11 ESTIMATING THE DIFFERENCE BETWEEN THE MEANS OF INDEPENDENT POPULATIONS

INTRODUCTION

In Part I, we established the importance of statistical distributions. In Part II, these distributions were shown to underlie the construction of confidence intervals around point estimates of parameters. Estimating a parameter of a single distribution has important uses. For example, in the early stages of research, one may wish to estimate mean digit span, mean number of depressive symptoms, or mean response times in specific populations. However, as research progresses, our interest inevitably shifts to how such parameters differ between populations.

In Chapters 9 and 10 we saw examples of estimating $\mu_1 - \mu_0$ using $m - \mu_0$. In some situations, it is plausible to assume we know μ_0. For example, assuming $\mu_{IQ} = 100$ may be plausible for a given population. Also, when a population is completely known, such as the distribution of reading scores obtained over 30 years, it seems quite reasonable to use the mean of this population as μ_0. In other cases, however, it is a wild stretch to assume we know μ_0. For example, it seems highly implausible to assume, as we did earlier, "that college students in general complete 24 words in 20 minutes ($\sigma = 4.8$) on a standardized crossword puzzle."

When the goal is to estimate the difference between two population means (μ_1 and μ_2), it is almost always best to obtain samples from each of the two populations and then use the difference between the two sample means, $m_1 - m_2$, to estimate $\mu_1 - \mu_2$. When we estimate $\mu_1 - \mu_2$, we say that the two populations (e.g., male versus female participants, or control versus experimental groups) represent two levels of the **independent variable** and the scores measured represent the **dependent variable**. In Chapters 11 through 15, the dependent variables we consider will be scale variables.

An **independent variable** is a qualitative or quantitative variable whose values define the two (or more) groups of interest.

A **dependent variable** is the variable for which we obtain scores.

THE TWO-INDEPENDENT-GROUPS DESIGN

Variables that have just two values are called *dichotomous variables*. In this chapter, our independent variable will be dichotomous. The left column of Table 11.1 shows examples of five different qualitative dichotomous variables and their values. The right column shows examples of five scale variables and their values.

Table 11.1 provides 25 different questions that can be asked. Five such questions can be formed by combining the dichotomous variable *sex* with the five scale variables. For example, how much do males and females differ on the following variables: *depression*, *fitness*, *intelligence*, *math ability*, and *belief in extrasensory perception* (ESP)? The same five questions can be asked for each of the four remaining dichotomous variables. For any scale variable (e.g., *fitness*), it is almost certain that the means of the two populations differ, TABLE 11.1 ■ Five Independent Variables and Five Dependent Variables

Dichotomous Independent Variable	Scale Dependent Variable
Smoking: smokers versus nonsmokers	*Depression:* scores on a clinical test
Sex: males versus females	*Fitness:* scores on a fitness test
Political views: conservatives versus liberals	*Intelligence:* IQ scores
Religious views: believers versus atheists	*Math ability:* scores on a math test
Experimental condition: treatment versus control	*Belief in ESP:* scores on an ESP questionnaire

as we saw in Chapter 9. Therefore, our goal is to estimate the size of such differences and place confidence intervals around them.

Experimental Versus Quasi-Experimental Studies

There are two ways that dichotomous groups can be formed. Some groups form naturally (e.g., smokers versus nonsmokers, age younger than 40 versus older than 40, Christian versus Muslim, depressed versus not depressed). In other cases, groups are formed by an experimenter (e.g., treatment versus control conditions). Studies involving experimenter-created groups are called **experimental** studies, and those involving naturally occurring groups are called **quasi-experimental** studies.

The critical difference between experimental and quasi-experimental studies is random assignment. One can't randomly assign individuals to levels of the first four dichotomous variables in Table 11.1 because they were formed naturally. However, one can randomly assign individuals to different conditions in an experiment. Random assignment permits inferences about causality that are much more straightforward than without it. For example, the mean score on a measure of cardiovascular health may be lower for smokers than for nonsmokers. However, we can't say that smoking caused this difference, because some other factor may have disposed people to both smoking and poor cardiovascular health. On the other hand, imagine selecting 1000 rats, dividing them randomly into two groups, and then raising one group in a smoky environment and the other in a smoke-free environment. Any measured differences in cardiovascular health following these two rearing methods may be reasonably attributed to the presence or absence of smoke.

An **experiment** involves random assignment of individuals to treatment conditions. In an experiment, it is plausible to conclude that group differences on the dependent variable are caused by the different treatment conditions.

A **quasi-experiment** compares two naturally occurring groups (i.e., groups not formed as a result of random assignment). In a quasi-experiment, it is not always plausible to conclude that group differences on the dependent variable are caused by group membership.

Real Versus Hypothetical Populations

There is an interesting (and provocative) difference between the populations considered in experimental and quasi-experimental studies that we touched on in Chapter 7 (see the section on the alternative hypothesis). The populations in quasi-experimental studies are easy to imagine, such as men versus women, smokers versus nonsmokers, Americans versus Canadians, or Sprague-Dawley rats versus Wistar rats, to name just a few randomly chosen examples. We can easily imagine drawing samples from each of these populations and computing the difference between the means of the samples.

Now let's think about a new treatment for alcohol abuse. We could choose a random sample of alcohol abusers, divide them at random into two groups, and administer the new treatment to one group but not the other. The untreated sample represents a random sample of alcohol abusers. What about the treated group? You may think it odd, but we will consider this group to be a sample from the population of alcohol abusers administered the new treatment. But, you might say, this group represents the entire population because no other individuals exist that have been administered this treatment. However, if things had been different and all alcohol abusers had been given this treatment, then our sample is one of the many random samples that could have been drawn from this population. Therefore, we can use the characteristics of this sample to infer properties of the **hypothetical population** of alcohol abusers who have been given this new treatment.

A **hypothetical population** is one that does not exist, but which could exist. Individuals in an experimental group that undergo a novel treatment can be considered a sample from a hypothetical population of all individuals that undergo the same treatment.

Confounding Variables

When we compare two samples of scores, it is important that there be no **confounding variables**. A confounding variable is one that affects the scores in our samples differently but is not the independent variable of interest. For example, if rats raised in the smoke-free environment were allowed more exercise than those raised in the smoke-filled environment, we would say that the effect of smoke is *confounded* with the effect of *exercise*. In this case, *exercise* is a confounding variable. Another example would be assigning the first 20 volunteers to the treatment condition of an experiment and the next 20 to the control condition. There may be something different about those who volunteer early and those who volunteer late that confounds our ability to assess the effect of the independent variable. In this case, *time to enroll* is the confounding variable.

A **confounding variable** is an uncontrolled variable that affects the scores in our samples differently.

We make all efforts to control for (i.e., eliminate) the effects of confounding variables in experiments, typically through random sampling from a population (e.g., alcoholics) and random assignment to experimental conditions (e.g., treatment and control). However, random sampling and random assignment do not guarantee that all confounds are eliminated. Consider a random sample from a psychology department participant pool. Because there are typically more females than males in such pools, a random sample will typically contain more women than men. When these participants are randomly assigned to the two conditions of an experiment, it is possible for all the men to be in one of the two conditions. No matter what the experimental manipulation is, the result will be confounded because one group is all women and the other is a mixture of men and women. In the long run, random assignment will cancel out confounds such as these, but any given sample may not.

Independent Versus Dependent Samples

In this chapter we will consider **independent samples**. This simply means that the scores in the two samples are not related in any systematic way. For example, if I choose 100 people and divide them randomly into two groups of 50, then the scores obtained from the individuals in the two samples are *independent* of each other. This means that knowing the score (on the dependent variable) of an individual in one sample tells you nothing about the score of any individual in the other sample.

Independent samples comprise scores that are completely unrelated to each other. This means that knowing the score of an individual in one sample tells you nothing about the score of any individual in the other sample.

However, the same 100 people could be *paired* based on their similarity on variables such as height, IQ, or extroversion. If each member of each pair is then assigned to one of the two groups, we would have individuals matched for specific characteristics in the two groups. In this case, we say there is a dependency between the members of the two groups, and the scores obtained from the individuals in the samples are dependent. This means that knowing the score (on the dependent variable) of an individual in one sample does tell you something about the score of the corresponding individual in the other sample.

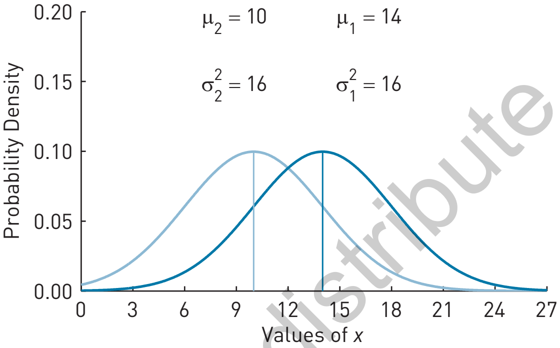
Another case of **dependent samples** is when two measurements are taken from each individual. For example, depression scores could be measured in individuals before and after treatment. These samples of before and after scores are dependent samples. When two or more scores are obtained from each individual, we say we have a *repeated-measures design*.

Dependent samples comprise scores that are related to each other in some way; either pairs of scores come from the same individual (repeated measures) or pairs of scores come from individuals matched on some characteristic (matched samples). This means that knowing the score of an individual in one sample tells you something about the score of the corresponding individual in the other sample.

Two Populations, Two Distributions

Figure 11.1 illustrates the state of affairs examined in this chapter. Each level of a dichotomous independent variable is associated with a distribution of scores on the scale dependent variable. One distribution in Figure 11.1 has a mean (μ_1) of 14 and the other has a mean (μ_2) of 10. Both distributions have a variance of 16. One of these distributions might correspond to females and the other to males. Or one distribution might correspond to a control group and the other to an experimental group. The scores on the dependent variable may correspond to cardiovascular health, belief in supernatural phenomena, anxiety, digit span, or anything else we choose to measure. Whatever the two distributions correspond to, our objective will be to estimate the difference between the two population/distribution means (i.e., $\mu_1 - \mu_2$).

FIGURE 11.1 ■ Two Populations



Two normal populations with the same variance but different means.

LEARNING CHECK 1

1. What is the primary difference between an experimental study and a quasi-experimental study?
2. Explain why it is possible to study a hypothetical population. Give an example.
3. What is the difference between an independent-groups design and a dependent-groups design?
4. For the purposes of this chapter, is it the independent or dependent variable that is dichotomous?
5. Can you imagine a two-independent-groups design in which both variables are dichotomous? If so, give an example.

Answers

1. Random assignment. We have random assignment in experiments but not in quasi-experiments.
2. In an experimental study, our treatment group may be the only group of individuals who've received the treatment in question. Nevertheless, for the purposes of estimation, we may think of this group as a sample from a population of individuals who've all received the same treatment. For example, we could estimate the mean heart rate of chimpanzees in space. If all chimps on earth had been launched into space, then we would have a population of heart rates from spacefaring chimps. If instead we choose a random sample of chimps, launch them into space, and measure their heart rates while they are in space, we would have a sample from the population of interest even though that population doesn't (currently) exist.
3. In an independent-groups design, knowing the score of an individual in one sample tells you absolutely nothing about any score in the other sample. In the dependent-groups design, there is an association between pairs of individuals in the two groups. The clearest example is the repeated-measures design, in which two scores are obtained from the same individuals at different times.
4. The independent variable is dichotomous.
5. Sure. The independent variable could be smoking status (smoker, nonsmoker) and the dependent variable could be cancer status (has cancer, doesn't have cancer). (This is not a situation that is covered in this chapter.)

AN EXAMPLE

Susan Cain (2012), a former Wall Street lawyer, wrote a well-received book titled *Quiet: The Power of Introverts in a World That Can't Stop Talking*, in which she argued that

iStock.com/Rawpixel

Western society, and the United States in particular, values the traits of extroverts more than those of introverts. She argues that introverts thrive in quiet settings where they are free to pursue their thoughts and nurture their creativity. One of Cain's strongest claims is that the group work favored in schools and industrial settings often works against the natural inclinations of introverts. In other words, introverts are not able to flourish when forced to work in groups at school or in open offices in the workplace.

Let's say that an industrial psychologist at an American university shares the view that open offices are detrimental to activities requiring insight and creativity, but she believes this is true for both introverts and extroverts. She would like to assess this idea using what she believes is a novel methodology. To assess problem solving and creativity, she will have participants solve riddles of the following sort:

> "Yesterday I went to the zoo and saw the giraffes and ostriches. Altogether they had 30 eyes and 44 legs. How many animals were there?"*
>
> "Marsha and Marjorie were born on the same day of the same month of the same year to the same mother and the same father, yet they are not twins. How is that possible?"**

Solving riddles requires insight and creative thinking; therefore, the researcher chose the number of riddles correctly solved in a 30-minute period as her dependent variable.

Our researcher's hypothesis is that more riddles will be solved in quiet settings than in noisy settings. In her experiment, all participants will be tested in a university classroom where students are learning statistics. One group will be tested when the class is writing an exam (quiet condition) and the second group will be tested while the students are engaged in group work (noisy condition).

Because our hypothetical researcher believes that the nature of the task itself requires quiet, she expects both introverts and extroverts to benefit from quiet. Therefore, when she selects participants, the researcher does not assess their positions on the introversion-extroversion continuum.

Twenty-two individuals were randomly selected from the department's participant pool and then divided at random into two groups of 11 participants each. One group was assigned to the quiet condition and the other group was assigned to the noisy condition. Each participant took the riddle test, and the number of correct solutions was counted. The results of the experiment are shown in Table 11.2 and summarized in Figure 11.2. The mean number of riddles solved in the quiet condition was 12 with a standard deviation of 4.94, and the mean number of riddles solved in the noisy condition was 9 with a standard deviation of 3.95.

The goal of the riddle-solving study was to estimate the difference between two population means (i.e., $\mu_1 - \mu_2$). The difference between the two sample means (i.e., $m_1 - m_2$) is the best point estimate of $\mu_1 - \mu_2$. The confidence interval around the difference between two sample means is computed as follows:

$$(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1 - m_2}). \tag{11.1}$$

*Answer: Each animal has two eyes, so there were 30/2 = 15 animals.

**Answer: They are triplets.

This confidence interval, like those we saw in Chapters 6, 9, and 10, involves a point estimate (i.e., a statistic) and a margin of error. The statistic in this case is the difference between the two sample means ($m_1 - m_2$), and the margin of error is $t_{\alpha/2}(s_{m_1-m_2})$.

We were reminded in Chapter 10 that the smallest sample size for which the sample variance can be computed is two. In the two-independent-groups design, we require two groups with at least two scores in each, because computing a confidence interval will require computing the variance for both samples. This means that $t_{\alpha/2}$ is based on the sum of the degrees of freedom *within* the two groups; i.e., $df_{\text{within}} = df_1 + df_2$ degrees of freedom. We can also express this as $df_{\text{within}} = n_1 + n_2 - 2$. (There is an even more interesting reason why we associate $t_{\alpha/2}$ with $df_1 + df_2$ degrees of freedom, but we will have to wait until Chapter 16 for this explanation.)

