9 SIGNIFICANCE TESTS

Problems and Alternatives

INTRODUCTION

You may be surprised to learn that statistics is currently seething with controversy. People do not disagree about basic things like sampling distributions. Rather, the controversy centers on the use of significance tests, which are by far the most widely used data analysis methods in psychology. People can get quite worked up about these issues (Bakan, 1966; Carver, 1978; Cohen, 1994; Lambdin, 2012; Rozeboom, 1960), so it can be very entertaining to read these debates.

When psychologists approach a research question, we reflexively form our questions in terms like "Does this intervention or experimental manipulation *work*?" When we state things this way, we really mean "Is there a statistically significant effect of our experimental manipulation?" Significance tests seem to provide an elegant way to make decisions about our questions, so what could be the problem? And, if there is a problem, what might be a better approach to data analysis?

In this chapter, we'll summarize some of the most frequently aired concerns about significance tests and then show how the routine use of estimation can go a long way toward addressing them. Rather than asking whether an intervention works (yes or no), it might be better to ask *how well* an intervention works.

SIGNIFICANCE TESTS UNDER FIRE

Given the prevalence of significance tests in psychology, you might think that all researchers endorse them. This is not so. Consider the following quote from Gerd Gigerenzer (2004), who has long criticized the use of significance tests in psychology:

> You would not have caught Jean Piaget [conducting a significance test]. The seminal contributions by Frederick Bartlett, Wolfgang Köhler, and the Noble laureate I. P. Pavlov did not rely on *p*-values. Stanley S. Stevens, a founder of modern psychophysics, together with Edwin Boring, known as the "dean" of the history of psychology, blamed Fisher for a "meaningless ordeal of pedantic computations" (Stevens, 1960, p. 276). The clinical psychologist Paul Meehl (1978, p. 817) called routine null hypothesis testing "one of the worst things that ever happened in the history of psychology," and the behaviorist B. F. Skinner blamed Fisher and his followers for having "taught statistics in lieu of scientific method" (Skinner, 1972, p. 319). The mathematical psychologist R. Duncan Luce (1988, p. 582) called null hypothesis testing a "wrongheaded view about what constituted scientific progress" and the Nobel laureate Herbert A. Simon (1992, p. 159) simply stated that for his

> research, the "familiar tests of statistical significance are inappropriate." (Gigerenzer, 2004, pp. 591–592)

There are some really strong words in this quote from people you've probably read about. So, let's try to understand where these criticisms come from. Because many criticisms of significance tests are connected to the publication process, we will have to say a few things about this before moving on to the criticisms.

The Publication Process

I mentioned in Chapter 1 that most of your professors think of themselves as researchers. A routine part of research is publishing our research results in academic journals so that others can discuss them. Publishing is generally fun and exciting because it is one of the most important ways of engaging in a public conversation about research questions that are interesting to us. However, publishing is not an optional part of the job for your professors. They are expected to publish regularly, and their job performance is based on the number (and quality) of the papers they publish. Researchers who don't, or can't, publish their research results will not succeed, and they may lose their jobs. Because research productivity determines how they are viewed by their universities and professional colleagues, there is tremendous pressure on professors to publish. To understand some of the concerns about significance tests, we need to think about the process that researchers go through to get the results of their research published in academic journals.

Figure 9.1 illustrates the publication process. A professor typically has a laboratory housed in her university. In collaboration with other professors, graduate students, and research assistants, she runs experiments and collects data. When she thinks the results of her experiments answer her research question, she writes a paper describing the experiments, the results, and what the results mean.

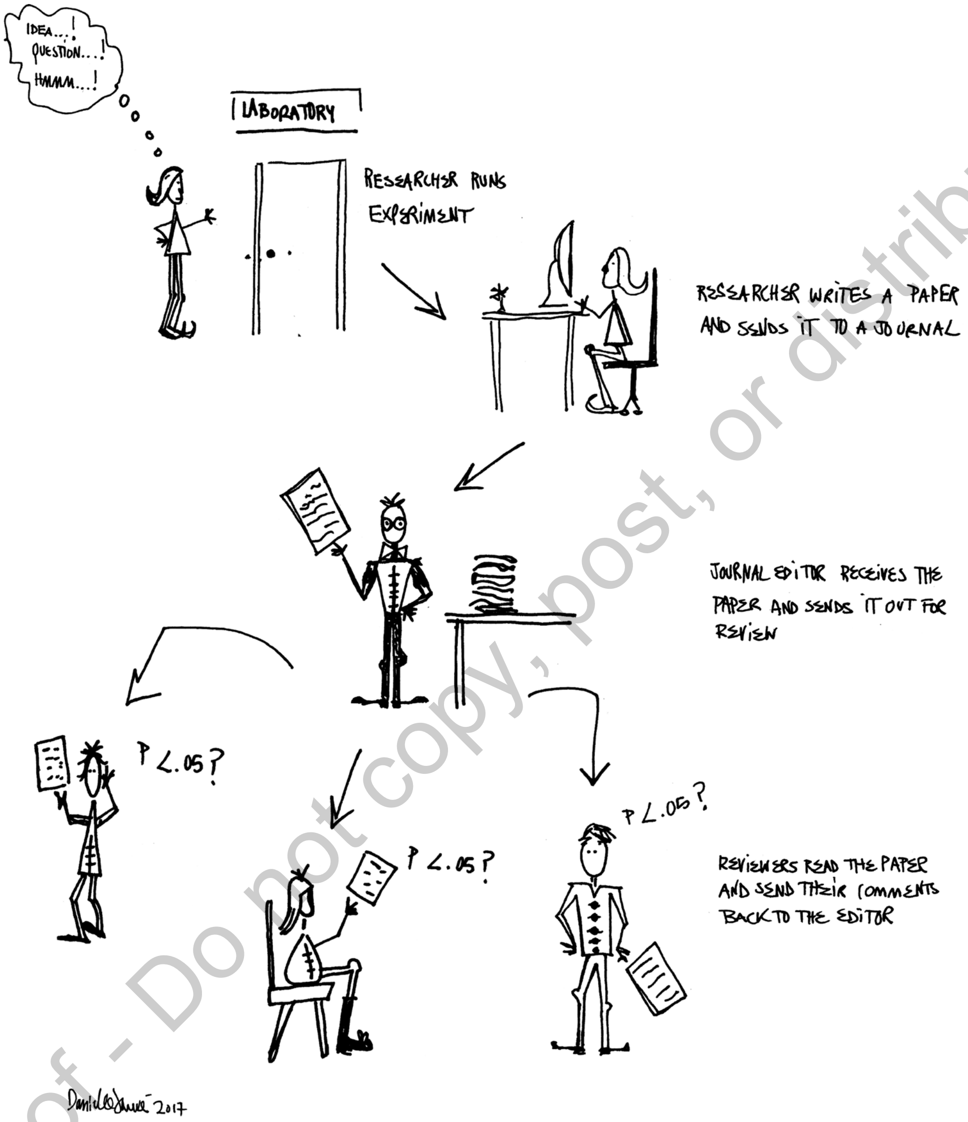
When the paper is finished, the professor sends it to a research journal, where it is assigned to an editor. It is the editor's responsibility to ensure that the journal publishes high-quality research. Therefore, the editor sends the paper to two or three experts in the relevant field and asks them to read the paper to make sure that the experiments were properly run, that the statistical analyses are sound, and that the conclusions make sense.

Because the experts reviewing the paper work in the same field as the author, they probably know her from her previous publications or from meeting her at conferences. However, the reviews are typically anonymous so that the reviewers can feel free to express any concerns they have about the quality of the research. The reviewers want to be thorough but fair. Their job is to provide useful comments in a report to the editor that will help him decide whether to accept the paper.

When the editor receives the reports from the reviewers, he may be able to make a decision to accept or reject the paper right away. Very often, however, the reviewers will find the paper interesting but needing improvement. For example, the author may have failed to acknowledge relevant research from another researcher. Or the reviewers may find the conclusions unconvincing and ask for more experiments to be run or more analyses to be conducted. In such cases, the editor may ask the author to do additional work and then submit a revised version of the paper. The revised paper may be sent back to the same reviewers to see if their concerns have been answered. There can be several rounds of revisions and reviews before the editor makes a final decision to accept or reject the paper.

If the paper is accepted, it will be published in the journal and other scientists will be able to read about the research. If the paper is rejected, it will not be published in that journal, and the author will have to either look for another journal to publish it or give up and store the paper away in a filing cabinet.

FIGURE 9.1 ■ The Publication Process


IDEA...!
QUESTION...!
HMMM...!
LABORATORY
RESEARCHER RUNS EXPERIMENT
RESEARCHER WRITES A PAPER AND SENDS IT TO A JOURNAL
JOURNAL EDITOR RECEIVES THE PAPER AND SENDS IT OUT FOR REVIEW
P < .05 ?
P < .05 ?
P < .05 ?
REVIEWERS READ THE PAPER AND SEND THEIR COMMENTS BACK TO THE EDITOR

A researcher runs an experiment and collects and analyzes data. She then writes a paper (manuscript) describing the experiment, the results, and what the results mean. She sends the paper to a journal, where it is assigned to an editor. The editor sends the paper to experts in the field and asks for their opinions on the merits of the paper. When the editor receives the reports from the experts, he makes a decision about whether to accept and then publish the paper or to reject it.

Figure courtesy of Danielle Sauvé.

Publication and Statistical Significance

At the heart of many research papers are claims such as A causes B. For example, we might claim that assuming a power pose for 2 minutes causes an increase in final exam grades, or that an additional 20 minutes of phonics instruction improves the reading scores of first-grade students. In the simplest case, such claims involve comparing two means (e.g., m and μ_0), computing a test statistic (e.g., z_{obs}), and determining its p-value under the null hypothesis. If $p < .05$, the result is considered statistically significant and the claim may be supported, assuming there are no problems that undermine the interpretation. Journal editors and reviewers often rely heavily on significance tests to judge whether claims are supported. In this way, statistical significance acts as a kind of filter that determines whether a paper is published and thus made available for other researchers to discuss. Unfortunately, many problems arise from the requirement for statistical significance.

CRITICISMS OF SIGNIFICANCE TESTS

The File-Drawer Problem

A major problem in psychology is the reluctance of journals and journal editors to publish papers in which statistically significant results have not been found. This form of **publication bias** means that many interesting results won't make it into the literature because the results were not supported by statistical significance. Such results may be filed away and thus not shared with other researchers, creating what we call the **file-drawer problem**.

Publication bias "occurs whenever the research that appears in the published literature is systematically unrepresentative of the population of completed studies" (Rothstein, Sutton, & Borenstein, 2006). One form of publication bias occurs when journals, editors, reviewers, and even authors favor publication of results that achieve statistical significance.

