CHAPTER 4

Hypothesis Testing, Power, and Control

A Review of the Basics

OBJECTIVES

After reading this chapter, students should be able to do the following for a research project they have conducted or proposed:

- Generate conceptual, research, and statistical hypotheses
- List and describe the steps we follow to test a null hypothesis
- Consider the implications of inferential error and statistical significance
- Describe the four outcomes of an inferential decision and provide the probability associated with each
- Discuss power and consider ways of increasing the power of their test of the null
- Discuss the importance of effect size and estimate effect size for a given test of the null
- Discuss the role of replication in research as it relates to power
- Discuss the influence of confounding and extraneous variables on inferential statements
- List and describe common ways of controlling extraneous variables

Are outgoing people happier than people who keep to themselves? Do people do better when they think they are being evaluated? Does stress management training help highly anxious people? We use **inferential** techniques to answer these kinds of questions. **Hypothesis testing** is the most common inferential technique used by psychologists.
Hypotheses are educated guesses about the relationships between variables. By educated guesses, we mean that these hypotheses come from previous research, theory, or logic. We categorize hypotheses by the level of the variables specified in the hypothesis.

**THREE LEVELS OF HYPOTHESES**

As researchers, we first hypothesize about concepts or theoretical constructs. Here are some examples of conceptual hypotheses:

- “People work harder when they think they are being evaluated.”
- “Outgoing people are happier than people who keep to themselves.”
- “Highly anxious people trained in stress management are less anxious than those not trained.”

These are conceptual hypotheses because they state expected relationships among concepts such as “work harder, outgoing, happy, and highly anxious.” As researchers, we talk about our research using these conceptual terms. To test a conceptual hypothesis, we need to operationalize these concepts, thereby creating the next level of hypothesis—the research hypothesis. Our research hypotheses for the three conceptual hypotheses above might be as follows:

- “People who are told that experts are rating their performance get higher scores on a test than people who are not told they are being rated.”
- “People who score high on a standard test of extraversion give higher ratings of happiness in their lives than do people who score low.”
- “People who score high on a standard test of anxiety have lower resting heart rates after stress management training compared to similar people who have received no training.”

These are research hypotheses because we have operationalized our concepts such that they are now measurable. We conduct research to test our research hypothesis. But when we do our statistics, we are not testing our research hypothesis directly; rather, we are testing a statistical hypothesis. A statistical hypothesis, in parametric hypothesis testing, states the expected relationship between or among summary values of populations, called parameters. There are two kinds of statistical hypotheses: the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$). And it is the null hypothesis that we actually test when we use inferential procedures.

**NOTE:** Statistical hypotheses for nonparametric tests (i.e., tests that are not looking for differences among summary values or parameters of populations) are somewhat different.

**Conceptual Exercise 4A**

A research methods student hypothesizes that people working in helping professions are more ethical than people working in business professions. Can you help her restate this hypothesis as a research hypothesis?
Once we have restated our research hypothesis in terms of a null and alternative hypothesis, we can then test the null.

**TESTING THE NULL HYPOTHESIS**

In science, we recognize that there is much more power in disconfirming a hypothesis than there is in confirming one. For example, let’s say you want to show that your spouse is faithful. To demonstrate or confirm this hypothesis, you present the fact that your spouse has never had an affair in your 10 years of marriage. The evidence does seem to confirm that your spouse is faithful, but maybe you should track his or her behavior for another 10 years just to be sure. By using confirmation evidence, it is impossible to know how much evidence is enough—10 years? 20 years? But what if we looked for disconfirming evidence? How much disconfirming evidence would you need to prove that your spouse is NOT faithful? A single affair is all it takes! Just one disconfirming piece of evidence is needed to disprove your hypothesis that your spouse is faithful. Although the null hypothesis is rarely specified in a research article, it is the hypothesis being statistically tested when you use inferential statistics.

You will recall from your statistics course that the null has that name because the researcher hopes to show that the null is not likely to be true (i.e., he or she hopes to *nullify it*). If the researcher can show that the null is not supported by the data, then he or she is able to accept an alternative hypothesis, which is the hypothesis the researcher postulated at the outset of the study. We often tell our statistics students that we cannot prove that our research hypothesis is true; we can only provide evidence that the null hypothesis is probably not!

Let’s say our research hypothesis is that people get higher scores on some test when they are told that their performance is being evaluated. Let’s follow the steps given below to test this hypothesis:

1. **State the null and the alternative.** Our null is that the mean score of people who are told they are being evaluated equals the mean score of people who are not told they are being evaluated. The null is a statement of no difference between groups—no treatment effect. This is the hypothesis you hope to nullify. Our alternative hypothesis is that the mean score of people who are told they are being evaluated is higher than the mean score of those who are not told they are being evaluated. This is the hypothesis you hope to confirm. Statistical hypotheses typically use Greek letters to refer to population parameters (i.e., summary values about populations). The Greek letter we use for population mean is $\mu$ (mu). In statistical notation, our null and alternative are as follows:

   $H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 > \mu_2$

2. **Collect the data and conduct the appropriate statistical analysis.**
3. **Reject the null and accept the alternative or fail to reject the null.**
4. **State your inferential conclusion.**
An inference is a statement of probability. An inference is not a statement made with certainty. It is a leap from a specific instance to a general rule. When we make an inference, we may be wrong, but inferential statistics provide us with an estimate of the probability that our inference is correct. In essence, when we reject a null hypothesis, we are saying that it is unlikely that the null is true. It is more likely that an alternative hypothesis is true. But remember, inferences are probability statements about the nature of populations that are based on information about samples. And when we make such statements, we can be wrong. After all, it could be that our sample was just unusual or was not representative of the population.