Knowing the point estimate and $t_{\alpha/2}$ leaves only the quantity $s_{m_1-m_2}$ to be explained. This quantity is the *estimated standard error* of the statistic $m_1 - m_2$. In the special case in which (i) sample sizes are the same and (ii) we can assume that the two populations have the same variance (i.e., $\sigma_1^2 = \sigma_2^2$), we compute the estimated standard error as follows:

$$s_{m_1-m_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \qquad (11.2)$$

where s_1^2 and s_2^2 are the two sample variances, and n_1 and n_2 are the two sample sizes.

With these definitions (and the assumptions that $n_1 = n_2$ and $\sigma_1^2 = \sigma_2^2$), we can compute a 95% confidence interval around the difference between the two sample means in Table 11.2. To make things clearer, we will use the subscripts q (quiet) and n (noisy) on sample means, variances, and sample sizes.

TABLE 11.2 ■ Raw Data

	Quiet	Noisy
	12	5
	4	7
	13	11
	11	3
	17	14
	21	10
	11	9
	5	8
	15	7
	14	8
	9	17
n	11	11
m	12	9
ss	244	156
s^2	24.4	15.6
s	4.94	3.95

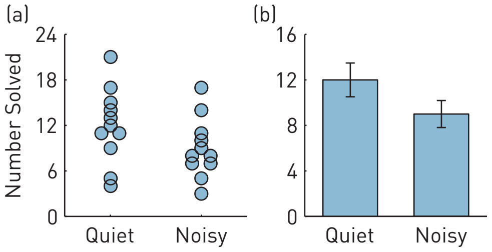
Step 1. Calculate $m_n - m_q$. $m_q = 12$ and $m_n = 9$, so $m_n - m_q = 12 - 9 = 3$.

Step 2. Calculate $s_{m_1-m_2}$ using equation 11.2:

$$s_{m_1-m_2} = \sqrt{\frac{s_q^2}{n_q} + \frac{s_n^2}{n_n}} = \sqrt{\frac{24.4}{11} + \frac{15.6}{11}} = 1.91.$$

Step 3. Determine $t_{\alpha/2}$. There are 11 + 11 − 2 = 20 degrees of freedom. Because this is a 95% confidence interval, $\alpha/2 = .025$. When we consult the *t*-table, we find that $t_{\alpha/2} = 2.086$. [The same thing can be accomplished using **T.INV.2T** in Excel by typing '=**T.INV.2T(.05, 20)**' into a cell in a spreadsheet.]

FIGURE 11.2 ■ Raw Data and Statistics



(a) Raw scores for participants in the quiet and noisy conditions. (b) The means of the two groups. Error bars represent ±*SEM*.

Step 4. Calculate the 95% confidence interval around the difference between the two means as follows:

$$\text{CI} = \left(m_q - m_n\right) \pm t_{\alpha/2}\left(s_{m_1-m_2}\right) = (12-9) \pm 2.086(1.91) = [-0.98, 6.98].$$

The results of this study might be reported as follows:

APA Reporting

The mean number of riddles solved in the quiet condition was M_q = 12 with a standard deviation of s = 4.94, and the mean number of riddles solved in the noisy condition was M_n = 9 with a standard deviation of s = 3.94. The difference between the two means was 3, 95% CI [−0.98, 6.98]. Although the results are somewhat imprecise, they support the idea that quiet conditions are more conducive to problem solving than noisy conditions.

The general form of this conclusion is familiar. The best point estimate of the difference in the number of riddles solved in the quiet and noisy populations is 12 – 9 = 3. In this example, we have 95% confidence that the true difference between the population means is in the interval [−0.98, 6.98]. Our confidence comes from knowing that 95% of intervals computed this way will capture $\mu_q - \mu_n$.

LEARNING CHECK 2

1. Bargh, Chen, and Burrows (1996) thought that exposure to words associated with age and fragility would have a subconscious effect on people and cause them to walk more slowly than individuals exposed to neutral words. In their experiment, Bargh et al. brought participants to a research lab and asked them to solve word problems. One group of 15 participants solved problems involving words such as *old*, *bingo*, and *Florida*. The other group of 15 participants solved problems involving words such as *thirsty*, *clean*, and *private*. Without their knowledge, the researchers measured participants' walking speeds when they left the research lab. Bargh et al. found that those exposed to words suggesting age and fragility walked at 2.63 mph, on average, with a variance of about 0.1933. This was slower than the speed of participants exposed to neutral words, who walked at 2.99 mph, on average, with a variance of about 0.2233. Compute the 95% confidence interval around the difference between these two means.

2. Do the results of this study support the idea that exposure to words associated with age and fragility has a subconscious effect on people and causes them to walk more slowly than individuals exposed to neutral words?

Answers

1. $m_1 = 2.63$, $s_1^2 = 0.1933$, $n_1 = 15$, $m_2 = 2.99$, $s_2^2 = 0.2233$, $n_2 = 15$. We first calculate $m_1 - m_2 = 2.63 - 2.99 = -0.36$. We then note that there are 15 + 15 – 2 = 28 degrees of freedom. Then, because this is a 95% confidence interval, we find that $t_{\alpha/2} = 2.048$ when we consult the t-table. The calculations for $s_{m_1-m_2}$ and $\text{CI} = (m_1 - m_2) \pm t_{\alpha/2}(s_{m_1-m_2})$ are shown in the table below.

2. Yes, the results support the proposal. Those exposed to words associated with age and fragility walked $m_1 - m_2 = 2.63 - 2.99 = -0.36$ mph slower than those exposed to neutral words, 95% CI [−0.70, – 0.02]. This result has become quite controversial in the last few years, and it's worth Googling.

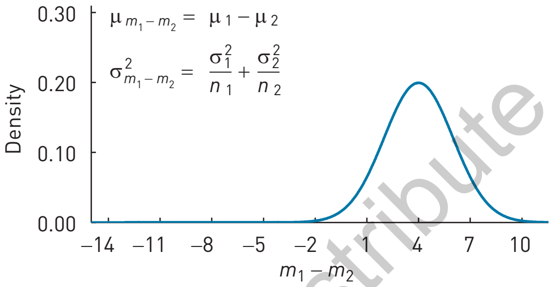
Calculate $s_{m_1-m_2}$	Calculate $\text{CI} = \left(m_1 - m_2\right) \pm t_{\alpha/2}\left(s_{m_1-m_2}\right)$
$s_{m_1-m_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $= \sqrt{\frac{0.1933}{15} + \frac{0.2233}{15}}$ $= 0.17$	$\text{CI} = \left(m_1 - m_2\right) \pm t_{\alpha/2}\left(s_{m_1-m_2}\right)$ $= -0.36 \pm 2.048(0.17)$ $= -0.36 \pm 0.34$ $= [-0.70, -.02]$

THEORETICAL FOUNDATIONS FOR THE (1–α)100% CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$

In the scenario discussed above, our hypothetical researcher was interested in the difference between two population means. One population comprises riddle scores obtained in quiet conditions and the other comprises riddle scores obtained in noisy conditions. Our researcher had in mind a situation like the one shown in Figure 11.1. She assumed that the two distributions have different means but she didn't know how different they are.

We will use the two distributions in Figure 11.1 when we need concrete illustrations in this section. Keep in mind that Population 1 has a mean of 14 and a variance of 16, and Population 2 has a mean of 10 and a variance of 16. Of course, these are things that we know but our hypothetical researcher did not.

FIGURE 11.3 ■ The Distribution of $m_1 - m_2$



The sampling distribution of the difference between two independent means. The two distributions in question are those shown in Figure 11.1. Population 1 has a mean of 14 and variance of 16, and Population 2 has a mean of 10 and variance of 16. If we consider all possible samples of size n_1 drawn from Population 1 and all possible samples of size n_2 drawn from population 2, then we can compute the difference between all possible pairs of these sample means (i.e., $m_1 - m_2$). The resulting distribution is shown. In this example, $n_1 = n_2 = 11$; therefore, $\sigma^2_{m_1-m_2} = 2.91$. and $\sigma_{m_1-m_2} = 1.71$.

The Sampling Distribution of $m_1 - m_2$

As in previous chapters, understanding a confidence interval requires understanding the sampling distribution of the statistic in question. In this case, the statistic is $m_1 - m_2$. To understand the distribution of this statistic, we must consider the means of *all possible* samples of size n_1 drawn from Population 1 and the means of *all possible* samples of size n_2 drawn from Population 2. Given these two (enormous) distributions of means, we can now imagine computing the difference between all possible pairings of the means from Populations 1 and 2. The resulting distribution is called the **sampling distribution of the difference between two means**, or the sampling distribution of $m_1 - m_2$ for short.

The **sampling distribution of the difference between two means** is a probability distribution of all possible differences between two sample means, m_1 and m_2, of size n_1 and n_2, respectively, drawn at random from two independent populations. In other words, it is the sampling distribution of all possible values of $m_1 - m_2$.

The parameters of the sampling distribution of $m_1 - m_2$ are directly related to the parameters of the distributions (populations) from which the samples were drawn. The mean of the distribution of $m_1 - m_2$ is

$$\mu_{m_1-m_2} = \mu_1 - \mu_2. \tag{11.3}$$

Because the sample mean is an unbiased statistic, the mean of all sample means drawn from Population 1 will be μ_1, and the mean of all sample means drawn from Population 2 will be μ_2. Therefore, on average, the difference between two sample means is equal to the difference between the two population means. This is illustrated in Figure 11.3, which shows that the mean of the distribution of $m_1 - m_2$ is $\mu_1 - \mu_2 = 14 - 10 = 4$.

The variance of the distribution of $m_1 - m_2$ is

$$\sigma^2_{m_1-m_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \tag{11.4}$$

This formula contains familiar components. If you think back to our discussion of the distribution of means in Chapter 5, you will remember that its variance is $\sigma^2_m = \sigma^2/n$. Equation 11.4 shows that the variance of the distribution of $m_1 - m_2$ is simply the variance of the sampling distribution of m_1 (i.e., σ_1^2/n_1) plus the variance of the sampling distribution of m_2 (i.e., σ_2^2/n_1).

Knowing the variance of the distribution of $m_1 - m_2$ allows us compute the standard error of $m_1 - m_2$ as follows:

$$\sigma_{m_1-m_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{11.5}$$

Let's take a concrete example and think about the distribution of $m_1 - m_2$ for samples drawn from the two populations shown in Figure 11.1. If $n_1 = n_2 = 11$, we can substitute numbers into equation 11.5 to obtain the following:

$$\sigma_{m_1-m_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{16}{11} + \frac{16}{11}} = 1.71.$$

This is the standard error of the distribution shown in Figure 11.3.

Looking back to our opening example, we computed a confidence interval around $m_1 - m_2$ using the usual definition $(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1-m_2})$, with the estimated standard error computed as in equation 11.2. A comparison of equations 11.2 and 11.5 shows that they are structurally identical, with σ_1^2 and σ_2^2 in equation 11.5 replacing s_1^2 and s_2^2 from equation 11.2. Therefore, $s_{m_1-m_2}$ is our best estimate of $\sigma_{m_1-m_2}$.

We've noted that using $s_{m_1-m_2}$ as defined in equation 11.2 is valid only when (i) the two sample sizes are the same, and (ii) it reasonable to assume that $\sigma_1^2 = \sigma_2^2$. Creating confidence intervals for $m_1 - m_2$ becomes slightly more complex when these conditions do not hold, but we don't need to worry about these additional complexities for the moment.

LEARNING CHECK 3

1. Calculate $\sigma^2_{m_1-m_2}$ for $\sigma_1 = 6$, $n_1 = 16$, $\sigma_2 = 5$, $n_2 = 10$.
2. Calculate $\sigma^2_{m_1-m_2}$ for $\sigma_1 = 20$, $n_1 = 16$, $\sigma_2 = 5$, $n_2 = 100$.
3. Calculate $\sigma^2_{m_1-m_2}$ for $\sigma_1 = 12$, $n_1 = 8$, $\sigma_2 = 9$, $n_2 = 2$.

Answers

1. $36/16 + 25/10 = 4.75$.
2. $\sqrt{400/16 + 25/100} = 5.02$.
3. $144/8 + 81/2 = 58.5$.

The Logic of a Confidence Interval for $\mu_1 - \mu_2$

We will now return to the familiar logic underlying confidence intervals. In Chapter 10 it was shown that when the sampling distribution of the mean is normal, it can be transformed to a *t*-distribution by applying the following transformation to each sample mean:

$$t = \frac{m - \mu}{s_m}. \tag{11.6}$$

Because $(1-\alpha)100\%$ of all *t*-scores fall in the interval $\pm t_{\alpha/2}$, we also know that $(1-\alpha)100\%$ of all possible $t \pm t_{\alpha/2}$ intervals will contain 0, which is the mean of the *t*-distribution. From this, we were able to show that $(1-\alpha)100\%$ of all possible intervals computed as

$$m \pm t_{\alpha/2}(s_m) \tag{11.7}$$

will capture the population mean, μ. There was a formal demonstration of this in Appendix 10.4 (available at study.sagepub.com/gurnsey).

We can apply the same logic to define the (1–α)100% confidence interval for $\mu_1 - \mu_2$. As with the distribution of means, we can transform the distribution of $m_1 - m_2$ into a t-distribution as follows:

$$t = \frac{(m_1 - m_2) - (\mu_1 - \mu_2)}{s_{m_1 - m_2}}. \tag{11.8}$$

If you compare equation 11.8 with equation 11.6, you will see that they are structurally identical. In equation 11.8, $m_1 - m_2$ replaces m, $\mu_1 - \mu_2$ replaces μ, and $s_{m_1 - m_2}$ replaces s_m. The result is a t-distribution with $df_{\text{within}} = df_1 + df_2$ degrees of freedom.

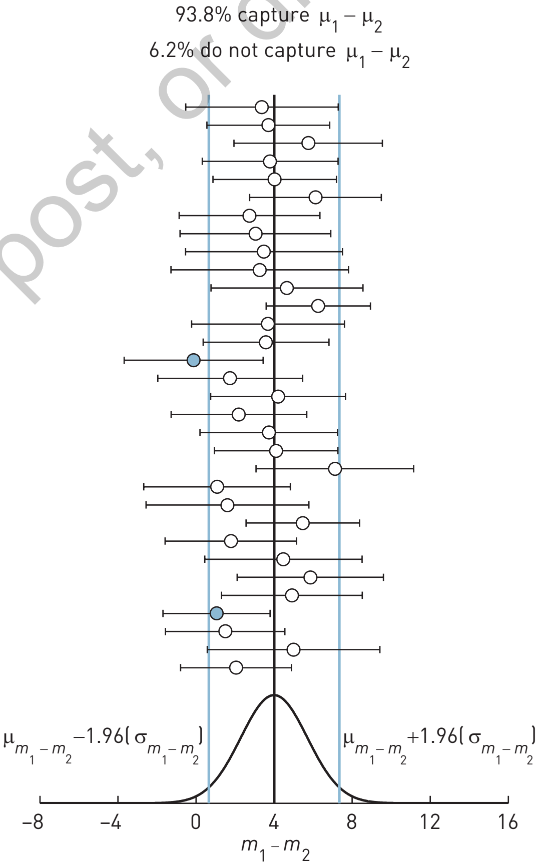
As before, because (1–α)100% of all t-scores fall in the interval $\pm t_{\alpha/2}$, we also know that (1–α)100% of all possible intervals $\pm t_{\alpha/2}$ will contain 0. From this, it follows that (1–α)100% of all possible intervals computed as

$$(m_1 - m_2) \pm t_{\alpha/2}\left(s_{m_1 - m_2}\right) \tag{11.9}$$

will capture $\mu_1 - \mu_2$. A formal demonstration of this is provided in Appendix 11.3 (available at study.sagepub.com/gurnsey).