The **file-drawer problem** refers to the large number of papers filed away in cabinets (or hard drives) because they were unpublished. As a consequence, many valid and worthwhile results are not available to guide and inform other researchers.

The file-drawer problem means that results in the published literature are not representative of all results obtained from studies addressing the same question. Imagine that 16 studies independently addressed the effectiveness of a particular treatment for attention-deficit/hyperactivity disorder (ADHD). Let's say that a quarter of the studies found a statistically significant reduction in ADHD symptoms, and the other three-quarters found reductions that weren't statistically significant. If only the statistically significant results are published, they will not represent the effectiveness of the treatment.

Later in this chapter we will see that the population effect size $[\delta = (\mu_1 - \mu_0)/\sigma]$ can be estimated from the sample mean $[d = (m - \mu_0)/\sigma]$. If we average the estimated effect sizes obtained in the four published studies, the resulting mean will overestimate the size of the effect in question. That is, the average of the four published effect sizes will be greater than the average of all 16 studies (including both published and unpublished). Averaging only effect sizes that made it through the $p < .05$ filter is like computing the class average on a statistics test from only those people whose grades exceeded a threshold of 75%.

Publishing only statistically significant results clearly distorts the literature and results in a misleading representation of the full body of evidence relating to a given question. This is a dangerous situation if the studies relate to health outcomes, such as the effectiveness of pharmacological treatments for depression (Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008).

Proliferation of Type I Errors

A publication bias favoring statistically significant results leads to the strong possibility that many if not most published results are Type I errors (Ioannidis, 2005). Let's do a thought experiment and consider a theory predicting that a daily dose of 1000 mg of vitamin C increases IQ. I doubt this theory is true because I just made it up. If several research groups (possibly funded by the vitamin C industry) studied this theory, then most studies would fail to find a statistically significant effect because the null hypothesis is true. However, it is inevitable that some studies will find statistically significant results; i.e., Type I errors. In a world in which publication bias favors statistically significant findings, these Type I errors would have a higher probability of being published than those that failed to reject the null hypothesis. If these Type I errors are published, then anybody reviewing the literature pertaining to this theory about vitamin C and IQ would conclude that it has been supported because they would not know about the many studies, hidden away in filing cabinets, that correctly retained the null hypothesis.

A second form of publication bias favors novelty. If I predicted, long before Carney et al. (2010), that holding a power pose for 2 minutes would increase final exam grades, I think most people would have found this prediction implausible. Therefore, if I ran the experiment and found no such increase, people would be unsurprised and it would probably be very hard to get the paper published. However, if the same experiment produced a statistically significant increase in grades, this would be viewed as an exciting new finding and the chances of being published would be much greater.

A publication bias favoring novelty is compounded by the fact that it is the policy of some journals not to publish replications of previously reported, statistically significant results. A notorious example of this happened recently when Bem (2011) published a paper in the *Journal of Personality and Social Psychology* titled "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." Here is how one of the experiments was described to the participants:

> [T]his is an experiment that tests for ESP. It takes about 20 minutes and is run completely by computer. First you will answer a couple of brief questions. Then, on each trial of the experiment, pictures of two curtains will appear on the screen side by side. One of them has a picture behind it; the other has a blank wall behind it. Your task is to click on the curtain that you feel has the picture behind it. The curtain will then open, permitting you to see if you selected the correct curtain. There will be 36 trials in all.
>
> Several of the pictures contain explicit erotic images (e.g., couples engaged in nonviolent but explicit consensual sexual acts). If you object to seeing such images, you should not participate in this experiment. (Bem, 2011, p. 409)

The novel twist in the experiment was that the window showing the picture was chosen at random by a computer *after* the participant had made his or her choice. Therefore, the choice is (arguably) about the future state of the world.* The null hypothesis in this case is that participants would have a 50% chance of guessing which of the two curtains hid the erotic image. However, it was found that participants guessed correctly 53% of the time, *on average*, and this turned out to be statistically significant. Bem concluded that these results constitute positive evidence that people can see or sense future events.

If this study truly demonstrated that people can "feel the future," it would be the most important experiment ever reported in human history, and every domain of science would have to be revised fundamentally in view of it. If any experiment calls out for replication to ensure that it is not a Type I error, it is this one. However, when a paper reporting an unsuccessful attempt to find the same results was submitted to the same journal, the editor rejected it, saying that it was the journal's policy to publish only original studies and not replications. The editor in question was quoted in the *New Scientist* as saying, "This journal does not publish replication studies, whether successful or unsuccessful" (Aldhous, 2011).

This episode illustrates the unfortunate fact that Type I errors are far easier to get into the literature than to remove from the literature. Science is supposed to be self-correcting, but when journals devalue replication, errors become difficult to correct.**

*I'm not sure why it wasn't taken as evidence for participants reading the current state of the random number generator in the computer through extrasensory perception.

**A bit of hopeful news here is that the public outcry over this event caused the *Journal of Personality and Social Psychology* to accept attempted replications of the Bem experiments (Galak, LeBoeuf, Nelson, and Simmons, 2012). Not surprisingly, Galak et al. did not find any evidence that people can "feel the future."

p-Hacking: The Quest for Statistical Significance

Because there is a publication bias favoring results that are statistically significant, researchers naturally feel pressured to find statistically significant results. This pressure does not necessarily result in dishonest behavior, but it can result in a multitude of questionable practices in which researchers make undisclosed adjustments to their data analysis procedure in efforts to produce statistically significant results. These practices are collectively known as ***p*-hacking**. Uri Simonsohn defined *p*-hacking as "trying multiple things until you get the desired result" (Nuzzo, 2014), where the desired result is statistical significance.

The term ***p*-hacking** refers to a multitude of questionable practices in which researchers make undisclosed adjustments to their data analysis procedure in efforts to produce statistically significant results.

We encountered a simple example of *p*-hacking in Chapter 7 when we noted that a researcher may decide, after the experiment has been run, to test for statistical significance using a one-tailed test rather than a two-tailed test as originally planned. Rex Kline (2013) refers to this as an instance of hypothesizing after the results are known, or HARKing.

A very common type of *p*-hacking involves running many experiments, or conducting multiple analyses on the same data, and reporting only the results that are statistically significant or those that are consistent with the predictions of the researcher. Such practices very often create a misleading picture of reality and increase the likelihood that Type I errors will make their way into the literature. This type of *p*-hacking is brilliantly illustrated by an xkcd cartoon that can be found at xkcd.com/882. The cartoon shows twenty significance tests, each one assessing the link between acne and jelly beans of a specific color. In one case (green jelly beans) a statistically significant result was found. If only this result is reported, as in the cartoon, then a reader has no idea about the other 19 analyses that did not produce statistically significant results. That is, the reader has no way of knowing that the result is almost certainly a Type I error. (If you ever need a break from studying statistics, visit xkcd.com for hours of top-notch entertainment. Some comics are NSFW.)

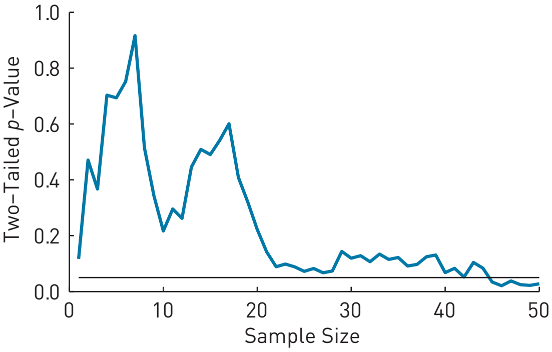
Another form of *p*-hacking is when researchers increase their sample size in hopes of eventually getting a statistically significant result. Let's think back to my theory about vitamin C and IQ. We will assume that IQ is normally distributed in college students, with a mean of 115 and standard deviation of 15; this is my null hypothesis distribution of scores. I then take a random sample of 25 students and put them on a high vitamin C diet for a month, after which I measure their IQs. I find that their mean IQ is 118.6, and from this, I compute $z_{\text{obs}} = (118.6 - 115/3 = 1.2, p = .12$. This result is not statistically significant, but the difference is in the predicted direction and the *p*-value is sort of small. I might conclude that my test lacked power because the sample size was too small. Therefore, I continue to add participants to my sample and test for statistical significance after each additional participant is added. Finally, I end the experiment when I find a statistically significant result.

We could call this the "run-until-statistically-significant" (RUSS) approach. Many researchers believe this is a perfectly legitimate strategy; I know I did for a long time. However, it is definitely not okay to do this. The problem with the RUSS approach is that even when the null hypothesis is true, there will be random variations in the statistic (and its associated *p*-value) as we add more participants to our sample. At some point, the statistic (e.g., z_{obs}) may become statistically significant because of these chance fluctuations.

The random fluctuations in *p*-values are illustrated in Figure 9.2. Starting with a sample of size 1, *drawn from the null hypothesis distribution*, we compute z_{obs} and record its *p*-value. We then add another score to the sample and test for statistical significance. We repeat this process until sample size is 50. Note that our 50 samples are not independent samples of different sizes. Rather, a sample of size *n* contains all the scores from the sample of size *n*−1, plus one more.

In Figure 9.2, sample size is plotted on the *x*-axis and *p*-values on the *y*-axis. As you can see, the *p*-values "wander around" as sample size increases. In this case, the *p*-value

FIGURE 9.2 ■ *p*-Hacking: RUSS



An illustration of how *p*-values can change as sample size increases. The sample is built up over time by drawing scores from the null hypothesis distribution. We start with a sample size of 1, and we add another randomly selected score on each of the remaining 49 trials. On each trial, z_{obs} and its two-tailed *p*-value are computed. The *p*-values change over time. In this example, the *p*-value drops below the $p < .05$ criterion for statistical significance (indicated by the black horizontal line) at trial 45. The RUSS strategy ends the experiment when *p* drops below the $p < .05$ criterion.

eventually drops below the $p < .05$ criterion for statistical significance, indicated by the black horizontal line. If data collection stops at this point, we might be inclined report that the obtained z_{obs} is statistically significant, $p < .05$. As we will see, however, the actual *p*-value associated with z_{obs} is much greater than .05.

Imagine running the process shown in Figure 9.2 10,000 times. The process stops when either (i) sample size is 50 or (ii) the sample produces a two-tailed z_{obs} for which $p < .05$. In this situation, about 30% of these 10,000 trials will stop because the $p < .05$ criterion has been obtained. The trials comprising this 30% are Type I errors because the scores were drawn from the null hypothesis distribution! The *p*-value associated with our z_{obs} statistic is clearly much greater than .05. Therefore, widespread use of the RUSS approach will inflate the number of Type I errors in the literature. As we've seen, Type I errors have an increased chance of being published, and once published it is difficult to publish replication studies to correct the error.