When we say that it is unlikely that the null is true, what do we mean by unlikely? This brings us to the heart of the matter—statistical significance!

**STATISTICAL SIGNIFICANCE**

“People perform significantly better when they think they are being evaluated than when they think they are not.”

“Outgoingness is significantly related to happiness.”

“Highly anxious people are significantly less anxious after receiving stress management training.”

You have no doubt seen many statements such as these during your reading of the research literature. But what do they really mean? Well, significant does not mean important. Significance is a statistical term, and statements such as these are statements of probability.

Here is what the examples given above really mean.

“It is highly unlikely that people perform the same when they think they are being evaluated than when they think they are not.”

“It is highly unlikely that outgoingness and happiness are unrelated.”

“It is highly unlikely that highly anxious people who receive stress management training are equally as anxious as those who received no such training.”

When we say that there is a significant difference between groups, we are saying that the probability that the groups are the same is low, very low. What do we mean by low? Well, researchers have agreed that low is less than 5% or 1%. These two levels are called significance levels, and we use the symbol \( \alpha \) (alpha) to refer to them. Alpha, then, is the level of significance chosen by the researcher to evaluate the null hypothesis. If the alpha level chosen was 5%, for example, then we are stating that the probability that the groups are the same is less than 5%. Therefore, we conclude they are not the same. Seems reasonable, don’t you think? Of course, as we have said, any inference is a statement of probability. When we say two groups are significantly different, we are pretty sure they are, but we are not certain. We could be wrong. With any inference, there is some probability that we could be wrong. And if we are wrong, we are making an inferential error.
INFERENTIAL ERRORS: TYPE I AND TYPE II

Imagine we have a coin and we are tossing it. We have tossed it 15 times, and it has turned up heads 15 times. What would you infer from this?

Well, there are two possibilities. One, that the coin is fair, and two, that the coin is not fair. We suspect that you, like us, would infer that this coin is bogus. Even if we have not calculated the actual probability of a fair coin showing heads 15 times in a row, we all know intuitively that this is not very likely to happen. In fact, the probability is close to zero ($p = 1/2^{15} \approx 0.00003$). We think it is a lot more reasonable to conclude that we have a bogus coin than to conclude we have a fair coin and that it turned up 15 heads in a row by fluke. Of course, it’s possible that’s what happened—a really rare fluke. And our conclusion that the coin is bogus would be wrong. We would have made the dreaded Type I error. So let’s examine this coin-tossing situation from a hypothesis testing point of view. We will begin with our null and alternative hypotheses in everyday words.

Null hypothesis in words: The coin is fair.
Alternative hypothesis in words: The coin is bogus.

Remember our goal in hypothesis testing is to reject the null and accept our alternative. You will recall that power in hypothesis testing is our ability to do just that—to reject false nulls. We have already collected our data. We got 15 heads in a row.

There are four possible inferential outcomes in our little study. Here they are:

1. The coin is fair, but we conclude it is bogus—This is a Type I error.
2. The coin is fair, and we conclude it is fair—This is a correct decision.
3. The coin is bogus, but we conclude it is fair—This is a Type II error.
4. The coin is bogus, and we conclude it is bogus—This is a correct decision. This is power.

The last outcome is our goal in hypothesis testing. We want to reject a false null. If our coin is indeed bogus, we want to conclude that it is. Of course, when we make a statistical inference, we never know if we are correct. We could be wrong. That is the nature of the beast. But we can do things to better our chances of being right.

In hypothesis testing language, the probability of making a Type I error is equal to alpha ($\alpha$), the level of significance chosen by the researcher (typically, $\alpha$ is set at .05 or .01). Displaying an amazing lack of imagination, statisticians call the probability of a Type II error beta ($\beta$), and because a Type II error can only occur if the null is false, beta is not set by the researcher. Let’s put our four outcomes in a table (see Table 4.1).

Researchers do not want to make errors in their inferences, but which error is more important? Well, that depends.

Several years ago, one of us was hired as a statistical consultant to do some research on the state of the terrain at a local ski area. Skiing Louise in Banff, Alberta, had recently put in some snowmaking equipment on the mountain. Parks Canada was concerned that this
equipment might have damaged the mountain terrain. If we wanted to determine the truth of the matter, we would have to assess every inch of the area (the population), but this was impossible. Instead, we sampled the terrain by assessing the density and diversity of the plant growth of randomly determined plots. We knew that there were four possible outcomes of our statistical analysis of the data. They are in Table 4.2.