Figure 11.4 shows that the logic underlying confidence intervals for $\mu_1 - \mu_2$ is exactly the same as the logic underlying confidence intervals for μ in Chapter 10. The sampling distribution of $m_1 - m_2$ at the bottom of Figure 11.4 was previously shown in Figure 11.3. It is a normal distribution with a mean of $\mu_1 - \mu_2 = 4$ and a standard error of $\sigma_{m_1 - m_2} = 1.71$. Each dot above the distribution represents the difference between two sample means ($m_1 - m_2$) and the arms around each dot define the 95% confidence interval. White dots are at the centers of intervals that capture $\mu_1 - \mu_2$, and filled dots are at the centers of intervals that do not capture $\mu_1 - \mu_2$. If we were to compute $(m_1 - m_2) \pm t_{\alpha/2}\left(s_{m_1 - m_2}\right)$ for all possible pairs of sample means, then exactly 95% of these intervals would capture $\mu_1 - \mu_2$.

FIGURE 11.4 ■ 95% Confidence Intervals



95% confidence intervals for $\mu_1 - \mu_2$ when the population standard deviations are unknown. The distribution at the bottom represents the sampling distribution of $m_1 - m_2$ for $\mu_1 = 14$, $\mu_2 = 10$, $\sigma_1 = \sigma_2 = 4$, and $n_1 = n_2 = 11$. The standard error of the distribution of $m_1 - m_2$ is $\sigma_{m_1 - m_2} = 1.71$. The light blue lines enclose the central 95% of the sampling distribution. Each dot represents the difference between two sample means ($m_1 - m_2$), and the arms around each dot represent the 95% confidence interval. White dots are the centers of confidence intervals that capture $\mu_1 - \mu_2$ and filled dots are the centers of confidence intervals that do not capture $\mu_1 - \mu_2$.

Estimating the Standard Error of the Difference Between Two Means

We saw earlier that computing a confidence interval around the difference between two means requires estimating the standard error of the difference between two means. The formula we use to do this depends on our sample sizes and assumptions we make about the two population variances. In this section, we will describe two ways to compute the estimated standard error of $m_1 - m_2$ when we can assume that the two populations have the same variance. When we aren't able to make this assumption, computing confidence intervals becomes a bit more complicated, and this method is described in Appendix 11.4 (available at study.sagepub.com/gurnsey).

Case 1: Equal Sample Sizes and Equal Population Variances

In the example that we worked through earlier, $\sigma_{m_1-m_2}$ was estimated using equation 11.2:

$$s_{m_1-m_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

As noted twice before, this way of estimating $\sigma_{m_1-m_2}$ is only appropriate when it is reasonable to assume that $\sigma_1^2 = \sigma_2^2$ and when sample sizes are the same ($n_1 = n_2$).

Case 2: Unequal Sample Sizes and Equal Population Variances

If we assume that $\sigma_1^2 = \sigma_2^2$ but *sample sizes are different*, then estimating $\sigma_{m_1-m_2}$ becomes more interesting and involves an intermediate step to estimate the variance common to the two distributions, which we denote σ^2. That is, subscripts on the population variances are unnecessary because they are assumed to be identical. The two sample variances, s_1^2 and s_2^2, are therefore independent estimates of σ^2.

In Chapter 5, we saw that parameter estimates based on large samples are more precise than estimates based on small samples. Therefore, when two samples are different sizes, the variance of the larger sample provides a more precise estimate of σ^2 than the variance of the smaller sample. However, the variance of the smaller sample can't be ignored. Therefore, to estimate σ^2, we combine s_1^2 and s_2^2 in a way that gives greater weight to the variance from the larger sample. This estimate of σ^2 is called s_{pooled}^2, because we *pool* two sample variances.

When s_{pooled}^2 has been computed, we use it to estimate $\sigma_{m_1-m_2}$ as follows:

$$s_{m_1-m_2} = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}}. \tag{11.10}$$

That is, s_{pooled}^2 replaces both σ_1^2 and σ_2^2 from equation 11.5. This leaves us with the relatively minor problem of computing s_{pooled}^2. There are two ways to do this. The first is more conceptual in nature and the second produces exactly the same result but is easier to compute by hand when you've been given the sums of squares for the two samples.

A **weighted sum** is a way of computing the mean of two or more statistics by multiplying each statistic by a weight related to sample size and then summing the products.

Estimating the Population Variances by "Pooling" the Sample Variances. Sample variances are pooled by making use of a so-called **weighted sum**. (We used weighted sums to compute GPA in Appendixes 1.1 through 1.3.) This means that we multiply s_1^2 by a weight and s_2^2 by a different weight and then add these products as follows:

$$s_{\text{pooled}}^2 = w_1\left(s_1^2\right) + w_2\left(s_2^2\right).$$

In the case of s_{pooled}^2, w_1 and w_2 are defined as

$$w_1 = \frac{df_1}{df_{\text{within}}}, \text{ and } w_2 = \frac{df_2}{df_{\text{within}}},$$

s_{pooled}^2 is the **pooled variance**. It is computed as a weighted sum of two separate estimates of σ^2.

where $df_{\text{within}} = df_1 + df_2$ as before. Therefore, these two weights always sum to 1; i.e., $w_1 + w_2 = 1$. The **pooled variance (s_{pooled}^2)** can be computed as a weighted sum as follows:

$$s_{\text{pooled}}^2 = \frac{df_1}{df_{\text{within}}}\left(s_1^2\right) + \frac{df_2}{df_{\text{within}}}\left(s_2^2\right). \tag{11.11a}$$

To illustrate the value of weighted sums, we will consider an extreme example. Imagine that we've drawn samples from Populations 1 and 2 from Figure 11.1. (Remember, for both distributions $\sigma^2 = 16$). Let's say $s_1^2 = 15$ and $s_2^2 = 25$. However, these two samples have very different sizes. There are $n_1 = 99$ scores in the sample drawn from Population 1 and $n_2 = 3$ scores in the sample drawn from Population 2. This means that $df_1 = 98$ and $df_2 = 2$. In this case,

$$w_1 = \frac{df_1}{df_{\text{within}}} = \frac{98}{100} = .98$$

and

$$w_2 = \frac{df_2}{df_{\text{within}}} = \frac{2}{100} = .02.$$

When we compute the weighted sum of the two sample variances, we have

$$s_{\text{pooled}}^2 = \frac{df_1}{df_{\text{within}}}\left(s_1^2\right) + \frac{df_2}{df_{\text{within}}}\left(s_2^2\right) = .98(15) + .02(25) = 15.2.$$

If we had computed the simple mean of s_1^2 and s_2^2, it would be $(15 + 25)/2 = 20$. Another way to express this average would be

$$.5\left(s_1^2\right) + .5\left(s_2^2\right) = .5(15) + .5(25) = 20.$$

So, when we give equal weight to s_1^2 and s_2^2, we obtain a poor estimate of σ^2, which in this example is 16. Therefore, when sample sizes are unequal, we compute the pooled variance as a weighted sum that gives greater weight to the variance from the larger sample.

Estimating the Population Variances by "Pooling" Sums of Squares. There is a second way to compute s_{pooled}^2 that is more useful for hand calculations. Remember that the sample variance (s^2) is the sum of squares divided by degrees of freedom ($s^2 = ss/df$). Therefore, the sum of squares is the sample variance multiplied by degrees of freedom ($ss = s^2{*}df$). With this in mind, we can show a simpler version of equation 11.11a:

$$s_{\text{pooled}}^2 = \frac{ss_1 + ss_2}{df_{\text{within}}}. \qquad (11.11\text{b})$$

The following sequence shows the equivalence of 11.11a and 11.11b:

$$s_{\text{pooled}}^2 = \frac{df_1}{df_{\text{within}}}\left(s_1^2\right) + \frac{df_2}{df_{\text{within}}}\left(s_2^2\right) = \frac{df_1\left(s_1^2\right)}{df_{\text{within}}} + \frac{df_2\left(s_2^2\right)}{df_{\text{within}}} = \frac{ss_1}{df_{\text{within}}} + \frac{ss_2}{df_{\text{within}}} = \frac{ss_1 + ss_2}{df_{\text{within}}}.$$

Therefore, if we've been given ss_1 and ss_2, we can compute s_{pooled}^2 in fewer steps using equation 11.11b.

Computing the Estimated Standard Error Using the Pooled Variance. Let's return to our opening example that estimated the effect of noise on riddle solving. Steps 1, 3, and 4 in the calculation of the confidence interval are exactly as before (so they are not recalculated). However, the calculation of $s_{m_1-m_2}$ is broken into two parts (steps 2a and 2b).

Step 2a. Compute s^2_{pooled} using equation 11.11b.

$$s^2_{\text{pooled}} = \frac{ss_q + ss_n}{df_{\text{within}}} = \frac{244+156}{10+10} = 20.$$

Step 2b. Calculate $s_{m_1-m_2}$ using equation 11.10.

$$s_{m_1-m_2} = \sqrt{\frac{s^2_{\text{pooled}}}{n_q} + \frac{s^2_{\text{pooled}}}{n_n}} = \sqrt{\frac{20}{11} + \frac{20}{11}} = 1.91.$$

If you compare $s_{m_1-m_2}$ calculated in Step 2b, you will see that it is exactly the same as $s_{m_1-m_2}$ computed in our original example. Therefore, the confidence interval computed in Step 4 will be the same in both cases.

Which Formula to Use? We've just seen that when sample sizes are the same, then equations 11.5 and 11.10 will produce exactly the same result, so *either one* can be used. However, when sample sizes are different, these two methods will produce different results, and only equation 11.10 will provide a valid estimate of $\sigma_{m_1-m_2}$. Therefore, when sample sizes are unequal, you *must* use s^2_{pooled} in the calculation of $s_{m_1-m_2}$.

Case 3: Unequal Population Variances

It is not always plausible to assume that $\sigma_1 = \sigma_2$. Sometimes interventions can change the variance in a distribution. For example, a weight loss treatment might reduce weight by the same percentage for each individual in a population. The distribution of weights would be smaller after the weight loss treatment than before the treatment. (See the discussion in Chapter 3 about the effect of multiplying the variance by a constant.)

If we cannot assume that $\sigma_1 = \sigma_2$, then confidence intervals are computed differently. First, the estimated standard error is computed using equation 11.2. However, the result will underestimate $\sigma_{m_1-m_2}$. To compensate for this underestimation, our second step is to reduce the degrees freedom to widen the confidence interval. That is, reducing the degrees of freedom increases $t_{\alpha/2}$. The amount by which the degrees of freedom are reduced depends on how different s_1^2 and s_2^2 are. Appendix 11.4 (available at study.sagepub.com/gurnsey) provides a detailed explanation of how to adjust the degrees of freedom.

Assumptions

For a confidence interval to be valid, a number of assumptions must be made.

1. The two populations we sampled are normal.
2. The two populations have the same variance.
3. The samples are random samples from the populations of interest.
4. The two samples are *independent*.

Normality

The assumption of normal populations is a common one in inferential statistics. We make this assumption because our margin of error was computed using $t_{\alpha/2}$, and the *t*-distribution assumes that the scores in our samples were drawn from normal distributions. Obvious violations of the normality assumption can be detected using QQ or PP plots or other methods discussed in Appendix 4.3 (available at study.sagepub.com/gurnsey).

Equal Variances

When we compute confidence intervals for the difference between two sample means, we often assume that the two samples were drawn from distributions having the same variance. We made this assumption in the riddle example above when we computed s^2_{pooled}. If the two populations in question were to have different variances, our confidence interval will generally be too narrow. The Welch-Satterthwaite correction factor (described in Appendix 11.4 available at study.sagepub.com/gurnsey) provides a way around this problem. Most statistical packages that compute confidence intervals around the difference between two means will provide both versions. (We will see this in Appendix 11.2.)

Random Sampling

Of course, we always have to assume random sampling from the populations of interest. The scenario at the beginning of the chapter stated that a random sample of 22 participants was drawn from the participant pool. These 22 individuals were then assigned to the two conditions (quiet and noisy) at random. Because of random selection and random assignment to conditions, there is no reason to suspect a systematic difference between the two groups; hence, there is no reason to suspect sampling bias.

The populations of interest were populations of scores (number of riddles solved) of individuals who performed the task in quiet and noisy conditions. We treat our samples as random samples from hypothetical populations that would exist if all members of the participant pool had solved riddles in a quiet environment, or if all members of the participant pool had solved riddles in a noisy environment.

Independence

Finally, we assume that there is no *association* between the scores in the two groups. In other words, the two samples are independent. In our riddle-solving example, the sample of 22 participants was divided into two equal groups at random. That is, pairs of participants were not matched on any task-relevant variables before being placed in different groups. For this reason, the two groups are independent.

EFFECT SIZE δ

In Chapters 9 and 10 we showed how to estimate Cohen's δ, which is the difference between two population means divided by their shared standard deviation. It is typically the magnitude or absolute value of δ that interests us (i.e., $|\delta|$) and not so much its sign. Nevertheless, we can compute δ in either of the following ways:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \qquad (11.12a)$$

or

$$\delta = \frac{\mu_2 - \mu_1}{\sigma}. \qquad (11.12b)$$

Therefore, we can estimate δ in either of the following ways:

$$d = \frac{m_1 - m_2}{s_{\text{pooled}}} \qquad (11.13a)$$

LEARNING CHECK 4

1. If $m_1 = 10$, $m_2 = 9$, $s_{m_2-m_1} = 1.6$, and $t_{\alpha/2} = 2$, calculate $(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1-m_2})$.
2. Calculate $\sigma^2_{m_1-m_2}$ for $\sigma_1 = 6$, $n_1 = 16$, $\sigma_2 = 5$, $n_2 = 10$.
3. Calculate s^2_{pooled} for $s_1 = 6$, $n_1 = 16$, $s_2 = 5$, $n_2 = 16$.
4. Calculate s^2_{pooled} for $s_1 = 6$, $n_l = 16$, $s_2 = 5$, $n_2 = 11$.
5. If $\alpha = .05$, $n_1 = 16$, and $n_2 = 10$, what is $t_{\alpha/2}$?
6. The 95% confidence interval is computed as $(m_1 - m_2) \pm 2.086(s_{m_1-m_2})$ What is the total number of scores in the two samples?
7. If $n_1 = 11$ and $n_2 = 13$, what proportion of all intervals computed as $(m_1 - m_2) \pm 1.717(s_{m_1-m_2})$ will capture $\mu_1 - \mu_2$?
8. Calculate the 95% confidence interval for $m_1 = 100$, $m_2 = 90$, $s_1 = 6$, $s_2 = 7$, $n_1 = 11$, $n_2 = 16$.
9. What four assumptions must be correct for $(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1-m_2})$ to be a valid confidence interval?