Binary Thinking

The quest for statistical significance produces a tendency to think of statistically significant results as important and meaningful, and results that are not statistically significant as unimportant and meaningless. As we will see at the end of the chapter, it is possible for two very similar results to differ in their statistical significance (e.g., $p = .04$ versus $p = .06$). It is silly to treat these two results as different in any meaningful way, but a reliance on statistical significance can seduce us into doing just that. Worse, psychologists in the past were instructed to think in exactly this way by the *APA Publication Manual*. The effects of this guidance are still with us.

Statistical Significance Versus Practical Significance

We've already discussed the important difference between statistical significance and practical significance. One effect of the $p < .05$ filter is that achieving statistical significance can become the goal of the research, rather than an aid to reasoning. Consequently, one might consider the job done once the null hypothesis has been rejected. We've seen, however, that statistical significance does not mean that the result is important or meaningful. Unfortunately, when the focus is on statistical significance, we may spend less time than we should considering the practical significance of a result. This can lead to very shallow research in which a large number statistically significant results "don't amount to a hill of beans," as Humphrey Bogart said to Ingrid Bergman in *Casablanca*. (You might have to Google this.)

It is a useful thought experiment to consider what would happen if researchers were not allowed to use significance tests (Harlow, Mulaik, & Steiger, 2016). In this case, a result would not carry weight just because it is statistically significant. Instead, researchers would have to explain things such as (i) why the magnitude of the difference is important, (ii) how it changes our view of a particular theory, or (iii) what it implies about the effect of some treatment on some population. I don't mean to imply that researchers don't do this at all. Rather, it is a question of balance. In my view, a result having low practical or theoretical significance can gain unwarranted stature if it is statistically significant.

Misunderstanding *p*-Values

A frequent complaint about significance tests is that many researchers don't know what statistical significance means. We saw in Chapter 7 that a *p*-value is a conditional probability. It is the probability of the obtained statistic, or one more extreme, occurring by chance when the null hypothesis is true. However, as noted by Cohen more than 20 years ago, many people who should know better are prone to committing the inverse probability error, which is the belief that the *p*-value represents the probability that the null hypothesis is true. As Cohen (1994) argues, the inverse probability error is really a case of wishful thinking. Researchers really want to know the probability of the null hypothesis being true, and they let their wishes lead them to believe that this is what *p* tells them, when it does nothing of the sort.

There are other misinterpretations of *p*-values documented by Oakes (1986), Haller and Kraus (2002), Carver (1978), and Kline (2004, Chapter 3), among many others. Among these misinterpretations are the following:

1. The belief that a *p*-value is the probability that you've made a Type I error. We saw in Chapter 8 that this is a confusion between *p* and α, arising from the hybrid Fisher-Neyman model. The correct interpretation is that α is the probability that *the test* will produce a Type I error, and *p* is the conditional probability of the statistic (derived from *the data*) under the null hypothesis. Unfortunately, someone believing that $p = .0062$ is the probability that a Type I error has been committed would have little reason to think that the study should be replicated.
2. The belief that 1–*p* is the probability that the same experiment will produce a statistically significant result if it is repeated. An innocent-sounding instance of this occurs when researchers say that small *p*-values indicate *reliable* (presumably replicable) results. It is true that there is a relationship between *p* and the probability of replication, but the relationship is not a simple one. The probability that a result will replicate is certainly not 1–*p*. The consequences of this mistaken belief can be quite bad. If you think that 1–*p* is the probability that the same experiment will produce a statistically significant result if it is repeated, then when $p = .0062$ you think the probability is .9938 that your experiment will replicate. This misunderstanding also might lead one to think that doing an actual replication would be a waste of time.
3. The belief that 1–*p* is the probability that the alternative hypothesis is true. The Fisher-Neyman hybrid model probably explains this confusion. In an acceptance procedure, one of two hypotheses is accepted. When this is merged with a significance test, then rejecting H_0 is equivalent to accepting H_1. So, what do we do with a *p*-value in this case, given that it plays no role in an acceptance procedure? Well, it seems just a short step to see the *p*-value as a measure of the probability that H_0 is true (the inverse probability error) and 1–*p* as the probability that H_1 is true. This is impossible; *p*-values are derived from the assumption that H_0 is true, with no consideration of H_1.

Misunderstandings about *p*-values are so widespread that the American Statistical Association, in a highly unusual move, published a statement about the meaning of *p*-values (Wasserstein & Lazar, 2016). It is not a good thing when a professional does not understand the most commonly used tool in his or her toolbox. We'd be concerned about a surgeon who uses the wrong end of a scalpel.

In summary, statistical significance is a kind of filter that determines which papers make it into the publicly available research literature. This means that the published literature may seriously misrepresent the size of some effect. We would have a much better sense of any given effect if we knew about all relevant results, not just those that passed through the $p < .05$ filter. Furthermore, the criterion for statistical significance ensures that if some effect *does not exist*, it is more likely that a paper reporting that the effect does exist (a Type I error) will enter the literature than a paper that says it doesn't exist (correctly retains the null hypothesis). In a world in which replication is not valued, Type I errors will linger in the literature for years, misleading many other researchers and thus wasting their time and slowing the progress of science. This problem is compounded by the p-hacking that occurs when researchers (who are under pressure to publish) selectively report data in order to present a story that ends with a rejected null hypothesis that supports their claims. An emphasis on statistical significance also leads to a black-and-white worldview in which only results that are statistically significant are treated as interesting and meaningful. Equating statistical significance with significance can lead to very shallow thinking. If we add to these points reports that many researchers misunderstand the meaning of a p-value, then I'd say we have a problem.

CONFIDENCE INTERVALS

I have seen speakers deliver eloquent summaries of problems like those listed above to rooms full of researchers who routinely use significance tests as part of their professional activities. The really interesting thing is that it is rare for researchers to rise and defend significance tests against these criticisms. The most common response is, "True, true, but what's the alternative?" This dynamic also exists in the statistics literature in general. Kline (2013) notes that there are hundreds of papers criticizing significance tests, and there are only a handful of defenses. So, what is the alternative?

For many, estimation with confidence intervals provides the healthiest alternative to significance tests. Significance tests and confidence intervals rest on the same theoretical foundations (i.e., sampling distributions) but differ in their objectives. Significance tests are designed *to decide* what a parameter *is not*, whereas confidence intervals are designed *to estimate* what a parameter *is*. When you reject the null hypothesis, you are rejecting a specific hypothesis about the mean of the distribution from which a sample was drawn. When you construct a confidence interval, you are providing your best estimate about the mean of the distribution from which a sample was drawn. When we move attention away from decision making based on the $p < .05$ criterion, we can focus on the size of an effect and its practical significance.

DECISIONS VERSUS ESTIMATION

Significance tests are designed to decide what a parameter *is not*, whereas confidence intervals are designed to estimate what a parameter *is*.

The Advantages of Confidence Intervals

In my view, (almost) any steps that eliminate the $p < .05$ filter from the publication process in psychology and related disciplines will produce improvements over the current state of affairs. Without the $p < .05$ filter, there would be no need for p-hacking or binary thinking, less of a file-drawer problem, and fewer Type I errors lodged in the literature. Furthermore, without the $p < .05$ crutch, there would be more emphasis on replication and greater effort to explain the practical significance of a result.

LEARNING CHECK 1

1. With regard to the proper interpretation of *p*-values, explain why the following are wrong:
   (a) Somebody tells you that 1–*p* is the probability that an experiment will replicate.
   (b) Somebody tells you that 1–*p* is the probability that H_1 is true.
   (c) Somebody tells you that *p* is the probability that you've made a Type I error.
2. Provide an example in which a statistically significant result is of no practical significance. (Be creative.)
3. A researcher tells you that he plans to run an experiment to test the effect of THC (the psychoactive agent in marijuana) on a problem-solving task known to have a mean of $\mu_0 = 25$ in the general population. He says he thinks 18 participants should be sufficient to conduct a two-tailed test of the null hypothesis $\mu_1 = \mu_0$. What questions might you ask him that would help him run a better experiment?
4. A researcher wondered whether verbal reasoning might improve following 15 minutes of listening to a Mozart sonata. After running the experiment with 21 participants, he found an improvement of 6 points on a verbal reasoning task, $z_{obs} = 1$, $p = .159$. He thought this was encouraging, but concluded that his sample was not large enough. Therefore, he continued adding participants until he reached 36 participants, at which point he again found a 6-point improvement, $z_{obs} = 1.7$, $p = .045$. Would you have anything to say to this experimenter about the legitimacy of what he's done?
5. A researcher wondered about the effects of alcohol consumption on driving ability. She chose a random sample of 21 college students and measured the number of driving errors they made (on a controlled driving track) once their blood alcohol level was elevated to .05 g/dL, which is below the legal limit for intoxication in most jurisdictions. She found an elevation in driving errors corresponding to $z_{obs} = 1.4$, $p = .081$. Because this increase was not statistically significant according to the $p < .05$ criterion for statistical significance, she concluded that a blood alcohol level of .05 g/dL posed no increased risk of driving accidents. Please comment on this conclusion.

Answers

1. (a) 1–*p* is the probability of obtaining a statistic less extreme than the one you obtained *when the null hypothesis is true*. 1–*p* is not derived from any consideration of the alternative hypothesis and so can't possibly be the probability that the result will replicate.
   (b) For the same reason given in (a), 1–*p* cannot be the probability that H_1 is true.
   (c) This claim confuses *p* with α. α specifies the probability of a Type I error. *p* is the probability of the obtained data (or data more extreme) occurring by chance when the null hypothesis is true. α is a property of the test, whereas *p* is a property of the data.
2. You can choose your own example because infinitely many silly tests can be imagined. If I take a random sample of 121 psychology students and measure the number of push-ups that each can do, I'm pretty sure that the difference between the mean of these 121 scores and $\mu_0 = 200$ would be statistically significant. I can't attach any meaning to this result.
3. I'd ask why he chose $n = 18$ participants. If he were to say, "Well, that seems like enough," you could then ask him to think about what effect size would be interesting to him, and then tell him a little bit about power analysis. On the other hand, he might have said he is only interested in an effect size of $\delta = .6$ or greater, and 18 participants gives him power = .8. In this case, he gets a pat on the head and a big gold star.
4. I would say: "Running the experiment until you find a statistically significant result increases the probability of obtaining statistical significance when the null hypothesis is true. You should have done a prospective power analysis, chosen your sample size, and stuck with it."
5. Failing to reject the null hypothesis does not mean that the null hypothesis is true. In this case, the data suggest that an increased blood alcohol level increases driving errors and in the real world would increase the probability of driving accidents, which often have devastating consequences.