Skiing Louise was more concerned with Type I error. It did not want us to infer that the mountain was damaged when it was not because it would have to do extensive remedial planting. It wanted us to reduce the probability of a Type I error. Parks Canada, on the other hand, had a different agenda. It was concerned about Type II error. It wanted to make sure that our inference that the mountain was okay was not a mistake. It wanted us to reduce the probability of a Type II error. As you can see, which error is more important depends on the agenda of the researcher. Scientists are objective and so will use procedures to keep both kinds of errors as low as possible.

Are you ready for one of our favorite examples? Romeo made a Type I error when he wrongly rejected the null hypothesis “Juliet is alive.” So sad. Julius Caesar made a Type II error when he failed to reject the null hypothesis “Brutus is my buddy.” Also, a tragic error (Evans, 2007).

The goal of hypothesis testing is to reject the null when we should, of course, and accept an alternative. After all, the alternative is the research hypothesis that we want to confirm. Techniques and statistical analyses that increase the probability of rejecting false nulls are said to be powerful techniques. Remember that rejecting false nulls is the goal of hypothesis testing. Let’s look more closely at power.

### Table 4.2 Outcomes of Our Analysis of the Mountain Terrain

<table>
<thead>
<tr>
<th>Our Decision</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>The Mountain Is Okay</em></td>
</tr>
<tr>
<td><em>The mountain is damaged</em></td>
<td>Type I error</td>
</tr>
<tr>
<td><em>The mountain is okay</em></td>
<td>Correct inference</td>
</tr>
<tr>
<td></td>
<td><em>The Mountain Is Damaged</em></td>
</tr>
<tr>
<td><em>The mountain is damaged</em></td>
<td>Correct inference</td>
</tr>
<tr>
<td><em>The mountain is okay</em></td>
<td>Type II error</td>
</tr>
</tbody>
</table>
For many social scientists, research is all about power. So what exactly do we mean by power?

A powerful test of the null is more likely to lead us to reject false nulls than a less powerful test. As researchers, we want to reject nulls that should be rejected. When we say “powerful test of the null,” we are not just referring to the power of the specific statistical test. A test of the null is more powerful when good research techniques are used, good designs are chosen, samples are drawn properly, and so on. However, in this chapter, we are focusing on the role of the statistical procedures in power.

Powerful tests are more sensitive than less powerful tests to differences between the actual outcome (what you found) and the expected outcome (not what you really expected but what you stated in the null hypothesis). Remember that \( \beta \), beta, is the symbol we use to stand for the probability of not rejecting a false null, a bad thing—a Type II error. Therefore, power or the probability of rejecting a false null must be \( 1 - \beta \), a very good thing. Remember our hypothesis that people who think they are being evaluated perform better? If in fact this is true and our analysis of our data led us to conclude this is true, we had a powerful test of our null. So, most researchers want to do things to increase power.

Remember that we reject the null only if the probability of getting the outcome we got was low, less than alpha, our chosen level of significance. Okay, then, let’s choose a more “favorable” alpha level and make it easier to reject our null hypothesis. Rather than using the typical significance level of \( \alpha = .05 \), let’s choose \( \alpha = .10 \). This would surely increase power because we can reject the null if our outcome is likely to occur by chance only 10%
of the time or less, instead of the more standard 5% or 1%. Sounds good, right? Well, yes
and no. You surely will increase power this way (i.e., you will reject more false nulls), but
sadly you will also increase Type I errors this way—you will reject more true nulls—a
conundrum? Maybe not.

Increasing power—that is our mission! First, let’s talk about how you measure your
variables. Being careful about how you measure your variables can increase power. For
example, the level of measurement of your variables will influence the type of statistical
analysis that you can use. Remember that some statistical analyses are more powerful than
others. Making sure that your variables can be measured on an interval or ratio scale is
wise because then you can use these more powerful parametric procedures. We discuss
these scales in detail in Chapter 5.

What about the design itself? Designs that provide good control over the extraneous vari-
able that might be operating on your participants’ behavior help increase power by
reducing the effects of those nuisance variables on your participants’ behavior. If we can
get rid of the “sloppy” variation, then we are better able to assess the effect of the variable
we really care about (i.e., our independent variable).

Are there any other ways to increase power? Yes there are. We could restrict our sample
to only a specific group of individuals. We might select only women between the ages of
18 and 24 years. This group might be more alike in terms of behavior than people in
general, and so we might reduce variability that way. There is nothing wrong with this
approach, but you need to recognize that you could not generalize your results to the pop-
ulation at large. You really could only make inferences about the population from which you
have taken your sample (i.e., women in this age group).

A simple way to increase power without the problems we have just discussed is to
increase your sample size. Increasing sample size will tend to decrease “sloppy” variabil-
ity. More technically, what happens is that what we call error variance is reduced, and so
the effects of the treatment, the independent variable (IV), are more easily seen. We have
some cautions about overdoing this. Check out the FYI below.

Finally, maximizing your treatment manipulation can increase power. For example, if
your hypothesis is that exercise will increase academic performance, don’t have your par-
ticipants just walk around the block once; have them engage in vigorous exercise for 45
minutes! You do not want to miss obtaining statistical significance because your treatment
was not strong enough.