Answers

1. $(10 - 9) \pm 2(1.6) = [-2.2, 4.2]$.
2. $36/16 + 25/10 = 4.75$.
3. $(36 + 25)/2 = 30.5$.
4. $15/25*36 + 10/25*25 = 31.6$.
5. 2.064 for $df_{\text{within}} = 16 + 10 - 2 = 24$.
6. When we look at the *t*-table, we find that 2.086 is associated with $df_{\text{within}} = 20$ when $\alpha = .05$; therefore, $n_1 + n_2 = df_{\text{within}} + 2 = 22$.
7. $df_{\text{within}} = 22$; therefore, $t_{\alpha/2} = 1.717$ corresponds to $\alpha = .1$, so this is the 90% confidence interval. Therefore, the proportion of all such intervals that will capture $\mu_1 - \mu_2$ is .9.
8. $(100 - 90) \pm 2.060(2.592) = 10 \pm 5.340 = [4.66, 15.34]$.
9. The two populations we sampled are normal, the two populations have the same variance, the samples are random samples from the populations of interest, and the two samples are independent.

or

$$d = \frac{m_2 - m_1}{s_{\text{pooled}}}. \tag{11.13b}$$

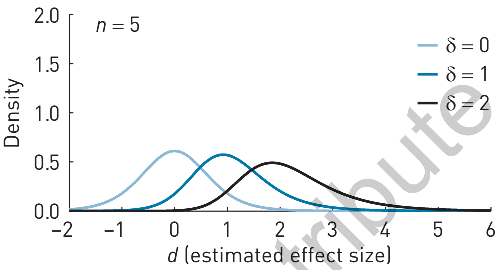
In equations 11.13a and 11.13b, s_{pooled} is simply the square root of s^2_{pooled}, as defined in equations 11.11a and 11.11b. That is, s_{pooled} estimates σ, which is the standard deviation common to the two distributions.

An Approximate Confidence Interval Around an Estimate of δ

As with any statistic, a point estimate is of limited use if we don't know the precision of the estimate. Therefore, we next show how to construct an approximate (1−α)100% confidence interval around *d*. Of course, computing a confidence interval requires knowing the sampling distribution of the statistic. Unfortunately, the sampling distribution of *d* is not usually normal, as illustrated in Figures 11.5 and 11.6. Figure 11.5 illustrates sampling distributions of *d*, computed as in equation 11.13a, for samples of size $n = 5$, drawn from two populations whose means are separated by $\delta = 0$ to 2. When $\delta = 0$, the sampling distribution of *d* resembles a *t*-distribution. As δ increases, the distributions shift to the right, because the average value of *d* increases as δ increases. More important is the fact that the distributions become increasingly skewed as δ increases.

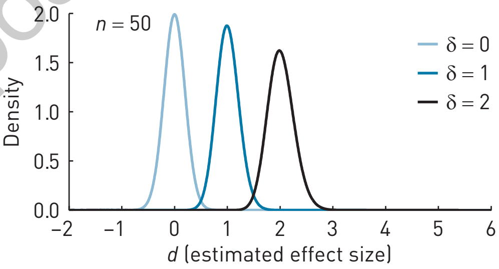
The skew in the distributions of *d* is more apparent when it is computed from small samples than when it is computed from large samples. Figure 11.6 illustrates sampling distributions of *d* for δ = 0 to 2 when both sample sizes are 50. As in Figure 11.5, the distributions in Figure 11.6 shift to the right as δ increases. In addition, the distributions in Figure 11.6 are narrower than those in Figure 11.5 because the samples are larger, and thus the estimates of δ are less variable. Finally, the right skew of the distributions (particularly δ = 2) is much less pronounced in Figure 11.6 than in Figure 11.5.

FIGURE 11.5 ■ Distributions of *d*



Sampling distributions for *d* when δ = 0, 1, and 2 when both sample sizes are 5. As δ increases, the distributions shift to the right and become wider and increasingly skewed.

FIGURE 11.6 ■ Distributions of *d*



Sampling distributions for *d* when δ = 0, 1, and 2 when both sample sizes are 50. As δ increases, the distributions shift to the right and become wider, but they are markedly less skewed than in Figure 11.5.

We can compute an approximate confidence interval around *d* using a simple extension of the method used in Chapter 10 as follows:

$$d \pm z_{\alpha/2}(s_d). \qquad (11.14)$$

The only part of equation 11.14 that we haven't yet explained is s_d, which is the approximate standard error of *d*. In Chapter 9, when we had one sample and σ was known, the standard error of *d* was

$$\sigma_d = \sqrt{\frac{1}{n}}.$$

In Chapter 10, when we had one sample and σ was unknown, the approximate estimated standard error of *d* was

$$s_d = \sqrt{\frac{d^2}{2df} + \frac{1}{n}}.$$

In the present case, we have two samples and σ is unknown. The approximate estimated standard error of *d* is

$$s_d = \sqrt{\frac{d^2}{2df_{\text{within}}} + \frac{1}{n_1} + \frac{1}{n_2}}. \qquad (11.15)$$

In equation 11.15, *d* is our estimate of δ, computed from our samples as described in equation 11.13a. As before, $df_{\text{within}} = df_1 + df_2$. Because of the nature of this approximation, we use $z_{\alpha/2}$ rather than $t_{\alpha/2}$. Confidence intervals computed using this approximation will always be slightly different from the exact confidence interval. For small values of *d*, the confidence intervals will be slightly narrower; for large values of *d*, the confidence interval will be slightly wider. Fortunately, these differences are often negligible, particularly as sample sizes increase.

Our Example (Continued)

We will now step through the calculation of an approximate confidence interval for an estimated effect size. For quick reference, Table 11.3 shows the data from the riddle study for which we computed the confidence interval around $m_1 - m_2$. We will use the same data to compute *d* and the 95% confidence interval around it.

TABLE 11.3 ■ Data From Riddle Study

	Quiet (q)	Noisy (n)
N	11	11
M	12	9
ss	244	156

Step 1. Compute s_{pooled}. Table 11.3 shows that $ss_{\text{q}} = 244$ and $ss_{\text{n}} = 156$. Because $n_1 = n_2 = 11$, $df_{\text{within}} = 11 + 11 - 2 = 20$. With this information, we can compute s_{pooled} as follows:

$$s_{\text{pooled}} = \sqrt{\frac{ss_{\text{q}} + ss_{\text{n}}}{df_{\text{within}}}} = \sqrt{\frac{244+156}{10+10}} = 4.47.$$

Step 2. Compute d, using equation 11.13.

$$d = \frac{m_{\text{q}} - m_{\text{n}}}{s_{\text{pooled}}} = \frac{12-9}{4.47} = 0.67.$$

Step 3. Compute s_d, using equation 11.15.

$$s_d = \sqrt{\frac{d^2}{2 * df_{\text{within}}} + \frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{0.67^2}{2*20} + \frac{1}{11} + \frac{1}{11}} = 0.4394.$$

Step 4. Compute the approximate 95% confidence interval around d using equation 11.14.

$$\text{CI} = d \pm z_{\alpha/2}(s_d) = 0.67 \pm 1.96(0.4394) = [-0.19, 1.53].$$

Therefore, the approximate 95% CI is [−0.19, 1.53]. Our confidence in this interval comes from knowing that approximately 95% of all confidence intervals computed in this way will contain δ.

According to Cohen's classification scheme, described in Chapters 9 and 10, our estimated effect size of 0.67 is between medium and large. Of course, this classification scheme should always be treated with some skepticism. In a later section we will reconsider a more quantitative approach to giving meaning to our estimate of δ.

An Exact Confidence Interval Around an Estimate of δ

In Appendix 10.3 we described how to compute an exact confidence interval using MBESS `ci.sm` in **R** for an estimated effect size (d) based on a sample mean and a known population mean. We can also compute an exact confidence interval for d when computed as in equations 11.13a and 11.13b using `ci.smd` from MBESS. The text in the code fragment below shows the arguments provided to `ci.smd`: `smd` is the standardized mean difference (i.e., d), `n.1` and `n.2` are the sample sizes, and `conf.level` is the confidence level. With values assigned to each of these arguments, we press return and `ci.smd` returns the lower and upper limits of the interval, [−`0.1986694`, `1.522964`] as well as the center of the interval. When it is rounded to two decimal places, the 95% confidence interval [−`0.20`, `1.53`] is very similar to the approximate interval computed in the previous section.

```
> ci.smd (smd =.67, n.1 = 11, n.2 = 11, conf.level =.95)
$Lower.Conf.Limit.smd
[1] −0.1986694
$smd
[1] 0.67
$Upper.Conf.Limit.smd
[1] 1.522964
```

U_3

In Chapters 8 through 10, we discussed Cohen's U_3 as a measure of overlap between two populations of scores. This allowed us to think in concrete terms about the effect of some treatment or intervention at the level of populations. We can use U_3 in the context of the two-independent-groups design just as we did for the one-sample design in earlier chapters.

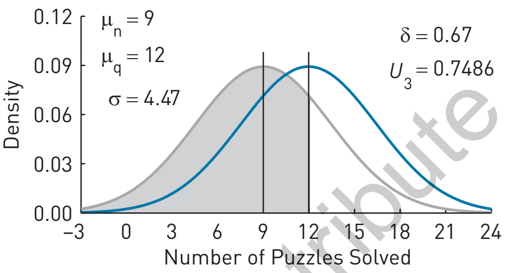
Figure 11.7 illustrates our best estimate of the relationship between the two populations. Both distributions are assumed to be normal, and our sample statistics are the best estimates of the population parameters. Our best estimate of μ_q is $m_q = 12$, our best estimate of μ_n is $m_n = 9$, and our best estimate of σ is $s_{\text{pooled}} = 4.47$. Therefore, our best estimate of δ is $d = 0.67$. U_3 is the proportion of the noise distribution below the mean of the quiet distribution (shown in gray in Figure 11.7) or, equivalently, the proportion of the quiet distribution above the mean of the noise distribution. When estimating δ from two independent samples, U_3 is calculated exactly as before:

$$U_3 = P(|d|),$$

where $P(|d|)$ is the proportion of the standard normal (z) distribution below the absolute value of d. In our example, our best estimate is that $U_3 = .7486$.

Let's think about the population of scores for the quiet condition. The distribution is assumed to be normal, with 50% of scores lying above and below its mean, μ_q. Our results suggest that if all individuals from this population had been tested under noisy conditions, the mean would have dropped by 0.67(σ). This means that 74.86% of scores would now fall below μ_q, as shown in the shaded gray region in Figure 11.7. Therefore, we estimate that testing in the noisy versus quiet conditions leads to an increase of $74.86 - 50 = 24.86\%$ of scores falling below μ_q. Conversely, testing in quiet conditions would result in 24.86% of scores falling above μ_n. This seems like a substantial effect on performance. But, like any statistical result, the practical significance of U_3 depends on one's perspective, as we'll see in a later section.

FIGURE 11.7 ■ Cohen's U_3



An illustration of our best estimate of the relationship between quiet and noisy populations. Because d = .67, we estimate that 74.86% of the noisy distribution lies below the mean of the quiet distribution.

LEARNING CHECK 5

1. What does s_{pooled} estimate?
2. What does $d = (m_1 - m_2)/s_{\text{pooled}}$ estimate?
3. For fixed sample sizes, how does the shape of the sampling distribution of d change as δ gets further from 0?
4. Compute the approximate 95% confidence interval for d for $m_1 = 100$, $m_2 = 90$, $s_1 = 8$, $s_2 = 8$, $n_1 = 20$, $n_2 = 20$.

Answers

1. σ, the standard deviation assumed to be common to the two populations in question.
2. $\delta = (\mu_1 - \mu_2)/\sigma$.
3. The sampling distribution of d becomes increasingly skewed as δ gets further from 0.
4. $d = (100 - 90)/8 = 1.25$.

$$s_d = \sqrt{d^2/(2df_{\text{within}}) + 1/n_1 + 1/n_2}$$
$$= \sqrt{.0206 + .1} = .3472.$$

$$\text{CI} = d \pm z_{\alpha/2}(s_d) = 1.25 \pm 1.96(.3472) = [0.57, 1.93].$$

SIGNIFICANCE TESTING

Although our focus has been on estimating $\mu_1 - \mu_2$ and δ, we've seen in previous chapters that many researchers follow the custom of significance testing to judge whether a treatment is effective. As always, the null hypothesis is that there is no difference between the means of the two populations under consideration. We saw in Chapter 10 that significance tests can be conducted with confidence intervals and *t*-statistics, so we will review these in turn.

Significance Testing With Confidence Intervals

In previous chapters, two-tailed tests of the null hypothesis with $\alpha = .05$ were conducted by asking whether μ_0 falls in the 95% confidence interval around *m*. In the present case, we are estimating the mean of the distribution of $m_1 - m_2$. According to the null hypothesis, our two populations have the same means ($\mu_1 = \mu_2$), so we can state it as

$$H_0\text{: } \mu_1 - \mu_2 = 0.$$

For a two-tailed test of H_0, our alternative hypothesis is simply

$$H_1\text{: } \mu_1 - \mu_2 \neq 0.$$

To conduct a two-tailed test of our null hypothesis, we ask if 0 falls in our confidence interval defined as $(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1-m_2})$. If it does, we retain H_0; if not, we reject H_0. In the riddle-solving example that we worked through above, the 95% confidence interval around $m_1 - m_2$ was [−0.98, 6.98]. Because this interval contains 0, we retain the null hypothesis. (Remember that retaining the null hypothesis does not mean that it is true.)

Hypothesis Testing With *t*-Statistics

Traditionally, significance tests are conducted with *t*-statistics rather than confidence intervals. However, the logic of the hypothesis test is the same. If the researcher predicts that quiet conditions will lead to more correct solutions than noisy conditions, then the null and alternative hypotheses would be as follows:

$$H_0\text{: } \mu_q - \mu_n = 0$$

$$H_1\text{: } \mu_q - \mu_n > 0$$

If the researcher makes no prediction about which condition will lead to more correct solutions, then the null and alternative hypotheses will be as follows:

$$H_0\text{: } \mu_q - \mu_n = 0$$

$$H_1\text{: } \mu_q - \mu_n \neq 0$$

In either case, t_{obs} is defined as follows:

$$t_{\text{obs}} = \frac{m_1 - m_2}{s_{m_1-m_2}}. \qquad (11.16)$$

When we fill in the quantities defined earlier, we find that

$$t_{\text{obs}} = \frac{m_q - m_n}{s_{m_1-m_2}} = \frac{12-9}{1.91} = 1.57.$$

To make a decision about t_{obs}, we would have to compare it with t_{critical}. As we saw in Chapter 10, t_{critical} depends on α, *df*, and H_1. Let's make the conventional assumption that $\alpha = .05$. The degrees of freedom are $df_{\text{within}} = n_q + n_n - 2 = 20$, exactly as in the 95% confidence interval we computed. Finally, assuming the directional alternative hypothesis, H_1 predicts that $\mu_q > \mu_n$, so we expect a positive value of t_{obs}. With this information we can consult the *t*-table for the *t*-score having α(100)% = 5% of the distribution above it when $df_{\text{within}} = 20$. In doing so, we find $t_{\text{critical}} = 1.725$. Because $t_{\text{obs}} = 1.57$ fails to exceed t_{critical}, we retain H_0.