The routine use of confidence intervals would go a long way to realizing these benefits. With confidence intervals, the focus is on the precision of a parameter estimate for some variable of interest. The actual size of a measured difference would be central to our thinking, rather than the probability of the difference under the null hypothesis. Because the focus would be on the size of a measured difference, we would not be drawn into binary thinking. Furthermore, any measured difference would be seen as a single estimate with unavoidable imprecision. Therefore, replication with an eye to combining measures from different studies would be far more common.

Significance Tests With Confidence Intervals

At a purely technical level, estimation with confidence intervals is a far more general method of data analysis than significance tests. In this section we'll see that once you have computed a confidence interval, you can test any null hypothesis of interest simply by asking whether μ_0 falls within the interval. For this reason, confidence intervals are the method recommended by the most recent versions of the *APA Publication Manual*. We will now revisit examples of significance tests from previous chapters from the perspective of computing confidence intervals.

Please note that *I am absolutely not recommending the routine use of confidence intervals to test null hypotheses*! This would be an egregious misuse of estimation. Rather, the main point is to emphasize that significance tests represent a minor feature of what can be done with confidence intervals. Understanding the secondary role of significance tests helps us recognize the poverty of the information they yield.

Whole Language Versus Phonics: Revisited

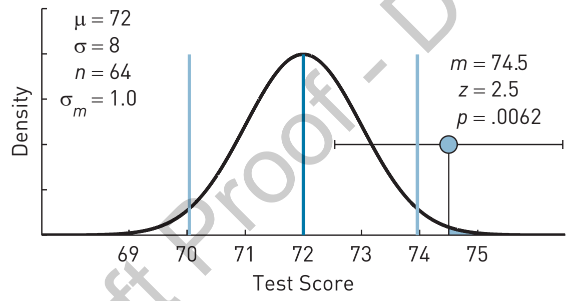
In the phonics example that opened Chapter 7, we were told that the mean of the null hypothesis distribution was $\mu_0 = 72$. We were also given the sample mean ($m = 74.5$), the population standard deviation ($\sigma = 8$), and sample size ($n = 64$). Using σ and sample size, we were able to compute $\sigma_m = \sigma/\sqrt{n} = 8/\sqrt{64} = 1$. This information is exactly what is needed to compute a confidence interval. The 95% confidence interval around m is as follows:

$$\text{CI} = m \pm z_{\alpha/2}(\sigma_m) = 74.5 \pm 1.96(1) = [72.54, 76.46].$$

Therefore, we have 95% confidence that the mean of the distribution from which the sample was drawn is in the interval [72.54, 76.46]. This confidence interval allows us to reject $\mu_0 = 72$ as a plausible hypothesis about the mean of the distribution from which the phonics-enriched students were drawn, *because 72 does not fall within it.*

Figure 9.3 illustrates the connection between confidence intervals and significance tests. The normal distribution in Figure 9.3 is the sampling distribution of the mean under the null hypothesis. Figure 9.3 compares two intervals: $\mu_0 \pm 1.96(\sigma_m)$ (the light blue vertical lines) and $m \pm 1.96(\sigma_m)$ (the confidence interval around the dot, representing the sample mean). The arms of both intervals are the same. Therefore, if m is outside the interval $\mu_0 \pm 1.96(\sigma_m)$, the confidence interval $m \pm 1.96(\sigma_m)$ will not capture μ_0. If m is within the interval $\mu_0 \pm 1.96(\sigma_m)$, the confidence interval $m \pm 1.96(\sigma_m)$ will capture μ_0. (Think back to the arms in the illustrations from Chapter 6 in the section on sampling distributions and confidence.)

FIGURE 9.3 ■ Estimation and Significance Tests



The connection between confidence intervals and two-tailed significance tests. The normal distribution is a distribution of sample means for size $n = 64$, drawn from a population of reading scores having a mean of $\mu_0 = 72$ and standard deviation of $\sigma = 8$. The mean of the sample is $m = 74.5$ and $\sigma_m = 1$. The dot plots the sample mean. The arms around the dot represent the 95% confidence interval. Notice that the interval does not capture μ_0, which is the mean of the distribution, according to the null hypothesis. Therefore, we can reject a two-tailed test of the null hypothesis at the $p < .05$ level.

Another way to see this relationship is to note that when a 95% confidence interval does not capture μ_0, then $z_{obs} = (m - \mu_0)/\sigma_m$ will be outside the interval ±1.96. If the 95% confidence interval captures μ_0, then $z_{obs} = (m - \mu_0)/\sigma_m$ will be within the interval ±1.96. Therefore, any time a 95% confidence interval does not capture μ_0, we can reject the null hypothesis at the $p < .05$ criterion for statistical significance.

If the 95% confidence interval around the sample mean does not capture μ_0 specified by H_0, we are faced with the same two possibilities we face in any significance test. Either H_0 is true, and this is one of those rare times that the sample mean falls a long way from the mean of the distribution (μ_0), or the confidence interval does not capture μ_0 because H_0 is false, and so μ_0 is not the mean of the distribution from which the sample was drawn. Of course, the second interpretation means we reject the null hypothesis. Therefore, in the phonics example, we could reject a two-tailed test of H_0 at the $p < .05$ level because the 95% confidence interval around the sample mean ($m = 74.5$) did not capture the mean specified by H_0 ($\mu_0 = 72$). More importantly, the confidence interval provides evidence of what μ_1 (i.e., $\mu_{phonics}$) *is*, not just what it *is not*.

Differences in IQ: Revisited

In the IQ example from Chapter 7, we were told that the mean IQ of Quebec residents was $\mu_{Que} = 100$ with a standard deviation of $\sigma = 15$. We were also told that the mean IQ of $n = 250{,}000$ people from Maine was $m = 100.06$. Using σ and sample size ($n = 250{,}000$), we were able to compute $\sigma_m = 0.03$. When we compute the 95% confidence interval around m, we find the following:

$$\text{CI} = m \pm z_{\alpha/2}(\sigma_m) = 100.06 \pm 1.96(0.03) = [100.0012, 100.1188].$$

As in the reading example, our confidence interval does not capture the known population mean, $\mu_{Que} = 100$. Therefore, we can reject a two-tailed test of H_0 at the $p < .05$ level. Besides revealing that 100 is an implausible hypothesis about the mean IQ of Maine residents, the confidence interval provides the best estimate of what μ_{Maine} is (i.e., 100.06). It also provides and interval around this mean, and we have 95% confidence that μ_{Maine} is in this interval (95% CI [100.0012, 100.1188]). As with the reading example, confidence intervals provide us with information about what the parameter in question *is*, in addition to what it *is not*.

Bench-Pressing 5-Year-Olds: Revisited

In the bench-pressing example from Chapter 7, we were told that the mean bench-pressing weight for 15-year-old European males was $\mu_{15YO} = 125$ pounds. We were also told that the mean bench-pressing weight for a sample of one hundred 5-year-old European males was $m = 10$ pounds, with sample standard deviation, $s = 2$. Using s and sample size, we were able to compute $s_m = s/\sqrt{n} = 2/10 = 0.2$. When we compute the approximate 95% confidence interval around m, we find the following:

$$\text{CI} = m \pm z_{\alpha/2}(s_m) = 10 \pm 1.96(0.2) = [9.61, 10.39].$$

As in the previous two examples, our confidence interval does not capture the known population mean, $\mu_{15YO} = 125$. Therefore, we can reject H_0 at the $p < .05$ level. In addition to revealing that 125 is an implausible hypothesis about μ_{5YO}, the confidence interval provides the best estimate of μ_{5YO} (i.e., 10). It also provides an interval around this estimated mean, and we have 95% confidence that μ_{5YO} is in this interval (95% CI [9.61, 10.39]). Again, this confidence interval provides us with information about what the parameter in question *is*, in addition to what it *is not*.

As a final point, note that in this example we used the sample standard deviation s in the calculation of s_m. The question of what s estimates does not arise in this situation. The sample standard deviation, s, clearly estimates σ_{5YO}. We didn't have to assume that σ_{15YO} and σ_{5YO} are the same, as we did in Chapter 7, so this makes things a little cleaner.

Retaining the Null Hypothesis

We noted in Chapters 7 and 8 that retaining H_0 does not mean that it is true. This is particularly clear when we test H_0 with a confidence interval. Think of the following null hypothesis in a case where $m = 26$, $\sigma = 8$, and $n = 16$:

$$H_0\colon \mu_0 = 25.$$

The 95% confidence interval around $m = 26$ is computed as follows:

$$\text{CI} = m \pm z_{\alpha/2}(\sigma_m) = 26 \pm 1.96(2) = [22.08, 29.92].$$

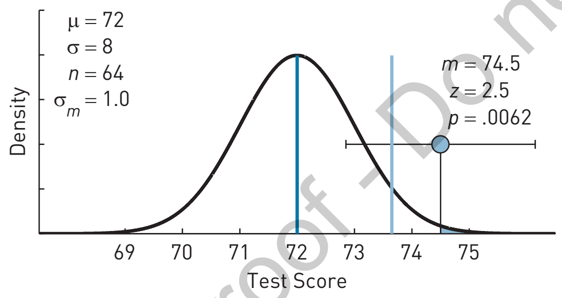
This confidence interval contains $\mu_0 = 25$, so we would retain H_0. In this example, it can be seen that many population means (μ) are consistent with this interval. That is, 25 is just one of many plausible hypotheses about the mean of the population from which the sample was drawn. Therefore, this example makes it very clear that retaining $H_0\colon \mu_0 = 25$ *does not mean* that 25 *is* the population mean. In fact, the most plausible estimate is that $\mu = m$.

*One- and Two-Tailed Tests (Optional Material)

In this section, we will see how to conduct one-tailed tests with confidence intervals. This material is presented for completeness only, and it can be skipped without doing damage to your understanding of the connection between confidence intervals and significance tests.

In the original phonics scenario, the researcher predicted that phonics instruction would improve reading scores. Therefore, her significance test was a one-tailed test in which she predicted $\mu_1 - \mu_0 > 0$. This raises the question of how to conduct a one-tailed significance test using confidence intervals. The easiest way to do this (when $\alpha = .05$) is to compute the 90% confidence interval around the mean, and then reject H_0 *only if* (i) the interval does not capture μ_0 and (ii) the sample mean is on the predicted side of μ_0. We will illustrate this procedure while referring to Figure 9.4.

FIGURE 9.4 ■ Estimation and Directional Tests



The connection between confidence intervals and one-tailed significance tests. This is a modified version of Figure 9.3. The distribution of sample means of size $n = 64$ was drawn from a population of reading scores having a mean of $\mu_0 = 72$ and standard deviation of $\sigma = 8$. The mean of the sample is $m = 74.5$ and $\sigma_m = 1$. The dot plots the sample mean. The arms around the dot represent the 90% confidence interval. Notice that the interval does not capture the μ_0, which is the mean of the distribution, according to the null hypothesis. Only if the alternative hypothesis predicts $\mu_1 - \mu_0 > 0$ could we reject a one-tailed test of the null hypothesis at the $p < .05$ level. If the alternative hypothesis predicts $\mu_1 - \mu_0 < 0$, we would have to retain the null hypothesis.