**FYI**

Is it possible to have too much power? Yes and no. In testing hypotheses, our aim is
to reject the false null hypotheses, and to do that, we need statistical power. But it is
possible to increase your power to such a degree that even very small, perhaps trivial,
effects become statistically significant. For example, small differences in mean perfor-
ance between different racial groups can be statistically significant if you increase
power by using extremely large samples. But a statistically significant difference does not mean that the difference is important. Put another way, it does not mean that you can predict an individual’s performance by knowing his or her race. Why not? Because there is too much variability within each racial group. It is said that men have greater upper body strength than women. Does that mean that if you’re a woman, you are not as strong as the average man? Not at all! You may be stronger than 80% of the population of men. So predicting one individual’s strength because she is a woman is a mistake. A statistically significant difference does not always mean an important difference; a necessary addition to hypothesis testing is to indicate the size of the effect, which we will talk about now.

**EFFECT SIZE**

So let’s say that we have found that a null hypothesis is false. But how big is the effect? Let’s say that in reality, the difference between the hypothesized value and the true value is not exactly zero as stated in the null. What if the true difference is statistically significant but, practically speaking, tiny? We need to ask ourselves, how large a difference should we expect? How large a difference really matters?

Powerful tests should be considered to be tests that detect large effects, that is, large differences between the null and the alternative. This brings us back to the difference between important and significant. A statistically significant effect is not necessarily important and may not be worth the attention of researchers. For example, a study might report statistically significant results that eating a chocolate bar 1 hour before an exam will improve your performance. That the researchers found statistical significance (meaning that the results are not likely due to chance) does not tell us anything about the importance of the effect. Imagine that the chocolate bar group had a mean performance of 82% compared to a no-chocolate bar control group mean of 80%. Are two percentage points likely to make much of a difference to anyone’s academic career? Is this effect important enough to provide all students with chocolate bars? Researchers should specify the minimum difference that is worth their attention and design their studies with that effect size in mind.

**Effect size** can be calculated in various ways. The American Psychological Association (APA) *Publication Manual* lists 13 different effect size estimates that are commonly used (APA, 2001). We will describe a couple of common approaches, but for more detail on effect size calculations, you should consult a statistics book. As indicated above, a statistically significant result does not necessarily mean that your effect is large. This is why it is important to report the group means, but it is also important to take variability into account. In the chocolate bar example, there was a mean difference of two percentage points; does that mean that you will improve two points? Not necessarily. There was a 2% difference between averages. Only if there was no variability in the groups and everyone in the chocolate bar group got 82% and everyone in the control group got 80% could you expect a 2% improvement in your performance, but that is not likely, is it?
Cohen’s $d$ is a common effect size calculation that looks at the difference between the means of two groups and takes into account the variability, using the pooled sample standard deviation.

\[ d = \frac{M_1 - M_2}{SD_{pooled}}. \]

The above method is commonly used with a $t$ test, but if you are investigating a relationship between variables, you may calculate a correlation coefficient such as Pearson’s product-moment correlation. The correlation coefficient $r$ tells you the strength of the relationship between your variables. For example, suppose you investigated the relationship between high school performance and performance in first-year university. You obtained the high school records of 500 students who just finished first-year university and you found $r(498) = .73, p < .01$. This indicates a pretty strong positive relationship between high school and university performance, but does it apply to you?

Should we use this finding to select students for university based on their high school performance? A simple method to calculate effect size for correlations is to square the correlation. $r^2$ indicates the proportion of the variance in the criterion variable that is accounted for by the predictor variable. In our example, .73 squared is .53. Therefore, we have accounted for .53 or 53% of the variability in first-year university performance by using high school performance as a predictor. That leaves a lot of unexplained university performance! Indeed, people often want to use correlation research to predict behavior, and they are often surprised when their variables are not very accurate. $r^2$ is also referred to as the coefficient of determination because it indicates how well you can determine the criterion variable by knowing the predictor variable (i.e., how well you can predict).

A related effect size calculation that is commonly used in analysis of variance (ANOVA) is eta-squared. Eta-squared is a calculation of the proportion of the total variability of the dependent variable that is accounted for by the independent variable. So, in an ANOVA, it is calculated by dividing the treatment sum of squares by the total sum of squares. It indicates the strength of the effect of the independent variable on the dependent variable, and because it is expressed as a proportion of the total variability, you can move the decimal place to the right two places and see what percent of the dependent variable is influenced by your independent variable.

\[ \eta^2 = \frac{SS_{treatment}}{SS_{total}}. \]

Yet more effect size estimates include Cramer’s $v$, which is suitable for the chi-square test for independence, and Cohen’s $w$ for chi-square goodness of fit.
NOTE: Cohen’s \( w \) has the same form as chi-square but uses proportions instead of frequency counts.

How large should your effect size be to consider it important? To some degree, this depends on your research question. But some generally accepted guidelines are as follows:

\[
v = \sqrt{\frac{\chi^2}{N_{tot}}}
\]

(Cramer’s \( v \))

\[
w = \sum \frac{(P_O - P_E)^2}{P_E}
\]

(Cohen’s \( w \))


We have shown you a few simple estimates of effect size, and in Chapter 13, we will calculate these for several examples, but in practice, statistical software packages all provide effect size and power estimates if you ask for them. We have included these calculations just to give you an idea of what effect size means and how to interpret the output from the computer programs.