Using the **T.DIST** function in Excel, we can determine that the exact proportion of a *t*-distribution (with 20 *df*) above 1.57 is $p = .0657$. Again, because $p = .0657$ is greater than $\alpha = .05$, we retain H_0. (For a two-tailed test, $p = 2{*}.0657 = .1314$.)

APA Reporting

The mean number of riddles solved for 11 participants in the quiet condition was $M_q = 12$, and the mean number of riddles solved for the 11 participants in the noisy condition was $M_n = 9$. The difference between the two means was 3, 95% CI [−0.98, 6.98]. A one-tailed test of the difference between the two means showed that the difference was not statistically significant, $t(20) = 1.57$, $p = .07$.

The Connection Between *d* and t_{obs}

The vast majority of research in psychology over the past 75 years or so has relied on significance tests to draw inferences about the effects of experimental manipulations. Until recently, very little emphasis has been placed on measures of effect size. As a consequence, you may read an older paper in which the authors report t_{obs} but don't mention the corresponding effect size. If you are in this situation, you can take heart in the following, very simple connection between *d* and t_{obs}:

$$d = t_{\text{obs}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{11.17}$$

Therefore, if an author provides t_{obs} and the sample sizes, one can easily recover *d* and then put a confidence interval around it.

The APA reporting section above provides the information required by equation 11.17 to recover *d*. We do this as follows:

$$d = t_{\text{obs}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.57\sqrt{\frac{1}{11} + \frac{1}{11}} = 0.67.$$

One of the unfortunate things about the past emphasis on significance tests is that t_{obs} may not be reported when it is not statistically significant. This means that we can't recover information about the effect size associated with the difference in question. Of course, this produces a very unbalanced representation of effect sizes in the literature.

INTERPRETATION OF OUR RIDDLE STUDY

The researcher in the riddle-solving experiment predicted that more riddles would be solved in a quiet setting than in a noisy setting. This prediction is supported by the data. Participants in the quiet condition solved three more riddles, on average, than those in

LEARNING CHECK 6

1. A researcher at the University of Dallas wonders what effect challenging cognitive activities has on outcomes on the Montreal Cognitive Assessment (MoCA) test of cognitive functioning. Thirty-six elderly men (ages 65 to 75) were selected at random and divided, at random, into a treatment group and a control group; there were 18 individuals in each group. The treatment group participated in a digital photography course for 3 hours a day for 4 weeks. The control group spent 3 hours a day watching television in the common room of a local senior home. At the end of the 4-week period, all 36 men were administered the MoCA. The results of the experiment were as follows: $m_T = 29.6$, $m_C = 27.2$, $s_T = 2.2$, and $s_C = 3.1$.

   (a) State the null and alternative hypotheses in symbols.

   (b) Assuming $\alpha = .05$, compute a confidence interval to test the null hypothesis.

   (c) If you were to conduct the hypothesis test using a *t*-test, what would be the value of t_{critical}?

   (d) Calculate t_{obs}.

   (e) Based on your calculations, should you retain or reject the null hypothesis?

   (f) Show how to compute an estimate of δ from t_{obs}.

Answers

1. (a) H_0: $\mu_{m_T - m_C} = 0$ and H_1: $\mu_{m_T - m_C} \neq 0$; no prediction was made.

   (b) $t_{\alpha/2} = 2.032$. $(m_1 - m_2) \pm t_\alpha(s_{m_1 - m_2}) = (29.6 - 27.2$ $2.032(.896) = [0.58, 4.22]$.

   (c) $t_{\text{critical}} = \pm 2.032$.

   (d) $t_{\text{obs}} = (29.6 - 27.2)/.896 = 2.679$.

   (e) $t_{\text{obs}} = 2.679$, $p = .011$. Reject H_0 (a *p*-value computed in Excel, $p < .02$, is also acceptable).

   (f) $d = t_{\text{obs}}\sqrt{1/n_1 + 1/n_2} = 2.679\sqrt{2/18} = .89$.

the noisy condition, 95% CI [−0.29, 6.29]. This represents a 33% improvement (12/9 = 1.33) and corresponds to an estimated effect size of 0.67, 95% CI [−0.19, 1.53]. From the estimated effect size of 0.67 we estimate that in the two populations, 74.86% of the noisy distribution lies below the mean of the quiet distribution.

Although the researcher's prediction is supported, the estimate of the difference between the two population means is somewhat imprecise. In fact, the 95% confidence interval contains 0. So, although 3 is our best estimate of the difference between the two population means, 0 remains one of many plausible values for $\mu_q - \mu_n$. Some would say that the difference is not statistically significant, but this should not be taken to mean that there is no evidence of a difference between the two population means.

This hypothetical experiment provided an analog of the real-world situation of working in a quiet or a noisy setting. The results suggest that for some types of work, a quiet setting may produce better outcomes. One could imagine many real-world contexts in which work quality would suffer in a noisy and active open office. Computer programming and architectural design might be examples of jobs that are done more effectively in quiet environments. Such jobs, like riddle solving, seem to require extended periods of focus, during which many choices are considered and the implications of each choice must be weighed. Interrupting these thought processes might lead to poor work.

On the other hand, people are different. It may be that introverts gravitate to quieter environments and perform better there, whereas extroverts seek out more dynamic environments. Therefore, it might be interesting to see whether both introverts and extroverts benefit from quiet environments on the riddle task, or whether introverts show greater benefit. The methods to address such questions go beyond the two-groups design and will be considered in Part IV of this book.

One mustn't forget, however, that there are jobs for which a dynamic open environment is essential. Think of comedy writers. So, although we have a (hypothetical) experimental example of quiet conditions yielding better outcomes, different results might be obtained for different populations of individuals and different types of tasks.

A Note on Sample Size

As with most examples, we have used small samples so that we will have manageable numbers to illustrate the calculations involved. However, it is important to remember that for both estimation and significance testing, it is possible to make rational and informed choices about sample sizes.

From the estimation point of view, sample size determines the margin of error. For the example we've been working with in this chapter, the margin of error associated with the confidence interval around $m_1 - m_2$ was shown above to be $t_{\alpha/2}(s_{m_1-m_2}) = 2.086 * 1.9069 = 3.98$. Appendix 11.5 (available at study.sagepub.com/gurnsey) shows that this margin of error is approximately 0.88(σ), or almost a full standard deviation wide. In some circumstances, this lack of precision might be acceptable. For example, in the present case we were simply asking if there is any evidence that quiet rooms are more conducive to riddle solving than noisy rooms. The results support the idea that there is an advantage to working in a quiet setting. However, because the samples are small, we are unable to state how big the effect is with much precision. Appendix 11.5 continues the discussion of how to choose an appropriate sample size to achieve a desired level of precision.

From the significance testing perspective, we use prospective power analysis to choose sample sizes. We can use G∗Power to determine sample sizes required to achieve a specified power, given an assumed effect size (δ). However, G∗Power can also be used to determine the power of the experiment *post hoc* (i.e., after the experiment has been run). The experiment described in our example has very low power. The G∗Power application shows that if the population effect size were δ = 0.67 (as estimated in our hypothetical study), then using two samples of size 11 would yield power of only .45. Furthermore, we can use G∗Power to determine what effect size can be detected with a specified power, for a given α and sample size. We call this the *sensitivity* of the experiment. For example, assuming a one-tailed test with α = .05 and power = .8, the choice of 11 participants per group suggests that the researcher was interested in an effect size of δ = 1.1. That is, the experiment has the sensitivity to detect an effect size of δ = 1.1 power* 100% of the time when sample sizes are 11 and α = .05. Appendix 11.6 (available at study.sagepub.com/gurnsey) illustrates how to use G∗Power to obtain both post hoc power and sensitivity.

PARTITIONING VARIANCE

Overview

In this section we will discuss the fascinating interconnectedness of statistics. This interconnectedness is a major theme that we will see often in the remainder of this book. We've already seen that d and t_{obs} are connected in a relatively straightforward way (equation 11.17). We'll now see that things go deeper than this.

The interconnectedness of statistics derives from the notion of partitioning variance. To see where we're going, let's think about merging the 11 scores in each of the two groups of our riddle study into a single group of 22 scores. Partitioning variance means that we can decompose the variability in this merged set of scores into two components. One component is related to the variability within each of the two subgroups of 11 scores. The other component is related to the difference between the two sample means. From this decomposition we will discover a new statistic that we will call r^2, which represents the proportion of variance in the merged group of scores that is attributable to the difference between the two group means. Although r^2 is derived quite differently from Cohen's *d*, we will see that they are two expressions of the same thing. We will also see how other statistics are connected to each other through r^2 and Cohen's *d*.

There will be quite a few formulas in this section, but introducing them is not an intellectual exercise to illustrate the interconnectedness of statistics. Rather, the research literature abounds with different statistics, and when we understand the connections between them, we are in a much better position to relate studies to each other and to see the order in what might otherwise seem a chaotic set of results. Seeing the unity in statistics puts us in a better position to understand an important statistical technique called meta-analysis, which will be introduced in the last section of this chapter.

Between- and Within-Group Variability

We could continue with the data from the riddle-solving example to illustrate the notion of partitioning variance. However, an example with fewer scores will be easier to work with. Therefore, we will use a concrete example involving cats and dogs.

Assume that we have a collection of three cats and three dogs and we've measured the weight of each animal. Figure 11.8a shows these weights. Cats are denoted by circles and dogs are denoted by squares. It should be clear from Figure 11.8 that cats weigh less than dogs on average. More importantly, the variability within the group of cats, and the variability within the group of dogs, is less than the variability in the two groups combined. It is this difference that we will focus on. To simplify the discussion, we will use sums of squares as a measure of variability.

iStock.com/GlobalP

In the discussion that follows, we will refer to weights as scores and denote the scores for cats and dogs as y_{cats} and y_{dogs}, respectively. The scores for the cats, $y_{\text{cats}} = \{6, 9, 12\}$, have a mean of $m_{\text{cats}} = 9$. The scores for the dogs, $y_{\text{dogs}} = \{19, 25, 31\}$, have a mean of $m_{\text{dogs}} = 25$. To compute the sums of squares within each group, we subtract the group mean from each score and then square and sum these deviation scores. The deviation scores for cats are $\{-3, 0, 3\}$, producing $ss_{\text{cats}} = 18$. The deviation scores for dogs are $\{-6, 0, 6\}$, producing $ss_{\text{dogs}} = 72$.

Within-group variability is the variability about the mean of a sample or population.

The sums of squares for cats and dogs reflect **within-group variability**. If we add the sums of squares for the two groups, we can denote the result as ss_{within}. Therefore, the total within-group variability is $ss_{\text{within}} = 18 + 72 = 90$.

Now let's merge our two groups of three scores into one group of six scores. The resulting merged set will be $y_{\text{total}} = \{6, 9, 12, 19, 25, 31\}$. This merged set of scores has a mean and a sum of squares, which we'll call m_{total} and ss_{total}, respectively. As shown in Figure 11.8a, $m_{\text{total}} = 17$. Subtracting 17 from each score in y_{total} produces deviation scores, $y_{\text{total}} - m_{\text{total}} = \{-11, -8, -5, 2, 8, 14\}$. Squaring and summing these deviation scores produces $ss_{\text{total}} = 474$.

Between-group variability reflects the variability in a collection of scores resulting from scores having been drawn from two or more populations.

We now have two sums of squares, ss_{within} and ss_{total}. The within-group variability makes up part of ss_{total} but not all of it. The difference between ss_{total} and ss_{within} is $ss_{\text{total}} - ss_{\text{within}} = 474 - 90 = 384$. We refer to this difference as **between-group variability** and denote it with ss_{between}. We will next see that ss_{between} is derived from the mean scores of cats and dogs.

To see more clearly what ss_{between} represents, we replace each score in y_{total} with the mean of the sample from which it came. This will produce the following collection of numbers: $y_{\text{means}} = \{9, 9, 9, 25, 25, 25\}$. The mean of these six numbers is $m_{\text{means}} = 17$; i.e., the mean of the means equals m_{total}. When we subtract m_{means} from each of the means, we obtain the following deviation scores: $y_{\text{means}} - m_{\text{means}} = \{-8, -8, -8, 8, 8, 8\}$. Squaring and summing these deviation scores produces $ss_{\text{means}} = 6{*}8^2 = 384$. (Notice, we've seen 384 before.) We will refer to ss_{means} as ss_{between} because it represents variability associated with the difference between group means.

From this example, we can see the following relationship:

$$ss_{\text{total}} = ss_{\text{between}} + ss_{\text{within}}$$

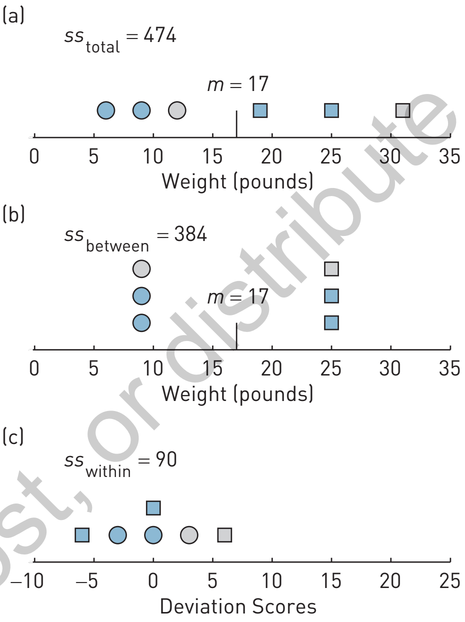
$$474 = 384 + 90.$$

That is, we have decomposed the total variability in our six scores (ss_{total}) into a part associated with variability within groups (ss_{within}) and a part associated with variability between groups (ss_{between}). This is what we mean by partitioning variance.

The numbers in Table 11.4 summarize the calculations we've just seen and relate to Figures 11.8a through 11.8c. The first column of Table 11.4 identifies the groups in question (i.e., cats and dogs). The second column (labeled "Subject") provides a number that we can use to refer to each individual. Each of the six scores shown in the third column comes from a specific individual (see Figure 11.8a). The fourth column shows the means of the corresponding groups; i.e., $m_{\text{cats}} = 9$ and $m_{\text{dogs}} = 25$ (see Figure 11.8b). The fifth column shows the deviation of each subject's score from its group mean (see Figure 11.8c). The deviations are denoted with the letter e because deviation scores are sometimes called error scores, and we don't want confusion with Cohen's d.

Table 11.4 shows that each individual's score can be broken down into two parts. One part is the group mean and the other is its deviation from the group mean. For example,

FIGURE 11.8 ■ Partitioning Variability



A graphical illustration of partitioning variability. (a) The six scores from Table 11.4 are shown; cats are circles and dogs are squares. (b) Each of these six scores is replaced with the group mean (9 or 25). (c) Each of the six scores is replaced with its deviation from its group mean.

TABLE 11.4 ■ Illustrating the Concept of r^2

Group	Subject	Scores (y)	Means (m)	$e = y - m$
Cats	1	6	9	−3
Cats	2	9	9	0
Cats	3	12	9	3
Dogs	4	19	25	−6
Dogs	5	25	25	0
Dogs	6	31	25	6
	Mean	17	17	0
	ss	474	384	90
		r^2	0.8101	

subject 4 (a dog) has a score of 19, which equals the group mean of 25, plus the score's deviation from the mean; i.e., −6. Therefore, we can say that $y_4 = m_4 + e_4 = 25 - 6 = 19$. Breaking down (or decomposing) each score into these two parts is the essential feature of partitioning the total variability into within-group and between-group variability.