First, in the phonics example, the 90% CI would be

$$\text{CI} = m \pm z_{\alpha/2}(\sigma_m) = 74.5 \pm 1.645(1) = [72.86, 76.15].$$

This confidence interval is shown in Figure 9.4. The interval does not capture μ_0, so that is the first requirement for concluding that there is a statistically significant difference between m and μ_0. The second requirement is that the sample mean must be on the side of μ_0 predicted by the alternative hypothesis. In this case, the alternative hypothesis predicts m will be greater than μ_0, so this requirement is satisfied as well. Therefore, we can reject the null hypothesis at the $p < .05$ level of statistical significance. As in the case of significance tests with z_{obs}, if the alternative hypothesis had been $\mu_1 - \mu_0 < 0$, we would have retained the null hypothesis.

Having shown that one can do significance tests with confidence intervals, I must reiterate that I don't recommend doing so. If you feel that you must do a significance test, then simply use a standard z-test; i.e., compute z_{obs}. That is, if you find yourself living in the high-contrast world of statistical significance, then use a test statistic such as z_{obs}, which is the coin of the realm.

LEARNING CHECK 2

1. A researcher conducts a two-tailed test of H_0: $\mu_1 - \mu_0 = 0$, where $\mu_0 = 25$. His alternative hypothesis is H_1: $\mu_1 - \mu_0 \neq 0$. In which of the following cases can he reject H_0 at the $p < .05$ significance level: (a) 95% CI [23, 26], (b) 95% CI [23, 24], (c) 95% CI [26, 27], and (d) 95% CI [24, 26]?
2. *A researcher conducts a one-tailed test of H_0: $\mu_1 - \mu_0 = 0$, where $\mu_0 = 25$. His alternative hypothesis is H_1: $\mu_1 - \mu_0 > 0$. In which of the following cases can he reject H_0 at the $p < .0$ significance level: (a) 90% CI [22, 27], (b) 95% CI [23, 24], (c) 90% CI [26, 27], and (d) 90% CI [18, 24]?
3. *A researcher conducts a one-tailed test of H_0: $\mu_1 - \mu_0 = 0$, where $\mu_0 = 25$. His alternative hypothesis is H_1: $\mu_1 - \mu_0 < 0$. In which of the following cases can he reject H_0 at the $p < .05$ significance level: (a) 90% CI [22, 27], (b) 95% CI [23, 24], (c) 90% CI [26, 27], and (d) 90% CI [18, 24]?

Answers

1. (a) No. The interval contains $\mu_0 = 25$.
   (b) Yes. The interval does not contain $\mu_0 = 25$.
   (c) Yes. The interval does not contain $\mu_0 = 25$.
   (d) No. The interval contains $\mu_0 = 25$.
2. (a) No. The interval contains $\mu_0 = 25$.
   (b) No. The interval does not contain $\mu_0 = 25$, but it is on the wrong side of μ_0.
   (c) Yes. The interval does not contain $\mu_0 = 25$, and it is on the predicted side of μ_0.
   (d) No. The interval does not contain $\mu_0 = 25$, but it is on the wrong side of μ_0.
3. (a) No. The interval contains $\mu_0 = 25$.
   (b) Yes. The interval does not contain $\mu_0 = 25$, and it is on the predicted side of μ.
   (c) No. The interval does not contain $\mu_0 = 25$, but it is on the wrong side of μ_0.
   (d) Yes, the interval does not contain $\mu_0 = 25$, and it is on the predicted side of μ_0.

ESTIMATING $\mu_1 - \mu_0$

There is an important point about significance tests that questions whether H_0 could ever be true. Consider pairs of populations, such as the IQs of men and women, the IQs of US citizens living east and west of the Mississippi River, or the heights of 20-year-old men born on odd and even days of the year. In none of these situations would we have any reason to expect a big difference in the means of the two populations. On the other hand, would we ever expect the difference between the two population means to be exactly 0? Not 0.1, 0.0032, or 0.0089, but exactly 0. It seems highly unlikely that the difference between any two populations, no matter how similar, will be exactly 0. If this is the case, there is a good argument that it makes little sense to test a null hypothesis that there is exactly zero difference.

If the null hypothesis that $\mu_1 = \mu_0$ has very little chance of ever being true, then a better approach to data analysis is to estimate the difference between μ_1 and μ_0. Therefore, to judge "how wrong" the null hypothesis is, we can estimate $\mu_1 - \mu_0$. Estimating $\mu_1 - \mu_0$ involves a minor modification to the method covered in the previous section.

In the phonics scenario, the known distribution of reading scores with $\mu_0 = 72$ and $\sigma = 8$ was the null hypothesis distribution. The distribution of reading scores under the alternative hypothesis has a mean (μ_1) that is unknown. However, it is assumed that the standard deviation (σ) of this distribution is equal to the standard deviation of the known (null hypothesis) distribution. From these observations it follows that the sampling distribution of the mean under the alternative hypothesis would have mean μ_1 and a standard error of $\sigma_m = \sigma/\sqrt{n}$.

To estimate $\mu_1 - \mu_0$, we must consider subtracting μ_0 from each mean in this distribution of means under the alternative hypothesis. Doing so will produce a distribution of the statistic $m - \mu_0$. The sampling distribution of $m - \mu_0$ will have mean $\mu_1 - \mu_0$ and standard error $\sigma/\sqrt{n}$. This standard error is just the standard error of the mean that we've been using all along. However, we will refer to it in a way that makes clear that it is associated with the statistic $m - \mu_0$. Therefore, the standard error of $m - \mu_0$ will be called $\sigma_{m-\mu_0}$, and its value is

$$\sigma_{m-\mu_0} = \sigma/\sqrt{n}. \tag{9.1}$$

A confidence interval around $m - \mu_0$ is computed as follows:

$$\text{CI} = (m - \mu_0) \pm z_{\alpha/2}(\sigma_{m-\mu_0}). \tag{9.2}$$

This is exactly like all confidence intervals we've computed before, except that now we're estimating $\mu_1 - \mu_0$ rather than μ_1.

Using the numbers from the phonics scenario, the 95% confidence interval around $m - \mu_0$ is computed as follows:

Step 1. Compute $m - \mu_0$.

$$m - \mu_0 = 74.5 - 72 = 2.5.$$

Step 2. Compute $\sigma_{m-\mu_0}$.

$$\sigma_{m-\mu_0} = \sigma/\sqrt{n} = 8/\sqrt{64} = 1.$$

Step 3. Find $z_{\alpha/2}$. Because this is the 95% confidence interval, $\alpha = .05$. Therefore, $\alpha/2 = .025$. Using the z-table, we find that 2.5% of the z-distribution falls below -1.96. Therefore, $z_{\alpha/2} = 1.96$.

Step 4. Compute $(m - \mu_0) \pm z_{\alpha/2}(\sigma_{m-\mu_0})$.

$$\text{CI} = (m - \mu_0) \pm z_{\alpha/2}(\sigma_{m-\mu_0}) = (74.5 - 72) \pm 1.96(1) = [0.54,\ 4.46].$$

We have 95% confidence that the true difference between μ_1 and μ_0 is in the interval [0.54, 4.46]. Our confidence comes from knowing that 95% of all intervals computed this way will capture $\mu_1 - \mu_0$. This means that our best estimate of the difference between the means of the whole language and phonics distributions is 2.5. However, a difference of 0.54 is just as likely as a difference of 4.46. The practical significance of this difference depends on all the factors discussed in Chapter 7.

To see that nothing mysterious has been done here, we can look at the situation slightly differently. The confidence interval around m was [72.54, 76.46] and the confidence interval around $m - \mu_0$ is [0.54, 4.46]. Therefore, the confidence interval around $m - \mu_0$ is simply equal to the confidence interval around m minus μ_0. That is, [72.54, 76.46] – 72 = [0.54, 4.46]; μ_0 has been subtracted from the upper and lower limits of the 95% confidence interval around m.

According to the null hypothesis, H_0: $\mu_1 - \mu_0 = 0$. Testing this null hypothesis is just a matter of asking whether 0 is in this interval $(m - \mu_0) \pm z_{\alpha/2}(\sigma_{m-\mu_0})$. Because it is not, a two-tailed test of the null hypothesis can be rejected at the $p < .05$ level of significance. However, this confidence interval also provides our best estimate of what $\mu_1 - \mu_0$ *is* (95% CI [0.54, 4.46]) in addition to what it is *not* (0). So, if the null hypothesis is never really true, a confidence interval around $m - \mu_0$ provides a simple method of estimating just how wrong it is.

LEARNING CHECK 3

1. A researcher with a long-standing interest in visual memory has collected data on the ability of university students to recall the details of five 30-second video clips. All participants were asked 100 questions about the videos; for example, How many people were in the video about the car theft? What color was the cottage in the video about the summer camp? What was the name on the store in the video about the hockey team? The average score on this test was $\mu_0 = 38$ with $\sigma = 10$. A random sample of 25 university students was shown the same five videos and asked the same 100 questions. Unlike all previous participants, these 25 participants were asked to answer the questions with their eyes closed (Nash, Nash, Morris, & Smith, 2015). The mean score for the sample of 25 students was $m = 42$. Use this information to compute the 95% confidence interval about an estimate of $\mu_1 - \mu_0$. Use this estimate to perform a two-tailed test of the null hypothesis in which H_0: $\mu_1 - \mu_0 \neq 0$.

Answers

1. The estimate of $\mu_1 - \mu_0$ is $m - \mu_0 = 42 - 38 = 4$. The standard error is $\sigma_{m-\mu_0} = \sigma/\sqrt{n} = 10/\sqrt{25} = 2$. Therefore, the 95% confidence interval is

$$\text{CI} = m \pm z_{\alpha/2}\left(\sigma_{m-\mu_0}\right) = 4 \pm 1.96(2) = [0.08,\ 7.92].$$

Because this confidence interval does not capture 0, we can reject the null hypothesis that H_0: $\mu_1 - \mu_0 = 0$ at the $p < .05$ level.

ESTIMATING $\delta = (\mu_1 - \mu_0)/\sigma$

Let's now return to the topic of power and effect size. In Chapter 8, the following points were made:

- Significance tests are used to decide whether two population means are different.
- If you have insufficient power to detect a difference, you may be wasting your time.
- Therefore, you should think about what effect size would be meaningful to you and then choose a sample size that provides sufficient power to reject the null hypothesis when it is false.