### Power and the Role of Replication in Research

Students often ask us about the problem that statistical inferences can be wrong. Even with good procedures and good statistical techniques, any inference can be wrong. If we do a study and show that an innovative teaching technique significantly improves student performance, we had better be confident in our inference if the superintendent of schools plans to adopt the new curriculum based on our work. This is where replication of research comes in. If we do another study and find the same thing, the probability that we have made an error of inference drops in a big way. In other words, our power (i.e., the probability that our rejection of the null is correct) has increased when we replicate the finding in a new study with different participants in a different setting, perhaps.

Many psychology students have heard of the study that involved asking people questions and watching which way they directed their gaze as they thought of the answer. It has long
been known that areas in the frontal lobes control movement of the eyes. This research involved monitoring lateral eye movements as people answered questions that required either language processing (left hemisphere) or spatial processing (right hemisphere). Although some researchers have replicated the result, there are many others who have failed to replicate it (e.g., Ehrlichman & Weinberger, 1979).

Could the research have been an example of a Type I error? Perhaps, but at the very least, this underscores the importance of replicability. After all, in 1949, the Nobel Prize in medicine was awarded to Egas Moniz for the transorbital prefrontal lobotomy. It is a real tragedy that practitioners did not wait for independent replication of his claims.

Conceptual Exercise 4C

For each example, indicate if there is a problem of power, effect size, or replication.

1. There was a large difference between the mean for the treatment group and the control group, but we did not achieve statistical significance.
2. There was a small difference between the treatment group and the control group, and we did not achieve statistical significance. Maybe we should increase our sample size.
3. We were amazed that we found a significant difference that has never been reported in the literature.

Research findings, significant or not, that are not replicable may not be important! Replicability is a fundamental requirement of research findings in all science. Replicability is one way we have to improve the validity of our research.

EXTERNAL AND INTERNAL VALIDITY

When we have replicated a finding with different participants or in different settings, we have demonstrated that our research has what is called external validity.

A study, then, is externally valid if the findings can be generalized to other populations and settings. Psychologists are typically interested in human behavior—all human behavior. A finding in your laboratory with your participants, your equipment, and your stimuli that does not replicate in other laboratories with other people and other materials does not have external validity. Research carried out under more natural conditions (i.e., in the field) tends to have greater external validity than research carried out under artificial conditions (i.e., in the laboratory). But research studies must also be internally valid. Internal validity refers to the validity of the measures within our study. Did our independent variable...
cause the changes we see in our dependent variable? Experiments conducted in the laboratory, where control is high, tend to have better internal validity than studies carried out in the field, where control of extraneous variables is more difficult. The internal validity of an experiment or study is directly related to the researcher’s control of extraneous variables.

**Confounding and Extraneous Variables**

An internally valid study is one where the change in the dependent variable is attributable to changes in the independent variable. If a variable that is not manipulated by the researcher affects the outcome, the study lacks internal validity.

An extraneous variable is a variable that may affect the outcome of a study but was not manipulated by the researcher. Such variables are not a problem as long as they are not systematically related to the IV, but they can be a big problem if they do vary in a systematic way. In one example, we talked about the influence of classroom technology on learning. When we compare two groups of students, one of which receives instruction with technology and the other receives instruction without technology, we do not want variables other than the IV (e.g., type of instruction) to influence learning. Some of these might be related to the students, such as differences in learning ability, motivation, gender, and so on. Others might be related to the instructional environment, such as room temperature, time of day, instructor, and so forth. The researcher is not investigating differences in learning ability, motivation, room temperature, or time of day, but nevertheless, these things are still variables that must be taken into account.

Sometimes, not often we hope, an extraneous variable is not controlled or accounted for and unfortunately gums up the works and produces what we call a spurious effect—a researcher’s worst nightmare. An extraneous variable can be a confounding variable if it is systematically related to the independent and dependent variables and, as a result, offers an alternative explanation for the outcome. The outcome is said to be spurious because it was influenced not by the independent variable itself but rather by a variable that was confounded with the independent variable. For example, consider our classroom technology study. Imagine that we decided to compare a group of students who used classroom technology with a group of students who did not. Imagine further that our tech group met at 8 a.m., and our nontech group met at 4 p.m. If we find that our tech group performed better than our nontech group, can we conclude that technology was the reason for the difference? No, we cannot. Even if we randomly assigned students to the two groups and controlled all sorts of things in this study, we have not controlled something that could be systematically related to both the independent and dependent variables (i.e., time of day). It is possible that early morning is a better time for learning, when students are alert and energetic, than late in the day, when students are tired and want to go home. In other words, the time of day difference between the tech and nontech groups is an alternative explanation for the effect and is confounded (i.e., systematically related) with the independent variable, technology.

