrote some of the most influential and important articles on this topic. He authored a very important and influential book (your stats teacher has it on his or her shelf!) that instructs researchers in how to figure out the effect size for a variety of different questions that are asked about differences and relationships between variables. And the book also gives some guidelines as to what different sizes of effects might represent for understanding differences. You remember that for our example, the effect size is .4.

What does this mean? One of the very cool things that Cohen (and others) figured out was just what a small, medium, and large effect size is. Much of these rules-of-thumb for interpreting effect sizes are based on examining thousands of real studies and seem what is normal. They used the following guidelines:

- A small effect size ranges from 0 to .2.
- A medium effect size ranges from .2 to .8.
- A large effect size is any value above .8.

Our example, with an effect size of .4, is categorized as medium. But what does it *really* mean?

With group comparisons, effect size gives us an idea about the relative positions of one group to another. For example, if the effect size is zero, that means that both groups tend to be very similar and overlap entirely—there is no difference between the two distributions of scores. On the other hand, an effect size of 1 means that the two groups overlap about 45% (having that much in common). And, as you might expect, as the effect size gets larger, it reflects an increasing distance, or lack of overlap, between the two groups.

Jacob Cohen's book *Statistical Power Analysis for the Behavioral Sciences,* first published in 1967 with the latest edition (1988, Lawrence Erlbaum) available in reprint from Taylor and Francis, is a must for anyone who wants to go beyond the very general information that is presented here. It is full of tables and techniques for allowing you to understand how a statistically significant finding is only half the story—the other half is the magnitude of that effect. In fact, if you really want to make your head hurt, consider that many statisticians consider effect size more important than significance. Imagine that! Why might that be?

REAL-WORLD STATS

Been to the doctor lately? Had your test results explained to you? Know anything about the use of electronic medical records? In this study, Noel Brewer and his colleagues compared the usefulness of tables and bar graphs for reporting the results of medical tests.

Using a z test, the researchers found that participants required less viewing time when using bar graphs rather than tables. The researchers attributed this difference to the superior performance of bar graphs in communicating essential information (and you well remember from Chapter 4, where we stressed that a picture, such as a bar graph, is well worth a thousand words). Also, not very surprisingly, when participants viewed both formats, those with experience with bar graphs preferred bar graphs, and those with experience with tables found bar graphs equally easy to use. Next time you visit your doc and he or she shows you a table, say you want to see the results as a bar graph. Now that's stats applied to real-world, everyday occurrences!

Want to know more? Go online or to the library and find . . .

Brewer, N. T., Gilkey, M. B., Lillie, S. E., Hesse, B. W., & Sheridan, S. L. (2012). Tables or bar graphs? Presenting test results in electronic medical records. *Medical Decision Making, 32,* 545–553.

Summary

The one-sample *z* test is a simple example of an inferential test, and that's why we started off this long section of the book with an explanation of what this test does and how it is applied. The (very) good news is that most (if not all) of the steps we take as we move on to more complex analytic tools are exactly the same as those you saw here. In fact, in the next chapter, we move on to a very common inferential test that is an extension of the *z* test we covered here, the simple *t* test between the means of two different groups.

Time to Practice

1. When is it appropriate to use the one-sample *z* test?
2. What does the *z* in *z* test represent? What similarity does it have to a simple *z* or standard score?
3. For the following situations, write out in words a research hypothesis:
   a. Bob wants to know whether the weight loss for his group on the chocolate-only diet is representative of weight loss in a large population of middle-aged men.
   b. The health department is charged with finding out whether the rate of flu per thousand citizens during this past flu season is comparable to the average rate during the past 50 seasons.
   c. Blair is almost sure that his monthly costs for the past year are not representative of his average monthly costs over the past 20 years.
4. Flu cases per school this past flu season in the Remulak school system ($n = 500$) were 15 per week. For the entire state, the weekly average was 16, and the standard deviation was 15.1. Are the kids in Remulak as sick as the kids throughout the state?
5. The nightshift workers in three of Super Bo's specialty stores stock about 500 products in about 3 hours. How does this rate compare with the stocking done in the other 97 stores in the chain, which average about 496 products stocked in 3 hours? Are the stockers at the specialty stores doing a "better than average" job? Here's the info that you need:

	Size	Average Number of Products Stocked	Standard Deviation
Specialty stores	3	500	12.56
All stores	100	496	22.13

6. A major research study investigated how representative a treatment group's decrease in symptoms was when a certain drug was administered as compared with the response of the entire population. It turns out that the test of the research hypothesis resulted in a *z* score of 1.67. What conclusion might the researchers put forth? Hint: Notice that the Type I error rate or significance level is *not* stated (as perhaps it should be). What do you make of all this?

(Continued)

(Continued)

7. Millman's golfing group is terrific for a group of amateurs. Are they ready to turn pro? Here's the data. (Hint: Remember that the lower the score [in golf], the better!)

	Size	Average Score	Standard Deviation
Millman's Group	9	82	2.6
The Pros	500	71	3.1

8. Here's a list of units of toys sold by T&K over a 12-month period during 2015. Were the sales of 31,456 for 1 month in 2016 significantly different from monthly sales in 2015?

Units Sold in 2015	
January	34,518
February	29,540
March	34,889
April	26,764
May	31,429
June	29,962
July	31,084
August	30,506
September	28,546
October	29,560
November	29,304
December	25,852

Student Study Site

SAGE edge™

Get the tools you need to sharpen your study skills! Visit **edge.sagepub.com/salkindfrey7e** to access practice quizzes, eFlashcards, original and curated videos, data sets, and more!