In summary, the total variability in our scores (column *y* in Table 11.4) is defined as

$$ss_{\text{total}} = \Sigma(y - m_y)^2.$$

The variability attributable to the difference between group means (column *m* in Table 11.4) is defined as

$$ss_{\text{between}} = \Sigma(m - m_y)^2.$$

And the variability not attributable to the difference between group means (column *e* in Table 11.4) is defined as

$$ss_{\text{within}} = \Sigma e^2.$$

Remember, the mean of the *e* scores is 0. The row labeled *ss* in Table 11.4 shows that $ss_{\text{total}} = 474$, $ss_{\text{between}} = 384$, and $ss_{\text{within}} = 90$.

Explained Variance as an Effect Size

We can now ask, what *proportion* of ss_{total} is attributable to group means? Well, that's easy. We define the proportion of ss_{total} attributable to group means as

$$r^2 = \frac{ss_{\text{between}}}{ss_{\text{total}}}.$$

In this case,

$$r^2 = \frac{384}{474} = .8101.$$

That is, 81.01% of the variability in our set of six scores is *attributable* to (or explained by) group means. Put differently, 81.01% of the variability in our set of six scores comes from between-group variability.

Furthermore, the variability not explained by means is

$$1 - r^2 = \frac{ss_{\text{within}}}{ss_{\text{total}}},$$

which in this case is

$$1 - r^2 = \frac{90}{474} = .1899.$$

That is, 18.99% of the variability in our set of six scores is *not attributable* to the differences in the group means. Put differently, 18.99% of the variability in our set of six scores comes from within-group variability.

The calculations we just worked through are cumbersome, and we did them to establish the general notion of partitioning variance, and r^2 in particular. It turns out that there is a very simple relationship between t_{obs} and r^2. It is just this:

$$r^2 = \frac{t_{\text{obs}}^2}{t_{\text{obs}}^2 + df_{\text{within}}}. \tag{11.18}$$

Although we haven't computed t_{obs}, the data in Table 11.4 give us enough information to do so. We saw that $m_{\text{cats}} = 9$ and $m_{\text{dogs}} = 25$. The deviation scores in column 5 of Table 11.4 allow us to compute the corresponding variances. That is, $s_{\text{cats}}^2 = \Sigma\{-3,\ 0,\ 3\}^2/2 = 9$ and $s_{\text{dogs}}^2 = \Sigma\{-6,\ 0,\ 6\}^2/2 = 36$. Therefore,

$$t_{\text{obs}} = \frac{m_{\text{dogs}} - m_{\text{cats}}}{s_{m_1 - m_2}} = \frac{m_{\text{dogs}} - m_{\text{cats}}}{\sqrt{\frac{s_{\text{dogs}}^2}{n_{\text{dogs}}} + \frac{s_{\text{cats}}^2}{n_{\text{cats}}}}} = \frac{25-9}{\sqrt{\frac{9}{3} + \frac{36}{3}}} = 4.1312.$$

If we put t_{obs} into equation 11.18, we find the following

$$r^2 = \frac{t_{\text{obs}}^2}{t_{\text{obs}}^2 + df_{\text{within}}} = \frac{4.1312^2}{4.1312^2 + 4} = .8101,$$

which is exactly what was computed previously as $r^2 = ss_{\text{between}}/ss_{\text{total}}$. Therefore, if we know t_{obs} and df_{within}, we can easily determine the proportion of variability in our scores explained by group means.

We can now think back to the riddle example that we've worked with throughout this chapter. Equation 11.18 provides a very simple way to determine the proportion of variability in our 22 scores explained by the difference between group means. Remember that t_{obs} for the riddle example was 1.57 and df_{within} was 20. When we put these numbers into equation 11.18 we find the following:

$$r^2 = \frac{t_{\text{obs}}^2}{t_{\text{obs}}^2 + df_{\text{within}}} = \frac{1.57^2}{1.57^2 + 20} = .1097.$$

This means that about 11% of the total variability in our 22 scores is explained by the difference between the group means, and the remaining 89% is explained by within-group variability; i.e., the vast majority of variability in our set of 22 scores represents within-group variability.

As noted at the beginning of this section, r^2 is an effect size. However, the interpretation of r^2 (or its square root r) as an effect size presents the same complications as the interpretation of d. An r^2 that is considered large in one field of study (e.g., social psychology) may be considered small in another (e.g., neuroscience). Therefore, there are no universal guidelines that allow us to say what are small, medium, and large effects when r is the measure of effect size. Well, there are almost no universal guidelines. As with d, Jacob Cohen provided guidelines that he found useful in his field of study. These are summarized in Table 11.5.

TABLE 11.5 ■ Cohen's Guidelines

Classification	r	r^2
Small	.10	.01
Medium	.25	.0625
Large	.40	.16

Note: Cohen's guidelines for effect size (r and r^2) are to be used as a last resort!

The Connection Between r^2 and d

Equation 11.17 showed that t_{obs} and d are directly related, and equation 11.18 shows that t_{obs} and r^2 are directly related. This suggests that r^2 and d should be directly related, and indeed they are. The relationship is this:

$$d = \sqrt{df_{\text{within}}\left(\frac{r^2}{1-r^2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}. \tag{11.19}$$

In the section on effect size (Cohen's d), we determined that our estimate of δ in the riddle study was $d = .67$. Therefore, if we substitute $r^2 = .1097$ into equation 11.19, we should obtain $d = .67$, which we do:

$$d = \sqrt{df_{\text{within}}\left(\frac{r^2}{1-r^2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{20\left(\frac{.1097}{1-.1097}\right)\left(\frac{1}{11}+\frac{1}{11}\right)} = 0.67.$$

You should work through these numbers yourself to confirm that the calculations are correct.

What we've just shown is really elegant. The standardized difference between two means (d) is directly related to the proportion of total variance explained by the difference between the two means (r^2). These are two units-free measures of how different the means of two distributions are. This result is not only elegant but also very useful.

The *F*-Statistic

Before drawing this discussion to a close, there is one last statistic to be thrown into the mix. This is the *F*-statistic (in honor of Sir Ronald Fisher) and it is widely used in the advanced analyses that are covered in Part IV of this book. However, *F* can be computed for the independent-groups design that we are currently considering. For the two-groups design, the *F*-statistic is defined as

$$F = \frac{r^2}{1-r^2}df_{\text{within}}. \tag{11.20}$$

For our riddle example,

$$F = \frac{r^2}{1-r^2}df_{\text{within}} = \frac{.1097}{1-.1097}20 = 2.46.$$

Interestingly, F is related to t_{obs} as follows:

$$F = t_{\text{obs}}^2 \text{ or} \tag{11.21}$$

$$t_{\text{obs}} = \sqrt{F}. \tag{11.22}$$

In the riddle example, $t_{\text{obs}} = 1.57$ and when squared (equation 11.21), we find $F = t_{\text{obs}}^2 = 1.57^2 = 2.46$, exactly as we found above using equation 11.20.

The Interconnectedness of Statistics

We've just worked through a lot of formulas that connect statistics to each other. Interconnectedness is one of the major themes that emerge from the study of statistics. This is fascinating from a purely intellectual point of view. In fact, this kind of interconnectedness is what makes mathematics beautiful. We can enjoy this on a small scale.

However, beyond aesthetic reasons, there are important practical reasons to understand these connections. Once understood, these interconnections allow us to compare the results of statistical analyses that have used different statistics. For example, different researchers may have run essentially the same experiment but reported the results using different statistics, including d, r^2, t_{obs}, and F. When we know how to translate between these four statistics, we can more easily compare the results of the experiments. Furthermore, we can put all results on a common scale and combine the results. In the next section we will see how this works.

LEARNING CHECK 7

1. A researcher at the University of Arizona was curious about how much male and female psychology students know about the local football team. (The Arizona Cardinals were formerly the St. Louis Cardinals, who, in their glory years, featured Jim Hart and Mel Gray as the core of an incredibly productive passing offense.) The researcher obtained three female and three male volunteers from the Psychology Department participant pool and asked each to name as many current Cardinals players as possible. The results of this small study showed that $y_{males} = \{8, 9, 10\}$ and $y_{females} = \{4, 5, 6\}$.

(a) What proportion of the variability in these two sets of scores is explained by sex?

(b) What proportion of the variability in these two sets of scores is not explained by sex?

(c) Convert r^2 to d.

(d) Compute the approximate 95% confidence interval around d.

(e) What does this confidence interval say about the null hypothesis that $m_{males} = m_{females}$?

Answers

1. (a) $r^2 = .8571$.

(b) $1 - r^2 = .1429$.

(c) $d = \sqrt{4\left(\frac{.8571}{.1429}\right)\left(\frac{2}{3}\right)} = 4.$

(d) $s_d = \sqrt{\frac{16}{8} + \frac{1}{3} + \frac{1}{3}} = 1.633$. CI = 4 ± 1.96(1.633) = [.80, 7.20].

(e) If the null hypothesis were true, we would expect 0 to fall within this interval. Because it does not, we can reject the null hypothesis.

META-ANALYSIS

All the examples we've worked through to this point have dealt with the results of single studies. This reflects how research is typically done. A researcher has a question that he or she would like answered. The study is conducted, the analysis is completed, and a paper is written describing the experiment, results, and conclusions. The paper is then submitted to a professional journal in the hope that it will be judged suitable for publication.

The examples in this and preceding chapters show that the results of individual studies can be rather imprecise. That is, the confidence intervals around our point estimates can be quite wide. The appendixes for Chapters 6 and 10 as well as Appendix 11.5 (available at study.sagepub.com/gurnsey) show that with proper planning, we can have some control over our margin of error. However, in practical terms there may be limits to how precise a single study can be. In any case, it is rare for a question of any importance to be settled by the results of a single study. Rather, we rely on replication and the cumulative weight of evidence arising from many related studies to draw our conclusions.

Meta-analysis is a quantitative method of data analysis that combines the results of many individual studies to obtain a more precise estimate of a population parameter.

One method used to combine the results of several studies is **meta-analysis** (Ellis, 2010; Hedges & Olkin, 1995; Hunter & Schmidt, 1990; Smith & Glass, 1977). The logic of the method is extremely simple: if the results of several individual studies are somewhat imprecise, then the results of these individual studies can be averaged to yield a more precise result. For example, if several studies had addressed the effects of a new treatment for depression, then the results of all such studies can be averaged to yield a more precise estimate of the benefits of the treatment. The logic is exactly the same as all estimation procedures we've done to this point. In single studies, we combine *measures taken from individuals* to estimate a population parameter. In meta-analysis, we combine the *results of studies* to get a more precise estimate of a population parameter.

Throughout this chapter we have worked with the example of the effect of noise (or quiet) on riddle solving. We found that the estimated effect size was d = .67. Let's say a second study of the same sort had been conducted, with 25 participants in each condition (quiet and noisy), and a statistically significant increase in the number of riddles solved was found, $t(48) = 3.7$, $p = .0006$. Perhaps a third study used the same method with 20 participants in each group. In this study too there was an increase in the number of riddles solved, which corresponded $r^2 = .25$. That is, the difference between sample means accounted for 25% of the variability in the scores. Finally, a fourth study was conducted with two groups of 16 participants and a statistically significant increase in the number of riddles solved was found, $F = 6.5$, $p = .016$.

Now we have four studies addressing the same question, but each has presented the results in a different way. However, we now know that all three measures can be converted to an estimate of δ. That is, all three results can be expressed as d. This is shown in Table 11.6. Using equation 11.17, we find that $t(48) = 3.7$ corresponds to $d = 1.05$. Using equation 11.22, we find that $r^2 = .25$ corresponds to $d = 1.13$. And, using equation 11.17, we find that $F = 6.5$ corresponds to $d = 0.90$; in this case, $t_{\text{obs}} = \sqrt{F}$.

TABLE 11.6 ■ Converting t_{obs} and r^2 to d for a Meta-Analysis

Study	Statistic	Value	Expressed as *d*	Transformation
Study 1	d	0.67	0.67	
Study 2	t_{obs}	3.70	1.05	$d = t_{\text{obs}} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Study 3	r^2	0.25	1.13	$d = \sqrt{df_{\text{within}}\left(\frac{r^2}{1 - r^2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
Study 4	F	6.50	0.90	$d = \sqrt{F} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Mean			0.94	
Standard deviation			0.20	

Now that all statistics have been converted to the same units (d), we can compute the mean and standard deviation of these four values of d. At the bottom of Table 11.6, we can see that the mean of the four d values is 0.94 with a standard deviation of 0.20. This is a meta-analysis. We have combined the results of four studies to obtain a more precise estimate of the true population effect size. Although we won't do it here, we could go on to compute a confidence interval around this average value of d. (We could also test it for statistical significance if we thought this would be useful, or if a journal editor forced us to.)

It is important to recognize a second important role of d in meta-analysis. Remember, d, like z, is a unitless measure. Therefore, the responses on many different dependent variables can be converted to d and then averaged. For example, in the four studies we just considered, the dependent variable was the number of riddles solved. However, different dependent variables could have been used. For example, a researcher might have measured the time required to solve five riddles, rather than the number solved in 30 minutes. Response times are supposed to reflect the same underlying psychological construct measured by the number of riddles solved. Therefore, participants who solve more riddles would be expected to solve a fixed number of riddles in a shorter time. So, on average, participants in the quiet condition would be expected to solve riddles faster than those in the noisy condition. The difference between these two *time-to-solve* means can be converted to d and combined in a meta-analysis with d derived from two *number solved* means.

Meta-analysis is a powerful tool for combining research results, and estimates of δ play an important role. First, many different *statistics* can be converted to d so that they are put on a common scale and averaged. Second, many different dependent variables, all reflecting the same underlying psychological construct, can be converted to d and averaged.

When you read about individual research results that seem interesting, you should always be thinking meta-analytically (Kline, 2013), which means

- thinking about how the dependent variable in one study relates to the dependent variables used in related studies, and
- what the weight of evidence suggests about the question being addressed in these related studies.

Thinking meta-analytically keeps us from thinking that the results of a single study are definitive, as we might be led to believe from a misinterpretation of p-values.

Because meta-analysis is a very general but simple analysis, we will wait until we've seen a few more statistical methods before returning to the question of how results obtained using these different methods can be combined.