Extraneous variables become controlled variables when the researcher takes them into account in designing the research study or experiment. Some of those variables are controlled by good experimental design and good research technique. Differences in room temperature, noise level, and testing apparatus can be controlled by using a soundproof, constant-temperature laboratory. And indeed, many researchers prefer laboratory research
because it is easier to control extraneous variables such as these. We want to control extraneous variables because they contribute variance to our dependent measures and cloud the results. We often call these variables *nuisance variables*. The better we can keep these variables out of our results, the better we can estimate the influence of the independent variable on the dependent variable.

Researchers need to think about the variables operating in their experiments that might contribute unwanted variability or, worse, be confounded with the IV. And then they need to control those sources of variance so that they can evaluate the effect of the IV.

As we have said, the internal validity of a research study depends on the researcher’s control over extraneous variables. The better we control other sources of variation, the better we can assess the effects of the variables we are interested in. Next we will look at several ways to do this.

**Controlling Extraneous Variables**

**Elimination**

One way to control extraneous variables is to get rid of them entirely. For example, if you think that ambient noise might influence the outcome of your study, then eliminate it by conducting your study in a noise-proof environment. The disadvantage of eliminating a variable is that you will not be able to generalize your results to those conditions that you eliminated. Would you have found the same results in an environment where ambient noise was present?

It is easier to eliminate extraneous variables in laboratory settings than in field settings. Laboratories can be soundproofed, but the natural environment cannot.

Lots of extraneous variables, particularly individual differences in ability, motivation, and so on, cannot be eliminated, of course. How would we eliminate room temperature from our study? Well, we couldn’t, of course, but we could hold it constant for all groups.

**Constancy**

Most, if not all, researchers use constancy. Instructions to participants are constant, the procedure (except for the IV) is constant, the measuring instruments are constant, and the questions on a questionnaire are constant. In Chapter 2, we discussed Knez’s (2001) study on lighting. He used constancy to control noise and room temperature in his study.

He also kept gender constant by including 54 women and 54 men.

But what if we thought gender might be interesting? Not as interesting as our primary IV but interesting nevertheless. We could make gender another variable to study, which indeed is exactly what Knez did. He included gender as a participant variable and included it as a variable of secondary interest in his analyses.

**Secondary Variable as an IV**

Often researchers make variables other than the primary IV secondary variables to study. Knez (2001) reported research indicating that men and women respond differently to various light conditions—and he claimed that gender is a moderating variable. He decided it was important to not only control for gender with an equal number of men and women in
each group but also to examine how gender differences affect the way the IV acts on the dependent variable (DV). This type of control is not only a design decision but also a statistical decision because the secondary variable will be included in the statistical analysis used to assess the effects of the primary and any secondary IVs on the DV.

Elimination, constancy, and making an extraneous variable a secondary variable are all techniques used to control extraneous variables that might affect the outcome of the study (i.e., variation in the dependent variable).

What if there are extraneous variables operating that we do not know about? This is a good question, and the answer is randomization. Perhaps the most common way to control extraneous variables is random assignment of participants (or subjects, if they are animals) to groups.

**Randomization: Random Assignment of Participants to Groups**

Before we can conclude that our manipulation of the IV, and nothing else, caused a change in responding, the DV, we must be confident that our groups were equivalent before we treated them. If they were not equivalent at the outset of our study, then it will be no surprise that they are different at the end of our study. Knez (2001) randomly assigned his participants to three light conditions, the three levels of his primary IV. Random assignment of participants to conditions of the study does not guarantee initial equivalence of groups, but we can be reasonably confident that our groups will be more or less similar if participants were truly randomly assigned to groups. This method of controlling extraneous variables is probably the most common procedure used by researchers, particularly when they have large samples. With smaller samples, another common method of dealing with unknown extraneous variables is to test all the participants in all the conditions—the within-participants or repeated-measures design.

**Repeated Measures**

What better way to control for unknown variation among participants than to use the same participants in all conditions, a **repeated-measures design**. Each participant acts as his or her or its (if we are talking about animals) own control because changes in the DV are compared within each participant (or subject, if we are talking about animals). Gender, IQ, motivation level, and so on are all held constant because comparisons are always made between measures from the same participant. A repeated-measures procedure cannot be used when the treatment conditions may result in lasting changes in responding. For example, you cannot train someone to solve problems using one strategy and then expect that person to forget the first strategy and use a different strategy when solving a second set of problems. Training cannot be unlearned. This issue is discussed in greater detail in Chapters 7 and 8.

What if we discover, after the data have been collected, that the groups were not initially equivalent despite our best efforts? Sometimes statistics can help.

**Statistical Control**

Sometimes we find that a variable we did not control seems to have affected our outcome. In such cases, we can use statistics to help us. If we are lucky enough to have measured the variable, we can treat it as a covariate and use statistical procedures (e.g., analysis of covariance...
To remove it from the analysis. Basically, we can use statistical procedures to estimate the influence of the extraneous variable on the dependent measure and remove this source of variation from the analysis. There are statistical techniques that allow us to take into account extraneous variables that have crept into our data that we discovered only after the fact; such techniques would be covered in most upper-level statistics texts.

**Conceptual Exercise 4D**

What method of control would you suggest for each of the following extraneous variables? Why?