These points strongly imply that δ should be a primary concern. Therefore, rather than using δ as part of a power analysis conducted before running a significance test, it would seem more direct to estimate δ from the sample data. In fact, it is a simple matter to estimate δ, and a confidence interval around the estimate can be constructed if we know its standard error.

In Chapter 8, the population effect size was defined as follows:

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}.$$

When the population standard deviation is known, δ is estimated with the statistic d as follows:

$$d = \frac{m - \mu_0}{\sigma}. \tag{9.3}$$

In equation 9.3, μ_1 has been replaced by *m*. As with any statistic, *d* is subject to sampling error. When σ is known, the standard error of *d* is the following:

$$\sigma_d = \frac{1}{\sqrt{n}}. \tag{9.4}$$

(There is an explanation of why this is the standard error of *d* in a later section.) To compute the (1–α)100% confidence interval around *d*, we would use our familiar formula:

$$d \pm z_{\alpha/2}(\sigma_d). \tag{9.5}$$

To illustrate the construction of a confidence interval around *d*, we will return to the power-pose example. In Chapter 8 we were told that the mean of a population of final exam grades was $\mu_0 = 75$ with a standard deviation of $\sigma = 10$. Let's assume that a sample of $n = 64$ students held a power pose for 2 minutes before their final exam and that the average grade on the exam was $m = 76$. With this information we can compute an estimate of δ and a confidence interval around the estimate as follows:

Step 1. Compute *d*.

$$d = \frac{76-75}{10} = \frac{1}{10} = 0.1.$$

Step 2. Compute σ_d.

$$\sigma_d = 1/\sqrt{n} = 1/\sqrt{64} = 0.125.$$

Step 3. Find $z_{\alpha/2}$. Because this is the 95% confidence interval, α = .05. Therefore, α/2 = .025. Using the *z*-table, we find that 2.5% of the *z*-distribution falls below –1.96. Therefore, $z_{\alpha/2} = 1.96$.

Step 4. Compute $d \pm z_{\alpha/2}(\sigma_d)$.

$$\text{CI} = d \pm 1.96\,(\sigma_d) = 0.1 \pm 1.96(0.125) = [-0.145, 0.345].$$

We have 95% confidence that δ lies in the interval [–0.15, 0.35]. Our confidence comes from knowing that 95% of all intervals computed this way will capture δ. This means that our best estimate is that the mean of the power-pose distribution is 0.1 standard deviations above the mean of the regular population. According to Cohen's classification scheme, this is a very small effect size. The confidence interval shows that a difference of –0.15 standard deviations, which means a drop in the average final exam grade, is just as likely as a difference of 0.35.

One way to approach the practical significance of this result is to ask what effect this intervention would have if applied to all members of the population. This question can be addressed with Cohen's U_3. Our best estimate of δ is 0.1. The *z*-table shows that $U_3 = P(0.1) = 0.5793$, which means that we estimate the proportion of the H_0 (power-pose) distribution above $\mu_0 = 75$ (the mean of the H_0 distribution) to be 0.5398. Therefore, it is estimated that adding 2 minutes of power posing before the final exam produces about a 4% increase [(0.5398 – 0.5)∗100 = 3.98%] in the number of students scoring above the previous mean of 75. This seems like an interesting result given how minimal the intervention was.

However, we should also consider the limits of the confidence interval. U_3 for the lower limit (–0.15) is 0.4424. This means about a 6% *decrease* [(0.4424 – 0.5)∗100 = –5.76%] in the number of students scoring above the previous mean of 75. U_3 for the upper limit (0.35) is 0.6350. This means about a 13% *increase* [(0.6350 − 0.5)*100 = 13.5%] in the number of students scoring above the previous mean of 75.

All things considered, these results seem rather imprecise and not very compelling. Our best estimate is that a 2-minute intervention will increase by 4% the number of students scoring above the previous mean. The lower limit of the confidence interval suggests that power posing could reduce by 6% the number of students scoring above the previous mean. This seems like a pretty big risk. The upper limit suggests that power posing could increase by 13% the number of students scoring above the previous mean. Although this suggests a large potential benefit, it must be weighed against the risk of a large cost. If one believes these results to be interesting, a study could be designed to achieve a far more precise estimate of δ using the "precision planning" method described in Appendix 6.3 (available at study.sagepub.com/gurnsey).

From the hypothesis testing point of view, the null hypothesis is that power posing has no effect on exam grades. We can state this as

$$H_0\colon \delta = 0.$$

If the null hypothesis were true, we would expect our interval to capture 0. The interval ([−0.15, 0.35]) does capture 0. Therefore, we retain the null hypothesis and say that the result is not statistically significant.

*A Detail

Let's return to the question of why $\sigma_d = 1/\sqrt{n}$. The following expression needs no explanation at this point:

$$\mathrm{CI} = m \pm z_{\alpha/2}\left(\sigma_m\right) = m \pm z_{\alpha/2}\left(\sigma/\sqrt{n}\right).$$

We also know from equation 9.3 that

$$d = \frac{m - \mu_0}{\sigma}.$$

We say that we've *standardized* the difference between m and μ_0 by dividing the difference by σ. (Remember, d is a kind of z-score, and z-scores are called standard scores.) We can standardize the limits of the 95% confidence interval about m in the same way. That is,

$$\frac{\left[m \pm z_{\alpha/2}\left(\sigma/\sqrt{n}\right)\right] - \mu_0}{\sigma}.$$

With a little manipulation, we can derive the following:

$$\begin{aligned}\frac{\left[m \pm z_{\alpha/2}\left(\sigma/\sqrt{n}\right)\right] - \mu_0}{\sigma} &= \frac{\left(m - \mu_0\right) \pm z_{\alpha/2}\left(\sigma/\sqrt{n}\right)}{\sigma} \\ &= \frac{m - \mu_0}{\sigma} \pm \frac{z_{\alpha/2}\left(\sigma/\sqrt{n}\right)}{\sigma} = \frac{m - \mu_0}{\sigma} \pm z_{\alpha/2}\left(1/\sqrt{n}\right) = d \pm z_{\alpha/2}\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Or, more simply,

$$\sigma_d = \frac{\sigma_m}{\sigma} = \frac{\sigma/\sqrt{n}}{\sigma} = 1/\sqrt{n} = \frac{1}{\sqrt{n}}.$$

The Connection Between z_{obs} and d

As a final note, it is important to recognize the connection between d and z_{obs}. We define z_{obs} as follows:

$$z_{\text{obs}} = \frac{m-\mu}{\sigma/\sqrt{n}},$$

but this is equivalent to

$$z_{\text{obs}} = \frac{m-\mu}{\sigma} * \sqrt{n}.$$

(You should plug numbers into examples like this to convince yourself that the statement is true.) The first term in this expression [i.e., $(m - \mu_0)/\sigma$] is the definition of d given in equation 9.2. Therefore,

$$z_{\text{obs}} = d * \sqrt{n}, \tag{9.6}$$

and

$$d = z_{\text{obs}}/\sqrt{n}. \tag{9.7}$$

Because $z_{\text{obs}} = d * \sqrt{n}$ we can see that no matter how small d is, as long as it is not exactly 0 it will become statistically significant if n is large enough. The value of equation 9.7 is that it allows one to determine an estimated effect size from a published report, even when a researcher has not reported it. The importance of this will be seen in the third part of this book when meta-analysis is discussed.

LEARNING CHECK 4

1. Let's say that the mean IQ for American adults living west of the Mississippi River is 100, with a standard deviation of 15. A random sample of 105,625 American adults living east of the Mississippi River is found to have a mean of 100.1.

   (a) Compute the 95% confidence interval around an estimate of δ.

   (b) Would this confidence interval lead you to reject a two-tailed test of the null hypothesis at the $p < .05$ level of significance? Why or why not?

   (c) Show how to convert the estimate of δ into z_{obs}.

Answers

1. (a) The estimate of δ is $d = (100.1 - 100)/15 = .006667$. The standard error of d is $1/\sqrt{n} = 1/\sqrt{105{,}625} = 0.0031$. Therefore, the 95% confidence interval is

   $$\text{CI} = d \pm Z_{\alpha/2}\,(\sigma_d) = .006667 \pm 1.96(0.0031) = [0.0006,\ 0.0127].$$

   (b) Because this confidence interval does not capture 0, we can reject a two-tailed test of the null hypothesis that H_0: $\delta = 0$ at the $p < .05$ level.

   (c) $z_{\text{obs}} = d * \sqrt{n} = (.1/15)\,(325) = 2.167.$

ESTIMATION VERSUS SIGNIFICANCE TESTING

Let's start with a few kind words for significance tests. Significance tests are used to make *decisions* about the null hypothesis. If we adopt the $p < .05$ criterion for statistical significance, then we seem to have a simple rule that makes research decisions very easy. A universal criterion for statistical significance could be seen as a referee in scientific debates. Without such a referee, judgments about a particular result may be determined by who can shout the loudest. A senior researcher might claim that a difference reported by a junior researcher is not important, while claiming that the same difference is important when she reports it. The apparent impartiality of $p < .05$ explains its role in the publication process.

Unfortunately, as discussed earlier in this chapter, many problems arise from adopting $p < .05$ as a universal criterion for statistical significance, and these far outweigh the potential value of $p < .05$ as an impartial referee. One of the major problems is the binary thinking that leads one to believe that when a statistic has an associated $p < .05$, the result is important and meaningful, and that when a statistic has an associated $p > .05$, the result is unimportant and meaningless. Binary thinking is put into stark relief when we consider the following question: What if H_1: $\mu_1 \neq \mu_0$, and $z_{obs} = 1.96$?

What if H_1: $\mu_1 \neq \mu_0$, and $z_{obs} = 1.96$?

If we adopt the $p < .05$ criterion for statistical significance, then $z_{critical} = \pm 1.96$ for a nondirectional test. So, what do we do if z_{obs} is exactly equal to 1.96, or exactly equal to −1.96? That is, should we treat 1.96 like 1.95 (not statistically significant) or 1.97 (statistically significant)? Of course, these two z-scores (1.95 and 1.97) are almost identical; if we hadn't heard about the $p < .05$ criterion for statistical significance, we wouldn't have thought for a moment that we should treat them differently. Unfortunately, psychologists of my generation were instructed to do exactly this. The 1974 edition of the *APA Publication Manual* provided the following guidance about the interpretation of p-values:

> Caution: Do not infer trends [*read as statistical significance*] from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. (p. 19)

This is an explicit instruction to treat $z_{obs} = 1.97$ and $z_{obs} = 1.95$ differently. In fairness, this (mis)advice may derive from Fisher (1926), who said:

> Personally, the writer prefers to set a . . . standard of significance at the 5 per cent point, *and ignore entirely all results which fail to reach this level*. (p. 504, emphasis added)

Imagine an experiment that tested the effects of regular exercise on the IQs of men and women. In the general population, the IQs of both men and women have $\mu = 100$, and $\sigma = 15$. Let's say a random sample of 100 men and a random sample of 100 women were put on the same exercise routine for 3 months. At the end of 3 months, the mean IQ for men was found to be $m_{men} = 102.955$, and the mean IQ for women was found to be $m_{women} = 102.925$. If you do the calculations, you'll find that $z_{obs} = 1.97$ for men, and $z_{obs} = 1.95$ for women. The guidance from the 1974 *APA Publication Manual* says we should report this result as follows:

> The effect of exercise on IQ was significant for men ($z = 1.97$, $p < .05$) but not for women ($z = 1.95$, $p > .05$).