SUMMARY

In this chapter we saw how to compute a confidence interval around the difference between two independent sample means when the population standard deviations are unknown. The confidence interval is computed as follows:

$$(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1 - m_2}),$$

where $m_1 - m_2$ is the difference between two means drawn from two independent populations. Population 1 has mean μ_1 and variance σ_1^2, and population 2 has mean μ_2 and variance σ_2^2. One can (theoretically) compute all possible values of m_1 from samples of size n_1 from population 1, and all possible values of m_2 from samples of size n_2 from population 2. If $m_1 - m_2$ is computed for all possible combinations of m_1 and m_2, the result will be *the sampling distribution of the difference between two means*. This distribution will have a mean of $\mu_1 - \mu_2$ and a variance of $\sigma_1^2/n_1 + \sigma_2^2/n_2$. If sample sizes are equal and it is assumed that $\sigma_1^2 = \sigma_2^2$, then $\sigma_1^2/n_1 + \sigma_2^2/n_2$

can be estimated with $s_1^2/n_1 + s_2^2/n_2$. If sample sizes are unequal but we can assume that $\sigma_1^2 = \sigma_2^2$, then we can compute

$$s^2_{\text{pooled}} = \frac{df_1}{df_{\text{within}}}(s_1^2) + \frac{df_2}{df_{\text{within}}}(s_2^2)$$

to estimate the variance (σ^2) common to the two populations. s^2_{pooled} can then be substituted into

$$\sigma_{m_1 - m_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

as follows to yield the estimated standard error of the difference between two means, $s_{m_1 - m_2}$:

$$s_{m_1 - m_2} = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}}.$$

We use $s_{m_1 - m_2}$ to compute our confidence interval. If σ_1^2 and σ_2^2 cannot be assumed to be identical, then we can use the Welch-Satterthwaite correction (described in Appendix 11.4 available at study.sagepub.com/gurnsey) to increase $t_{\alpha/2}$ appropriately.

The information used to compute a confidence interval around $m_1 - m_2$ can also be used to estimate δ and to place an approximate confidence interval around this estimate. Specifically,

$$d = \frac{m_1 - m_2}{s_{\text{pooled}}}$$

estimates δ. An approximate $(1-\alpha)100\%$ confidence interval around d is computed as follows:

$$d \pm z_{\alpha/2}(s_d)$$

where

$$s_d = \sqrt{\frac{d^2}{2df_{\text{within}}} + \frac{1}{n_1} + \frac{1}{n_2}}.$$

The estimated effect size d is related to t_{obs} as follows:

$$d = t_{\text{obs}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

This is a very useful connection because many research studies report only t_{obs}. So, if you know t_{obs} and the size of the two samples, you can recover d and place an approximate confidence interval around it.

Confidence intervals, $(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1 - m_2})$, can be used to test the null hypothesis that H_0: $\mu_1 - \mu_2 = 0$. If 0 does not fall within the interval, we can reject a two-tailed test of the null hypothesis with a significance level of α. The traditional method of testing the null hypothesis is to compute

$$t_{\text{obs}} = \frac{m_1 - m_2}{s_{m_1 - m_2}}$$

and reject H_0 if t_{obs} exceeds the t_{critical} value that is established based on α, df_{within}, and H_1.

The concept of partitioning variance is important for two reasons. The first is that r^2 is an alternative estimate of a population effect size, which we will call ρ^2 in later chapters. r^2 is the proportion of variability in our two samples, treated as a single set of scores, that is explained by the difference between the two group means. r^2 is related to t_{obs} in a very simple way:

$$r^2 = \frac{t^2_{\text{obs}}}{t^2_{\text{obs}} + df_{\text{within}}}.$$

r^2 is also related in a simple way to d. Specifically,

$$d = \sqrt{df_{\text{within}}\left(\frac{r^2}{1 - r^2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

In statistics, everything is related. Because t_{obs} and r^2 can be converted to d, we can combine the results of studies using different statistics and different dependent variables in a meta-analysis.

KEY TERMS

between-group variability	284
confounding variable	264
dependent samples	264
dependent variable	262
experiment	263
hypothetical population	264
independent samples	264
independent variable	262
meta-analysis	290
pooled variance (s^2_{pooled})	272
quasi-experiment	263
sampling distribution of the difference between two means	269
weighted sum	272
within-group variability	284

EXERCISES

Definitions and Concepts

1. What is the difference between an experiment and a quasi-experiment?
2. Please explain the concept of a hypothetical population.
3. What is the difference between independent and dependent samples?

True or False

State whether the following statements are true or false.

4. For the two-groups design, the independent variable is dichotomous and the dependent variable is quantitative.
5. The mean heart rate of a sample of 32 Wistar rats living on the international space station is an estimate of the mean heart rate of Wistar rats living on the international space station.
6. It is impossible to estimate the mean of a hypothetical population.
7. Statistical significance implies high practical significance.
8. $s_{m_1-m_2}$ is the standard error of the difference between two means.
9. $\sigma^2_{m_1-m_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.
10. If $n_1 = n_2 = 5$, and $s_1 = s_2 = 15$, then $\sigma_{m_1-m_2} = 6$.
11. If $n_1 = n_2 = 5$, and $s_1 = s_2 = 15$, then $s_{\text{pooled}} = 15$.
12. If $n_1 = n_2 = 11$, and $t_{\text{obs}} = 4$, then $r^2 = .6154$.
13. If $n_1 = n_2 = 11$, and $t_{\text{obs}} = 4$, then $d = 1.706$.

Calculations

14. For each of the following parameters, compute the mean and variance of the sampling distribution of the difference between two means.
    (a) $\mu_1 = 10$, $\mu_2 = 12$, $\sigma_1 = 6$, $\sigma_2 = 5$, $n_1 = 15$, $n_2 = 15$
    (b) $\mu_1 = 10$, $\mu_2 = 12$, $\sigma_1 = 6$, $\sigma_2 = 5$, $n_1 = 22$, $n_2 = 15$
    (c) $\mu_1 = 20$, $\mu_2 = 6$, $\sigma_1 = 6$, $\sigma_2 = 25$, $n_1 = 15$, $n_2 = 4$
    (d) $\mu_1 = 18$, $\mu_2 = 30$, $\sigma_1 = 28$, $\sigma_2 = 5$, $n_1 = 4$, $n_2 = 15$
    (e) $\mu_1 = 4$, $\mu_2 = 8$, $\sigma_1 = 16$, $\sigma_2 = 8$, $n_1 = 32$, $n_2 = 64$
15. For each of the following scenarios, (i) calculate the 95% confidence interval around $m_1 - m_2$, (ii) compute the approximate 95% confidence interval around d, (iii) compute t_{obs}, and (iv) compute r^2.
    (a) $m_1 = 10$, $m_2 = 12$, $s_1 = 6$, $s_2 = 5$, $n_1 = 15$, $n_2 = 15$
    (b) $m_1 = 9$, $m_2 = 11$, $s_1 = 5$, $s_2 = 5$, $n_1 = 10$, $n_2 = 15$
    (c) $m_1 = 8$, $m_2 = 7$, $s_1 = 4$, $s_2 = 3$, $n_1 = 9$, $n_2 = 15$
    (d) $m_1 = 7$, $m_2 = 8$, $s_1 = 3$, $s_2 = 4$, $n_1 = 8$, $n_2 = 15$

Scenarios

16. Twenty university students (chosen at random from students enrolled at the University of Vermont) took part in an experiment testing the effect of watching FOX News versus CNN on knowledge of world events. The 20 participants were divided into two equal groups of 10 participants each. One group was assigned to watch FOX and the other was assigned to watch CNN. Participants then watched their assigned station (FOX or CNN) for 3 hours a night for 4 weeks. At the end of the 4-week period they were administered a test of world knowledge. The results were as follows: $m_{\text{FOX}} = 110$, $m_{\text{CNN}} = 99$, $s_{\text{FOX}} = 10$, $s_{\text{CNN}} = 6$.
    (a) Compute the 95% confidence interval around $m_{\text{FOX}} - m_{\text{CNN}}$.
    (b) What does it mean to have 95% confidence in this interval?
    (c) Use the 95% confidence interval to test the null hypothesis H_0: $\mu_{\text{FOX}} - \mu_{\text{CNN}} = 0$ against the alternative hypothesis H_1: $\mu_{\text{FOX}} - \mu_{\text{CNN}} \neq 0$.
    (d) What do you conclude from these data?
17. This font is called Comic Sans MS. An honors student at Concordia University theorized that reading comprehension is reduced when

text is presented in Comic Sans MS rather than the more commonly used Helvetica. She chose to test her theory in a population of university students using a widely available test of reading comprehension. The student drew a random sample of 42 university students in Canada, which she then divided, at random, into two groups of 21. One group was given the reading test formatted using Comic Sans MS, and the other was given the test formatted in Helvetica. After she collected the results from the 42 students, she found the following: $m_{CSMS} = 110$, $m_{Helvetica} = 122$, $s_{CSMS} = 16$, $s_{Helvetica} = 20$.

(a) What proportion of the variability is explained by the difference between the two means?

(b) Calculate the approximate 95% confidence interval around *d*.

(c) If H_1: $\mu_{CSMS} - \mu_{Helvetica} < 0$, can we reject the null hypothesis at the $\alpha = .05$ significance level? Explain your answer.

18. Researchers at Concordia University have a theory that drinking beer while studying will reduce test anxiety and result in improved test grades. To test their theory, the researchers chose 16 students at random from those currently enrolled in an introductory statistics course at Concordia. Half of the students were asked to consume two bottles of beer during their normal studying periods, and the other half were asked to refrain from alcohol consumption while studying. The students then wrote the test in their regular class period. The results were as follows: $m_{Beer} = 75$, $m_{NoAlc} = 70$, $s_{Beer} = 8$, $s_{NoAlc} = 6$. Conduct a test to assess the researchers' theory about the relationship between beer drinking and test performance.

(a) Assuming $\alpha = .05$, is there a statistically significant difference between the means of the two groups? Explain your answer.

(b) What proportion of the variability in the test scores is explained by between-groups variability?

(c) Would Cohen consider this a small, medium, or large effect size?

19. A professor of linguistics was dismayed by the prevalence of the word "like" in contemporary language. Thinking this was a generational difference, he measured the number of times a random selection of young and old subway passengers used the word "like" in a 5-minute conversation. (This study did not have the ethical approval of his university's ethics committee and verges on the creepy.) The results were as follows: $m_{young} = 60$, $m_{old} = 75$, $s_{young} = 15$, $s_{old} = 17$. There were 25 individuals in the group of older subway passengers and 36 in the group of younger subway passengers.

(a) Compute the 95% confidence interval around $m_{old} - m_{young}$.

(b) If his alternative hypothesis was H_1: $\mu_{old} - \mu_{young} < 0$, would he be able to reject the null hypothesis based on this confidence interval?

(c) Compute the approximate 95% confidence interval around his best estimate of δ.

APPENDIX 11.1: ESTIMATION AND SIGNIFICANCE TESTS IN EXCEL

Confidence Intervals and t_{obs} for $m_1 - m_2$

In previous appendixes we introduced all the Excel functions needed to conduct the analyses described in the body of this chapter. Figure 11.A1.1 shows how to compute a confidence interval around $m_1 - m_2$ and how to calculate t_{obs}. In cells **B2** to **B7**, we are provided with values for m_1, m_2, s_1, s_2, n_1, and n_2. We are also provided with the value for α in cell **B9**. The within-groups degrees of freedom (df_{within}) are computed in cell **B8** as $n_1 + n_2 - 2$.

From the numbers given, m_1–m_2 is calculated in cell **B11**, and

$$s^2_{pooled} = \frac{df_1}{df_{within}}(s_1^2) + \frac{df_2}{df_{within}}(s_2^2)$$

is calculated in cell **B12**. s^2_{pooled} is used in the calculation of

$$s_{m_1-m_2} = \sqrt{\frac{s^2_{pooled}}{n_1} + \frac{s^2_{pooled}}{n_2}},$$

in cell **B13**. To compute a confidence interval around $m_1 - m_2$ requires $t_{\alpha/2}$, which is computed in cell **B14** using the **T.INV** function. The lower and upper bounds of the confidence interval are computed in cells **B16** and **B17**.

A two-tailed test of the null hypothesis can be conducted by asking whether 0 falls in the 95% confidence interval around $m_1 - m_2$. One can also compute

$$t_{\text{obs}} = \frac{m_1 - m_2}{s_{m_1 - m_2}}$$

as shown in cell **B19**. The **T.DIST** function is used in cell **B20** to compute the one-tailed p-value associated with t_{obs}. The two-tailed p-value is computed in cell **C21**.

Confidence Intervals for *d*

Cells **B2** to **B12** of Figure 11.A1.2 contain exactly the same quantities as cells **B2** to **B12** in Figure 11.A1.1. In cell **B13**, s_{pooled} is computed simply as the square root of s^2_{pooled}. In cell **B14**,

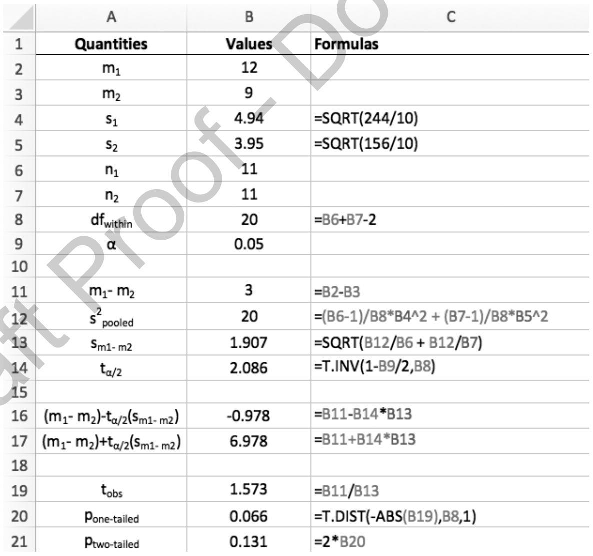
$$d = \frac{m_1 - m_2}{s_{\text{pooled}}}$$

is computed, and the estimated standard error of d is computed in cell **B15** as

$$s_d = \sqrt{\frac{d^2}{2df_{\text{within}}} + \frac{1}{n_1} + \frac{1}{n_2}}.$$

NORM.S.INV is used in cell **B16** to compute $z_{\alpha/2}$. The lower and upper limits of the confidence interval around d [i.e., $d \pm z_{\alpha/2}(s_d)$] are computed in cells **B18** and **B19**.

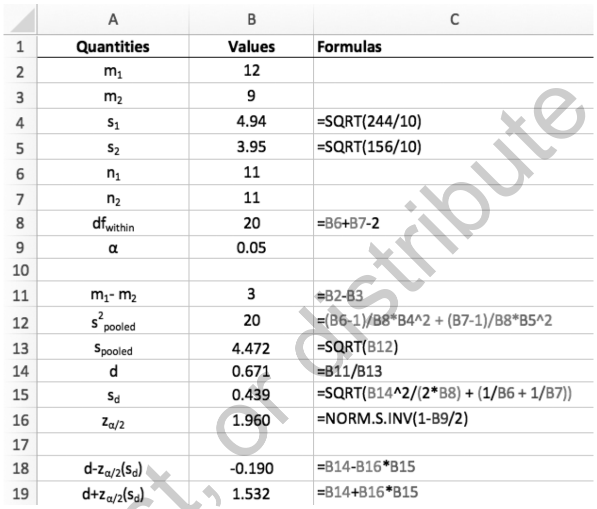
FIGURE 11.A1.1 ■ Confidence Intervals and t_{obs}



	A	B	C
1	**Quantities**	**Values**	**Formulas**
2	m_1	12	
3	m_2	9	
4	s_1	4.94	=SQRT(244/10)
5	s_2	3.95	=SQRT(156/10)
6	n_1	11	
7	n_2	11	
8	df_{within}	20	=B6+B7-2
9	α	0.05	
10			
11	m_1- m_2	3	=B2-B3
12	s^2_{pooled}	20	=(B6-1)/B8*B4^2 + (B7-1)/B8*B5^2
13	$s_{\text{m1- m2}}$	1.907	=SQRT(B12/B6 + B12/B7)
14	$t_{\alpha/2}$	2.086	=T.INV(1-B9/2,B8)
15			
16	(m_1- m_2)-$t_{\alpha/2}$($s_{\text{m1- m2}}$)	-0.978	=B11-B14*B13
17	(m_1- m_2)+$t_{\alpha/2}$($s_{\text{m1- m2}}$)	6.978	=B11+B14*B13
18			
19	t_{obs}	1.573	=B11/B13
20	$p_{\text{one-tailed}}$	0.066	=T.DIST(-ABS(B19),B8,1)
21	$p_{\text{two-tailed}}$	0.131	=2*B20

Confidence intervals and t-tests for the two-independent-groups design.