1. Gender of trainer when type of training is IV
2. Breed of rat when diet is IV
3. Intelligence, attitude, capability, and motivation of child when reinforcement strategy is IV
4. Fitness when exercise program is IV

**CHAPTER SUMMARY**

The most common **inferential** approach in psychology is **hypothesis testing**. A **conceptual hypothesis** is a statement about the expected relationship between conceptual variables. A **research hypothesis** is a statement about the expected relationship between measurable, or **operationalized**, variables. A **statistical hypothesis** is a statement about the relationship between statistical properties of data.

The **null hypothesis** is a statistical hypothesis about the value of a parameter, the relationship between two or more parameters, or the shape of a distribution and is the hypothesis that the researcher hopes to reject. The **alternative hypothesis** specifies something different than what is stated in the null and is the hypothesis the researcher hopes to confirm.

A **statistically significant** finding means that the relationship specified in the null is very unlikely to be true, and so the researcher rejects the null and accepts the alternative. Testing the null hypothesis follows these steps: State the null and alternative, collect the data and conduct the statistical test, reject the null or fail to reject the null, and state the conclusion.

A **statistical inference** is a statement of **probability**; as such, there is always some probability that the inference is incorrect. A **Type I error** occurs when our statistical test leads us to reject a true null. A **Type II error** occurs when our statistical test leads us to fail to reject a false null.

Powerful tests of the null are tests that have a high probability of rejecting false nulls. **Power** can be increased by controlling extraneous variables, using interval or ratio scales for the dependent measure, choosing good research designs, using appropriate sampling procedures, choosing powerful significance tests, choosing lenient alpha levels, increasing
sample size, reducing participant variability, maximizing the treatment manipulation, and replicating the findings.

**Effect size** estimates should be included along with other statistics in any research report. Common effect size estimates include *Cohen's $d$*, suitable for *t* tests; $r^2$; the *coefficient of determination*, suitable for Pearson's *r* correlation test; and *eta-squared*, suitable for ANOVA.

**Replicability** of research findings is an important part of the research process. A research finding that has been duplicated with different participants or in different settings is *externally valid*. When an IV has been shown to cause change in the DV, the study has **internal validity**. **Confounding** and *extraneous variables* must be controlled for a study to be internally valid.

Extraneous variables can be controlled by *elimination, constancy, making a secondary variable an IV, randomization*, using a *repeated-measures design*, and statistical techniques such as *ANCOVA*.

**Answers to Conceptual Exercises**

**Conceptual Exercise 4A**

There are numerous ways of operationalizing the variables in this conceptual hypothesis.

We might operationalize the concept "helping profession" to include clinical psychologists, psychiatric nurses, and psychiatrists, for example.

We might operationalize "business professions" as bankers, CEOs, and retailers.

How would we operationalize "ethical"? Well, we could create a series of ethical dilemmas, the solutions to which have already been rated by experts on some scale of ethicality (Is that a word? We are not sure).

You can see, we hope, that the operational definitions you choose for your research project are up to you. They might be good (i.e., valid), or they might be bad. We suggest you read the literature to find out what other researchers do to operationalize any constructs you want to use in your project.

**Conceptual Exercise 4B**

1. Your statistical decision was correct. You did not reject the null, and that was correct because the null was true (i.e., there was no difference).
2. Oh dear. You rejected the null, but the null was true. You have made a Type I error.
3. Oh dear again. Your statistical test did not find a significant difference, and so you did not reject the null. Sadly, the null was false. You have made a Type II error.
4. Good on you. The null is false, and your conclusion, based on your significance test, was just that.

**Conceptual Exercise 4C**

1. Hmm, well, it appears that the treatment effect was large. Probably we have a power problem, which could have been caused by all sorts of things. Perhaps we used a weak
statistical test, or maybe we had poor control over extraneous variables. Probably we need to redo our study with more attention to control of variables and reconsider the significance test we should use.

2. Maybe so, but maybe the effect is not very important. We might achieve a significant result if we increase our sample size in a big way, but would it be important?

3. Aha, time to replicate. Let’s not report a finding, even if significant, that is contrary to the existing literature on the topic. If we obtain the same result a second time, with different participants at a different time and in a different setting, perhaps we should go ahead, report our contradictory finding, and feel confident about it.

Conceptual Exercise 4D

1. There are various ways to control for gender. Perhaps the simplest way is to hold it constant.

2. Constancy is probably the best way to control breed of rat.

3. Randomization or repeated measures would be appropriate.

4. Randomization or constancy would be appropriate.

FAQ

Q1: Why do we need inferential statistics?

A1: We need inferential statistics whenever we want to make statements about populations based on sample data. We have not included everyone in our sample, so we cannot be 100% sure that our conclusions will generalize to the population. Inferential statistics provide us with a probability estimate of how likely it is that we are incorrect in our inference. If the probability that our outcome would have occurred by chance is low (e.g., \( p < .05 \) or .01), we can be more confident that our inference is correct—not 100% certain, but more certain. Only by measuring the entire population can you be certain that you know the true state of affairs.

Q2: Why do we test the null hypothesis and not our research hypothesis?