This kind of binary thinking has been passed from one generation of psychologists to the next, so it is no wonder that some engage in p-hacking to try to get statistically significant results.

Fortunately, things have changed, and more recent editions of the APA manual provide better guidance. The sixth edition (published in 2010) stated the following:

> Because confidence intervals combine information on location [*read as point-estimates*] and precision and can often be directly used to infer significance levels [*read as can be used to test* H_0], they are, in general, the best reporting strategies. The use of confidence intervals is therefore strongly recommended. (p. 34)

According to this guidance, we could report the results of our experiment in the following way:

APA Reporting

Following the 3-month exercise regime, the mean IQ for men was $M = 102.96$ (95% CI [100.02, 105.90]) and the mean IQ for women was $M = 102.93$ (95% CI [99.99, 105.87]). The mean IQ in the general populations of men and women is 100. Therefore, both after-exercising sample means are associated with increases in IQ. The effect of exercise on IQ yielded an effect size of $d = .197$ for men (95% CI [.001, .393]) and $d = .195$ for women (95% CI [–.001, .391]). Therefore, exercise leads to very similar improvements for both men and women.

This report gives a much clearer sense of the results of the experiment. In this situation, it doesn't seem to be particularly important that one confidence interval contains $\mu_0 = 100$, and the other doesn't.

What if H_1: $\mu_1 < \mu_0$, and $z_{obs} = 2$?

Another problem we've already encountered within the Fisher-Neyman hybrid model of significance testing is the question of what to do with a large z_{obs} that is inconsistent with our directional alternative hypothesis. For example, let's say that a review of the literature suggested to a researcher that extensive video gaming would impair the ability of college students to solve crossword puzzles. (The speculation is that video gaming would make a person more impulsive and this would lead to fewer correct words on the crossword puzzle.) The researcher happens to know that college students in general complete $\mu_0 = 24$ ($\sigma = 4.8$) words in 20 minutes on a standardized crossword puzzle. The researcher selects a random sample of $n = 36$ college students and has them play a video game for 30 minutes, after which they are given 20 minutes to complete the standardized crossword puzzle. The variable of interest is the number of correct responses on the crossword puzzle.

In this situation, the null hypothesis is that H_0: $\mu_1 - \mu_0 = 0$ and a directional alternative hypothesis that H_1: $\mu_1 - \mu_0 < 0$ has been chosen to increase the power of the experiment. (As noted in Chapter 8, one-tailed tests are more powerful than two-tailed tests, if your prediction is correct.) The alternative hypothesis states that following video game playing, the participants are expected to get fewer words correct on the crossword puzzle. Because this is a directional test that predicts $m - \mu_0 < 0$, $z_{critical} = -1.645$, with $\alpha = .05$; i.e., the $p < .05$ criterion for statistical significance. After running the experiment, the researcher finds that the average number of words completed is $m = 25.6$, which corresponds to $z_{obs} = 2$. Now she faces a conundrum; z_{obs} is large, but she is only allowed to reject H_0 if $z_{obs} < -1.645$. Should she say there is no statistically significant effect of video game playing on crossword performance? This would be an odd conclusion because $z_{\text{obs}} = 2$ would be unusual if H_0 were true.

This conundrum occurs only because the Fisher-Neyman hybrid model forces us to make a *decision* based on the statistic. The simplest way to deal with these data is to avoid the decision-making language of significance tests and say something like the following:

APA Reporting

Following 30 minutes of video gaming, the mean number of correctly completed crossword items was $M = 25.6$, 95% CI [24.03, 27.17]. This is an improvement over the population mean of $\mu_0 = 24$ and corresponds to an effect size of $d = .33$, 95% CI [.01, .66]. The magnitude of effect size is in line with past research, but its sign is opposite of what was expected. Therefore, further analysis is required to determine what features of this experiment may have differed from those of past experiments.

Decisions Versus Measurements

In this chapter and in Chapters 6 to 8, we have considered two approaches to inferential statistics: estimation and significance testing. Even though these approaches rest on exactly the same foundations, they differ fundamentally in emphasis. Estimation emphasizes what a parameter *is*. The hybrid model of significance testing emphasizes what a parameter *is not*. The estimation approach is the more general of the two, because once we have an estimate of what the parameter is (e.g., $m_{\text{men}} = 102.96$, 95% CI [100.02, 105.90]), we can, if we wish, make a statistical judgment about what it is not (e.g., we can reject the null hypothesis that $\mu_{\text{men}} = 100$).

The questions raised in the two preceding sections ("What if H_1: $\mu_1 - \mu_0 \neq 0$, and $z_{\text{obs}} = 1.96$?" and "What if H_1: $\mu_1 - \mu_0 < 0$, and $z_{\text{obs}} = 2$?") arise from a focus on *decision making*. When we establish a criterion for statistical significance (α, z_{critical}, or m_{critical}), it leads us to binary thinking: either H_0 is true or it is false; exercise either affects IQ or it does not; either the earth is round or it is not. That is, the answers are either yes or no, black or white. From the estimation point of view, the questions are more along the lines of "How wrong is H_0?," "How big is the effect of exercise on IQ?," and "How round is the earth?" Estimation does not lure us into binary thinking.

Addressing any research question should be thought of as part of an ongoing effort to understand a phenomenon of interest to researchers, or to society in general. Therefore, no single study should be seen as providing a verdict on such questions. Ideally, evidence accumulates as more studies are run, and researchers eventually form a consensus about the size of some effect. The hybrid model of significance testing, with its focus on decision making, can obscure the fact that no single study is definitive. When making a decision about H_0, one can be led to feel that the decision is final, particularly if we misunderstand p-values. With estimation, however, our view is more like that of a pollster, who finds that on August 10, 2016, 48% of decided voters prefer Hillary Clinton and 40% prefer Donald Trump. This result is seen as a single, more or less imprecise estimate. A better estimate would come from combining many such imprecise estimates. In statistics, we refer to the combination of many imprecise estimates as *meta-analysis*. We will discuss meta-analysis in Part III of this book. There we will see that it is easier to combine parameter estimates in a meta-analysis than the results of significance tests. For a meta-analysis to be valid, however, it is important to combine all relevant results, not just those that have made it into the literature through the $p < .05$ filter.

SUMMARY

Throughout their existence, significance tests have been criticized by many distinguished psychologists. The criticisms include the following points:

- The requirement for statistical significance for publication leads to publication bias and the file-drawer problem, causing the published literature to misrepresent the true effect of an intervention or the true difference between two populations.
- The requirement for statistical significance for publication also makes it more likely that Type I errors will be published.
- The requirement for statistical significance encourages *p*-hacking, which also distorts the literature and increases the probability that Type I errors will be published.
- A focus on obtaining small *p*-values is like the tail wagging the dog and distracts from the important questions of practical significance.
- The quest for small *p*-values promotes a tendency toward binary thinking, in which only statistically significant results are viewed as important and meaningful.
- Many researchers misunderstand *p*-values and thus misinterpret their results.

Many view confidence intervals as the healthiest alternative to significance tests because the emphasis is taken off decision making and placed on estimation. In a world in which estimation is standard practice, estimates would enter the literature without having to pass through the $p < .05$ filter of statistical significance. Furthermore, confidence intervals can be used to conduct significance tests. If μ_0 does not fall in the interval

$$m \pm z_{\alpha/2}(\sigma_m),$$

then we can reject H_0 at the $p < \alpha$ level of significance.

We can also estimate the difference between two population means ($\mu_1 - \mu_0$) using the statistic $m - \mu_0$. This statistic has a standard error of $\sigma_{m-\mu_0} = \sigma/\sqrt{n}$. A confidence interval around $m - \mu_0$ is calculated as

$$\left(m - \mu_0\right) \pm z_{\alpha/2}\left(\sigma_{m-\mu_0}\right).$$

We can not only estimate $\mu_1 - \mu_0$ but also test the null hypothesis that $\mu_1 - \mu_0 = 0$ by asking whether 0 falls in the confidence interval.

When we think of the hybrid model of significance testing that involves power analysis, we become aware of the central role that δ should play in our thinking about research questions. δ is estimated by

$$d = \left(m - \mu_0\right) / \sigma.$$

Of course, *d* is subject to sampling error, so a confidence interval can be computed around *d*. The standard error of *d* is $\sigma_d = 1/\sqrt{n}$, so the $(1-\alpha)100\%$ confidence interval around *d* can be computed as

$$d \pm z_{\alpha/2}\,(\sigma_d).$$

Many researchers feel that estimating, μ, $\mu_1 - \mu_0$, or δ is healthier and more informative than testing a null hypothesis about μ_0. Because confidence intervals for μ, $\mu_1 - \mu_0$, or δ can be used to test hypotheses about μ_0, $\mu_1 - \mu_0$, or δ, they represent a more general approach to data analysis. Estimation avoids the binary (black-and-white) thinking that we may fall into with significance tests and lends itself more readily to meta-analysis, which we'll cover in Part III.

KEY TERMS

file-drawer problem	205	*p*-hacking	207	publication bias	205

EXERCISES

Definitions and Concepts

1. What is the definition of a *p*-value?
2. Why does the file-drawer problem happen?
3. Why should one not use the run-until-statistically-significant procedure?
4. How will the literature be biased if only statistically significant results ($p < .05$) are published?
5. Why is it important for replications to be published?
6. Give three examples of *p*-hacking.
7. Explain why failing to reject the null hypothesis does not mean that the null hypothesis is true.

True or False

State whether the following statements are true or false.

8. If I retain H_0, then it is true.

9. If I reject H_0, then H_1 is true.

10. I ran an experiment in which H_1: $\mu_1 - \mu_0 < 0$. I obtained $z_{obs} = -1.43$. Therefore, my study is not worth submitting for publication.

11. $\mu_0 = 22$ and my 95% confidence interval around m is [21, 28]. Therefore, I should retain a two-tailed test of H_0, assuming $\alpha = .05$.