FIGURE 11.A1.2 ■ Confidence Intervals for *d*



	A	B	C
1	**Quantities**	**Values**	**Formulas**
2	m_1	12	
3	m_2	9	
4	s_1	4.94	=SQRT(244/10)
5	s_2	3.95	=SQRT(156/10)
6	n_1	11	
7	n_2	11	
8	df_{within}	20	=B6+B7-2
9	α	0.05	
10			
11	m_1- m_2	3	=B2-B3
12	s^2_{pooled}	20	=(B6-1)/B8*B4^2 + (B7-1)/B8*B5^2
13	s_{pooled}	4.472	=SQRT(B12)
14	d	0.671	=B11/B13
15	s_d	0.439	=SQRT(B14^2/(2*B8) + (1/B6 + 1/B7))
16	$z_{\alpha/2}$	1.960	=NORM.S.INV(1-B9/2)
17			
18	d-$z_{\alpha/2}$(s_d)	-0.190	=B14-B16*B15
19	d+$z_{\alpha/2}$(s_d)	1.532	=B14+B16*B15

Approximate confidence intervals for *d*.

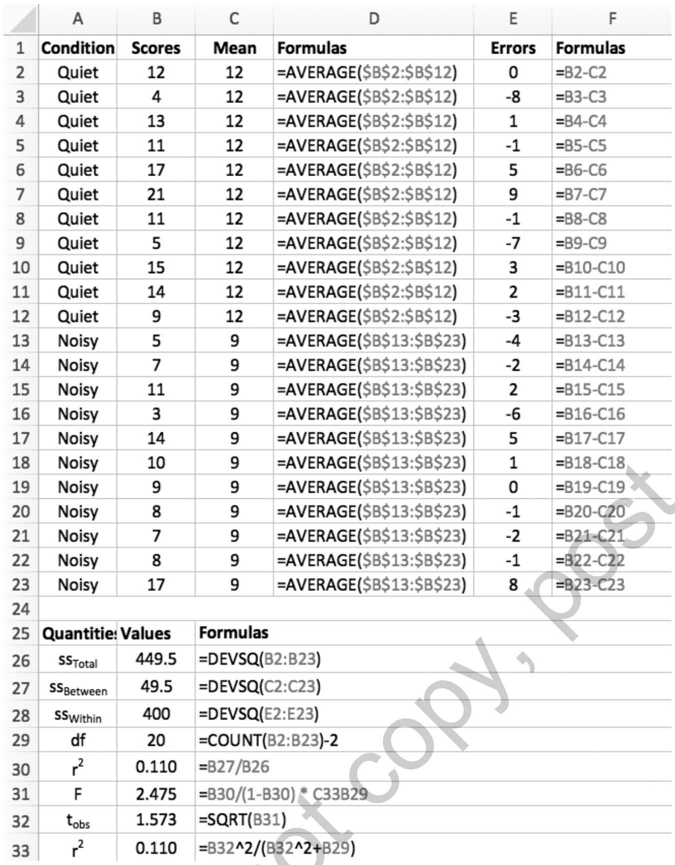
Partitioning Variance

Figure 11.A1.3 illustrates how variability in the dependent variable is partitioned into within- and between-groups sources. Cells **B2** to **B23** are the raw data from Table 11.2. The mean of each condition is computed in cells **C2** to **C23** using the **AVERAGE** function. (The '**$**' signs shown to the right in column **D** have to do with absolute referencing, which was introduced in Appendix 2.1.) Cells **E2** to **E23** show the deviations of scores from their group mean. For example, cell **B13** shows a score of 5 in the noisy condition, which has a mean of 9. Therefore, the deviation score is $5 - 9 = -4$, as shown in cell **E13**.

The **DEVSQ** function has been used to compute ss_{total} (cell **B26**), ss_{between} (cell **B27**), and ss_{within} (cell **B28**). Note that $ss_{\text{total}} = ss_{\text{between}} + ss_{\text{within}} = 49.5 + 400 = 449.5$. Cell **B30** shows the calculation of r^2 as

$$r^2 = \frac{ss_{\text{between}}}{ss_{\text{total}}}.$$

FIGURE 11.A1.3 ■ Partitioning Variance



	A	B	C	D	E	F
1	Condition	Scores	Mean	Formulas	Errors	Formulas
2	Quiet	12	12	=AVERAGE($B$2:$B$12)	0	=B2-C2
3	Quiet	4	12	=AVERAGE($B$2:$B$12)	-8	=B3-C3
4	Quiet	13	12	=AVERAGE($B$2:$B$12)	1	=B4-C4
5	Quiet	11	12	=AVERAGE($B$2:$B$12)	-1	=B5-C5
6	Quiet	17	12	=AVERAGE($B$2:$B$12)	5	=B6-C6
7	Quiet	21	12	=AVERAGE($B$2:$B$12)	9	=B7-C7
8	Quiet	11	12	=AVERAGE($B$2:$B$12)	-1	=B8-C8
9	Quiet	5	12	=AVERAGE($B$2:$B$12)	-7	=B9-C9
10	Quiet	15	12	=AVERAGE($B$2:$B$12)	3	=B10-C10
11	Quiet	14	12	=AVERAGE($B$2:$B$12)	2	=B11-C11
12	Quiet	9	12	=AVERAGE($B$2:$B$12)	-3	=B12-C12
13	Noisy	5	9	=AVERAGE($B$13:$B$23)	-4	=B13-C13
14	Noisy	7	9	=AVERAGE($B$13:$B$23)	-2	=B14-C14
15	Noisy	11	9	=AVERAGE($B$13:$B$23)	2	=B15-C15
16	Noisy	3	9	=AVERAGE($B$13:$B$23)	-6	=B16-C16
17	Noisy	14	9	=AVERAGE($B$13:$B$23)	5	=B17-C17
18	Noisy	10	9	=AVERAGE($B$13:$B$23)	1	=B18-C18
19	Noisy	9	9	=AVERAGE($B$13:$B$23)	0	=B19-C19
20	Noisy	8	9	=AVERAGE($B$13:$B$23)	-1	=B20-C20
21	Noisy	7	9	=AVERAGE($B$13:$B$23)	-2	=B21-C21
22	Noisy	8	9	=AVERAGE($B$13:$B$23)	-1	=B22-C22
23	Noisy	17	9	=AVERAGE($B$13:$B$23)	8	=B23-C23
24						
25	Quantitie	Values	Formulas			
26	SS_{Total}	449.5	=DEVSQ(B2:B23)			
27	$SS_{Between}$	49.5	=DEVSQ(C2:C23)			
28	SS_{Within}	400	=DEVSQ(E2:E23)			
29	df	20	=COUNT(B2:B23)-2			
30	r^2	0.110	=B27/B26			
31	F	2.475	=B30/(1-B30) * C33B29			
32	t_{obs}	1.573	=SQRT(B31)			
33	r^2	0.110	=B32^2/(B32^2+B29)			

Partitioning the variance in scores from two independent samples.

Cell **B31** shows the calculation of the *F*-statistic as

$$F = \frac{r^2}{1-r^2} df.$$

And as noted in this chapter, *t* is the square root of *F*, as shown in cell **B32**. (Note that 1.573 is the value of t_{obs} calculated in the chapter.) Finally, r^2 is calculated a second time in cell **B33** using the formula

$$r^2 = \frac{t^2_{obs}}{t^2_{obs} + df}.$$

APPENDIX 11.2: ESTIMATION AND SIGNIFICANCE TESTS IN SPSS

To compute a confidence interval around the difference between independent sample means in SPSS, we first enter our data as a single column of numbers. In Figure 11.A2.1a the column labeled N_Solved contains the scores from Table 11.2 arranged in a single column. To the right, in the column labeled Group, are numbers used to identify the groups. In this example, 1s corresponds to the quiet condition and 2s correspond to the noisy condition.

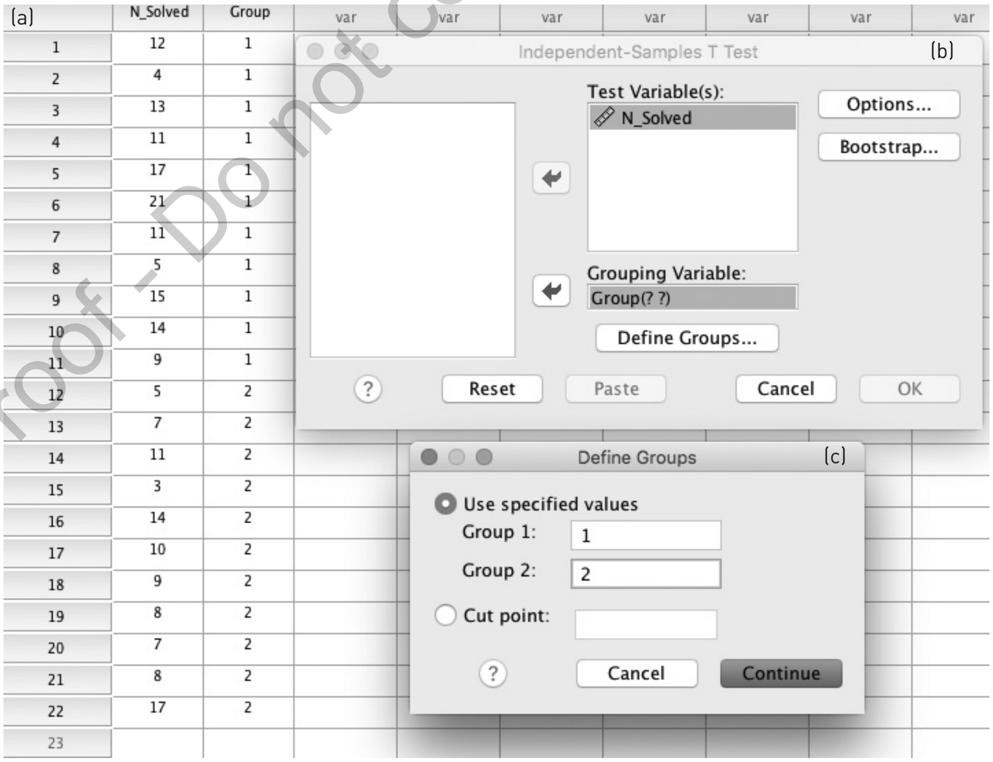
To compute a confidence interval and significance test for these two groups of scores, we choose the Analyze→Compare Mean→Independent-Samples T Tests . . . menu. When this has been chosen, the Independent-Samples T Tests dialog appears, as shown in Figure 11.A2.1b. The variable N_Solved has been moved into the Test Variable(s): region and Group has been moved into the Grouping Variable: region. Before we can proceed, we are required to identify the groups (the numbers in the Group variable) in the analysis. To do this, we click on the Define Groups . . . button, and the Define Groups dialog appears. As shown in Figure 11.A2.1c, 1s are associated with Group 1 and 2s are associated with Group 2. (These are values that I entered.) With this done, we click Continue to return to the Independent-Samples T Tests dialog, and there we click OK to proceed with the analysis.

The output of the analysis is shown in Figure 11.A2.2. There are two rows of numbers. The top row shows the analysis when the population variances are assumed to be equal (i.e., $\sigma_1 = \sigma_2$), and the second row shows the analysis when population variances are not assumed to be equal. The first two columns show the results of Levene's test for equal variances. We won't discuss this analysis, but it is in essence a kind of significance test. If the *p*-value (Sig.) associated with the computed statistic (F) is small, then a statistically significant difference in the two sample variances exists.

The next column (t) shows t_{obs} computed exactly as described in the body of the chapter. The degrees of freedom are $n_1 + n_2 - 2$ when equal variances are assumed. When equal variances are not assumed, the degrees of freedom are adjusted downward using the Welch-Satterthwaite procedure described in Appendix 11.4 (available at study.sagepub.com/gurnsey). In this case, there has been very little adjustment because the sample variances are very similar. The two *p*-values [Sig (2-tailed)] are slightly different because they are based on the same t_{obs} (1.573) but slightly different degrees of freedom.

The difference between the two means (Mean Difference) and the estimated standard error (Std. Error Difference) used to compute t_{obs} are shown in the next two columns. These two quantities do not depend on whether the population variances are assumed to be equal. The last two columns show the 95% confidence intervals for the difference between two means. The interval computed in

FIGURE 11.A2.1 Independent-Groups Analysis



	N_Solved	Group
1	12	1
2	4	1
3	13	1
4	11	1
5	17	1
6	21	1
7	11	1
8	5	1
9	15	1
10	14	1
11	9	1
12	5	2
13	7	2
14	11	2
15	3	2
16	14	2
17	10	2
18	9	2
19	8	2
20	7	2
21	8	2
22	17	2
23		

Test Variable(s): N_Solved

Grouping Variable: Group(? ?)

● Use specified values

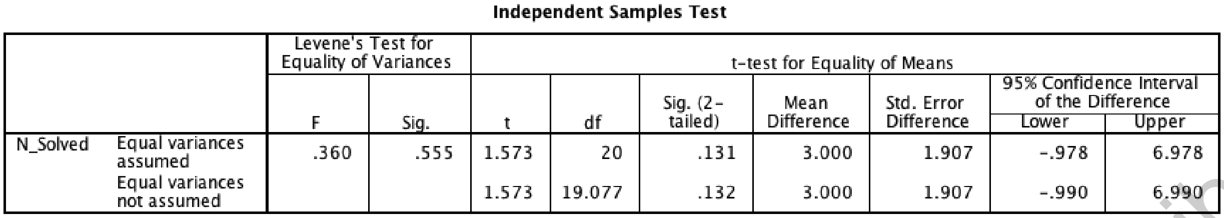
Group 1: 1

Group 2: 2

○ Cut point:

Conducting an independent-groups analysis in SPSS.

FIGURE 11.A2.2 ■ Independent-Groups Output



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
N_Solved	Equal variances assumed	.360	.555	1.573	20	.131	3.000	1.907	-.978	6.978
	Equal variances not assumed			1.573	19.077	.132	3.000	1.907	-.990	6.990

Conducting an independent-groups analysis in SPSS.

the first row is exactly the same as the one computed in the chapter. The second row shows the Welch-Satterthwaite corrected confidence interval. In this case, $t_{\alpha/2}$ is based on 19.077 degrees of freedom rather than 20, which means $t_{\alpha/2}$ is a little larger than in the first row and hence the confidence interval is a little wider.