A2: We test the null hypothesis because it is much easier to disprove than it is to prove. So instead of gathering endless data to confirm our research hypothesis, we set out to disconfirm a null hypothesis. One instance of disconfirmation, and we can reject the null and accept the alternative hypothesis (i.e., our research hypothesis). But perhaps more important, we test the null because the null provides us with a clear expectation of outcomes (called the sampling distribution). We then compare our outcome with what the null would predict. After all, we all know exactly how a fair coin behaves. But a biased coin could be slightly biased toward heads, moderately biased toward tails, extremely biased toward heads, and so on. Remember that our alternative hypothesis specifies that the null is false, and in a one-tailed test in what direction, but it does not specify the expected size of the difference.
Q3: Why is inference always stated as a probability? Isn’t science interested in proving things?
A3: In science, we are interested in finding empirical support for our research hypotheses. Because we work with samples and not entire populations, we have to make statistical inferences, and in doing so, there is always a possibility that our inference is wrong. Even research with statistical significance will have a nonzero probability that the results are a fluke and that another outcome may arise using a different sample. That is why the scientific community waits for independent researchers to replicate a finding before getting too excited.

Q4: How can you increase your statistical power?
A4: (1) Increase your sample size, (2) increase your alpha level, (3) decrease the variability among your participants, and (4) increase the magnitude of your treatment.

Q5: Why does the APA publication manual suggest including effect size estimates with your inferential statistic?
A5: Statistical significance tells us that the results are not likely a fluke and that they are probably replicable. However, significance does not indicate the size of the treatment effect or the strength of a correlation. Both pieces of information are needed to assess the importance of a study.

Q6: What is the coefficient of determination?
A6: The coefficient of determination is a measure of effect size that tells us how much variability in the criterion variable we can account for by its relationship with the predictor variable. It ranges from 1, where we have accounted for all the variability and can predict with perfect accuracy, to 0, where we have accounted for no part of the criterion variable, and our predictor does not predict at all (coefficient of determination = $r^2$).

Q7: Why is replication important in science?
A7: A single study that reports statistical significance may be a Type I error. After all, depending on our alpha level, Type I errors are expected to occur about 5% or 1% of the time. Replication is important to reduce the likelihood that a research outcome was just a fluke or Type I error. Having the research replicated by an independent group of researchers also guards against any researcher bias influencing the outcome.

Q8: I remember something about one-tailed or two-tailed t tests. What is the difference?
A8: In performing a t test, you compare two means and test a null hypothesis that they do not differ. Your alternative hypothesis can be either that the means differ and you do not specify which mean will be larger than the other (two-tailed), or you may specify that one mean will be larger than the other (one-tailed). A one-tailed test has more power, so there is strong incentive for using a one-tailed test. In using a one-tailed test, you are only looking for a difference between the means in one direction, so you must have good reason for predicting the direction of the difference (usually based on previous research). The reason a one-tailed
test has more power is that the critical value of \( t \) is smaller because your region of rejection is all in one tail rather than divided into each end of the \( t \) distribution. See Chapter 13 for more detail on one- and two-tailed tests.

**CHAPTER EXERCISES**

1. What is an inference?
2. What is the relationship between alpha level and statistical power?
3. What are the four outcomes of any inferential decision?
4. List three ways to increase power.
5. Why is effect size important?
6. Your research hypothesis is that women have a keener sense of smell than men. You measure the minimum amount of perfume that can be detected in a room as your measure of olfactory acuity. Based on your sample of women and men, you find that women do have a better sense of smell. Your inferential statistic has a \( p < .02 \), which means the difference was not likely due to chance. You decide to reject the null hypothesis and accept the alternative. You conclude that women have a keener sense of smell. What are the chances that you made a Type I error? What are the chances you made a Type II error? How would you know for certain that your decision was correct?
7. You read a study that found a strong positive correlation between high school performance and university performance (\( r = .68 \)). Thinking back on your own experience, you did really well in high school but flunked out of your first year in university. Just how well does high school performance predict performance in university?
8. Green hand syndrome is an embarrassing condition where a person’s hands turn green in social situations. You have reason to hypothesize that drinking a single cup of tea may be an effective treatment. How would you test this hypothesis? What alpha level would you select (include a rationale)? How would you interpret your results if your effect size is small?
9. What method of control is being used in the following?
   a. A researcher concerned about motivation, fatigue, ability, interests, and so on of her participants randomly assigns them to groups
   b. A researcher concerned about IQ differences decides to select only people with IQs between 95 and 105 as participants in his study
   c. A researcher concerned about gender effects uses only female participants in her study
   d. A researcher concerned about gender effects creates groups with equal numbers of men and women and decides to analyze any differences

**CHAPTER PROJECTS**

1. With a search term of your choice, find three empirical research articles. Report the inferential statistics from each article and include the degrees of freedom and the \( p \) value. Describe why the researchers selected those specific tests. What was the reported effect size? For their statistically significant findings, do you think their effect is important?
2. With a search term of your choice, find three empirical research articles. Describe the methods the researchers used to increase the power of their analysis. Can you think of other ways to increase power in each study?

3. Create a research project of your own and describe how you would operationalize the variables. Describe your IV, DV, and any control procedures you would need to use.

REFERENCES


Visit the study site at www.sagepub.com/evansmprstudy for practice quizzes and other study resources.