12. $\mu_0 = 5$ and my 95% confidence interval around $m - \mu_0$ is [–1, 6]. Therefore, I should retain a two-tailed test of H_0, assuming $\alpha = .05$.

13. $\mu_0 = 22$ and my 95% confidence interval around m is [18, 21]. Therefore, I should retain a two-tailed test of H_0, assuming $\alpha = .05$.

14. $\mu_0 = 5$ and my 95% confidence interval around $m - \mu_0$ is [1, 6]. Therefore, I should reject a two-tailed test of H_0, assuming $\alpha = .05$.

15. If $\mu_0 = 5$ and my 95% confidence interval around $m - \mu_0$ is [1, 7], then $m = 4$.

16. If $\mu_0 = 5$ and my 95% confidence interval around $m - \mu_0$ is [1, 7], then $m = 9$.

Calculations

17. If $\mu_0 = 32$, $\sigma_0 = 10$, $m = 27$, and $n = 25$, answer the following:
    (a) Calculate the 95% confidence interval around m.
    (b) Calculate the 95% confidence interval around $m - \mu_0$.
    (c) Calculate the 95% confidence interval around d.
    (d) Assuming the $p < .05$ criterion for statistical significance, explain why these intervals would or would not allow you to reject the null hypothesis when H_1: $\mu_1 - \mu_0 \neq 0$.

18. If $\mu_0 = 16.8$, $\sigma_0 = 2.4$, $m = 18.3$, and $n = 36$, answer the following:
    (a) Calculate the 95% confidence interval around m.
    (b) Calculate the 95% confidence interval around $m - \mu_0$.
    (c) Calculate the 95% confidence interval around d.
    (d) Assuming the $p < .05$ criterion for statistical significance, explain why these intervals would or would not allow you to reject the null hypothesis when H_1: $\mu_1 - \mu_0 \neq 0$.

19. If $\mu_0 = 100$, $\sigma_0 = 15$, $m = 101$, and $n = 225$, answer the following:
    (a) Calculate the 95% confidence interval around m.
    (b) Calculate the 95% confidence interval around $m - \mu_0$.
    (c) Calculate the 95% confidence interval around d.
    (d) Assuming the $p < .05$ criterion for statistical significance, explain why these intervals would or would not allow you to reject the null hypothesis when H_1: $\mu_1 - \mu_0 \neq 0$.

20. If $\mu_0 = 100$, $\sigma_0 = 15$, $m = 101$, and $n = 225$, answer the following:
    (a) Calculate the 90% confidence interval around m.
    (b) Calculate the 90% confidence interval around $m - \mu_0$.
    (c) Calculate the 90% confidence interval around d.
    (d) Assuming the $p < .05$ criterion for statistical significance, explain why these intervals would or would not allow you to reject the null hypothesis when H_1: $\mu_1 - \mu_0 > 0$.

21. If $\mu_0 = 100$, $\sigma_0 = 15$, $m = 102$, and $n = 225$, answer the following:
    (a) Calculate the 90% confidence interval around m.
    (b) Calculate the 90% confidence interval around $m - \mu_0$.
    (c) Calculate the 90% confidence interval around d.
    (d) Assuming the $p < .05$ criterion for statistical significance, explain why these intervals would or would not allow you to reject the null hypothesis when H_1: $\mu_1 - \mu_0 > 0$.
    (e) Assuming the $p < .05$ criterion for statistical significance, explain why these intervals would or would not allow you to reject the null hypothesis when H_1: $\mu_1 - \mu_0 < 0$.

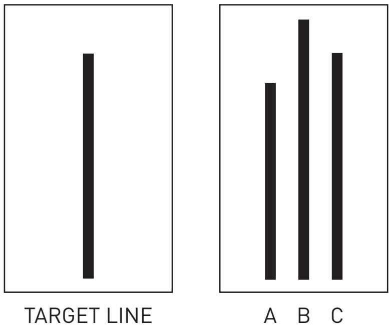
22. If $\mu_0 = 75$, $\sigma_0 = 15$, $m = 80$, and $n = 25$, answer the following:
    (a) Calculate the 90% confidence interval around m.
    (b) Calculate the 90% confidence interval around $m - \mu_0$.
    (c) Calculate the 90% confidence interval around d.
    (d) Assuming the $p < .05$ criterion for statistical significance, explain why these intervals would or would not allow you to reject the null hypothesis when H_1: $\mu_1 - \mu_0 > 0$.
    (e) Assuming the $p < .05$ criterion for statistical significance, explain why these intervals would or would not allow you to reject the null hypothesis when H_1: $\mu_1 - \mu_0 < 0$.

Scenarios

23. A difference threshold is the smallest difference between two levels of a stimulus that can be discriminated with a given level of accuracy. Let's say that on average, university students require a 3% increase in weight to discriminate one weight from another. For example, if a standard stimulus weighs 100 grams, then a comparison stimulus would need to be 3 grams heavier (i.e., 103 grams) for the difference to be noticeable. A researcher wondered how much difference thresholds could be changed by hypnotic suggestion. Each member of a random sample of 100 university students was given a suggestion under hypnosis that his or her sensory sensitivity had increased. Following hypnosis, difference thresholds for weights were measured in all participants. The standard stimulus was 100 grams and the mean difference threshold after hypnosis was $m = 2.5$ grams, with an estimated standard deviation of $s = 2$. Compute an approximate 95% confidence interval around $m - \mu_0$ and use this interval to determine whether a two-tailed test of H_0 is statistically significant at the $p < .05$ level.

24. Cell phone usage is a fact of life for many people these days, and there is some evidence of separation anxiety when individuals cannot access their phones. Let's say that the mean systolic blood pressure for cell phone users in possession of their phones is 110 mmHg with a standard deviation of 18. We wonder if the anxiety associated with cell phone separation will lead to increased blood pressure. A random selection of 36 cell phone users was contacted and asked to complete a survey about cell phone use. They were asked to leave their phones in a locked cabinet in another room while they completed the survey. At the end of the survey, their blood pressure was taken and the mean was found to be 128 mmHg. Compute the 95% confidence interval around $m - \mu_0$ and state whether a two-tailed test of H_0 is statistically significant at the $p < .05$ level.

25. People often feel pressure to conform to the attitudes and behaviors of groups. There are experiments showing that people will even deny the evidence of their own senses to conform to group behaviors (Asch, 1951). Imagine that 16 students in a psychology department participant pool volunteered for a study about perceptual judgments. They were asked to show up to a large classroom to take the test. On the day of the test, there were about 64 students in the classroom altogether. All 64 people were shown a large number of stimuli and asked which of three alternatives (A, B, C) matched a target line (e.g., the line on the left). Twenty-one different stimuli were shown. For each stimulus, the experimenter asked for a show of hands in response to the questions "How many think A is the match?" "How many think B is the match?" and "How many think C is the match?"



Unbeknownst to the 16 volunteers, the majority of people in the classroom were confederates of the researcher. Whenever the obvious match was C (as in this example), all of the confederates raised their hands in response to a non-match (e.g., B). The experimenter (and colleagues) noted how many times each of the 16 volunteers raised their hands along with the majority, and thus gave an obviously wrong answer. They found that on average the volunteers gave the wrong answer on m = 5.1 times out of 7 opportunities, with a standard deviation s = 1.2. Compute the approximate 95% confidence interval around m. If you were to compute an approximate confidence interval around $m - \mu_0$, what would μ_0 be? What do you think this result says about conformity? Would more or less than 95% of all such intervals be expected to capture $\mu_1 - \mu_0$?

26. It is known that the way a question is posed can affect the answer given (Loftus & Palmer, 1974). This is particularly important in court trials. Imagine that a large population of American adults had seen a movie of a traffic accident in which a car ran a stop sign and struck another. After viewing the movie, they were asked to judge how fast the car (that ran the stop sign) was going when it contacted the other car. On average, they judged that it was going μ = 32 mph, with a standard deviation σ = 6. In a subsequent study, a random sample of 25 participants from the same population were shown the movie and asked how fast the car was going when it smashed into the other car. On average, they judged that it was going 42 mph (51.5 kph). Compute the 95% confidence interval around d. What value of z_{obs} does d correspond to? Do you think that it would be easier for a non-statistician to understand a confidence interval around d or a confidence interval around $m - \mu_0$?

27. Does memory for line drawings change with age (Harwood & Naylor, 1969)? A standardized test shows that when young adults (mean age 24 years) have learned to recognize 20 line drawings of common objects, they recognize about 75% (σ = 5) of these in a surprise test 4 weeks later. The same experiment examined the recognition performance of 36 older adults (mean age 71.2 years). The mean percentage of recalled items in the older participants was found to be lower than the mean percentage recalled in the younger participants. A one-tailed test (μ_0 = 75, σ = 5) showed that this difference was statistically significant, $z_{obs} = -2.4, p = .0082$. Compute the 95% confidence interval around d. Use the information given, as well as the d-statistic you just computed, to determine the mean recognition rate of the older adults. (*Hint:* Remember that d represents change in standardized units.) Do you judge the decrease in recall ability to be severe?

28. Early work by Hetherington and Ranson (1940) suggested a role for the ventromedial hypothalamus in the regulation of food intake. It is known that Wistar rats weigh 275 grams on average, with a standard deviation of 12. A random sample of four Wistar rats was selected and subjected to a surgery that lesioned (destroyed) their ventromedial hypothalamus. After a 2-month period, during which these brain-damaged rats had free access to unlimited food, each was weighed. The mean weight of these brain-damaged rats was greater than that of the average Wistar rat, $z_{obs} = 91.7, p < .05$. Compute the 95% confidence interval around d. What was the mean weight of the four brain-damaged rats? Do you think these results are of any practical significance?

29. Here is a strange passage of text taken from an interesting study published in the early 1970s (Dooling & Lachman, 1971).

> With hocked gems financing him, our hero bravely defied all scornful laughter that tried to prevent his scheme. "Your eyes deceive," he had said. "An egg, not a table, correctly typifies this unexplored planet." Now three sturdy sisters sought proof. Forging along, sometimes through calm vastness, yet more often over turbulent peaks and valleys, days became weeks as many doubters spread fearful rumors about the edge. At last from nowhere welcome winged creatures appeared, signifying momentous success.

Imagine that a large population of psychology students had been read this passage and were asked to recall as many words as they could. The mean number of correctly recalled words was 13.23, with a standard deviation of σ = 4.2. A random sample of 16 psychology students was read the same passage but, unlike for the larger group, the passage was preceded by the title "Christopher Columbus." The mean number of words recalled by the 16 participants was m = 15.67. Compute the 95% confidence interval around $m - \mu_0$. What do you make of these results? Can you think of any situations in which these results would be seen as having practical significance?