33 Power and the Factors Affecting It

Term: power

Learning Objectives:

- Know the mathematical relationship between Type 2 error and power
- Understand the effect that various factors have on power
- Understand the interrelationship of statistical significance, effect size, and power
- Use graphs or tables to determine power, given known values of other variables
- Use graphs or tables to determine effect size, given known values of other variables
- Use graphs or tables to determine necessary sample size for a desired power, given known values of other variables

What Is Power?

In previous modules, we discussed statistics for testing null hypotheses. Each of the statistics assumed the null hypothesis to be true. To reject the null hypothesis, we looked for a big enough difference between treated and untreated groups. In Module 13, you learned that Type 1 error, α, occurs when there is really no treatment effect in the populations, but we nevertheless find one in our samples. Thus, with Type 1 error, we incorrectly reject the null hypothesis. This situation is shown in Figure 33.1.

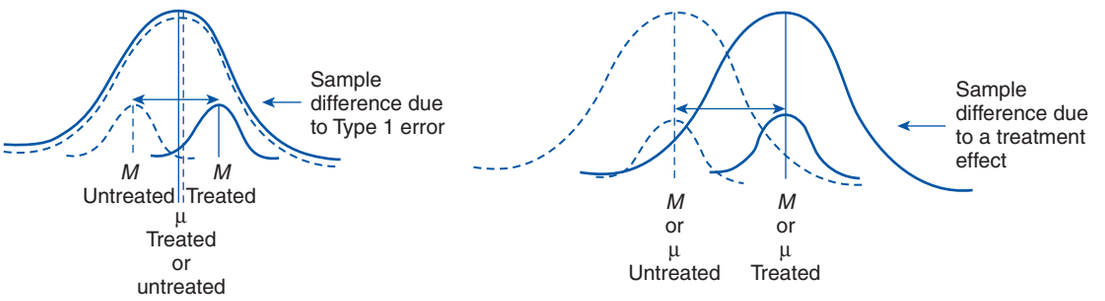

Figure 33.1 Rejecting the Null Hypothesis: Type 1 Error Versus a Treatment Effect

So is that enough? No, that's not enough. Experimental studies need to be able to correctly detect not only when there is *not* a real difference in treatments but also when there *is* a real difference in treatments. We are moving to a new topic—Type 2 error and its inverse, power.

The alternative hypothesis states that the null hypothesis is false. In other words, there really is a treatment effect in the underlying populations. In Module 13, you learned that Type 2 error, β, occurs when there really is a treatment effect in the populations, but we do not find it in our samples. With a Type 2 error, we incorrectly retain the null hypothesis. This situation is shown in Figure 33.2.

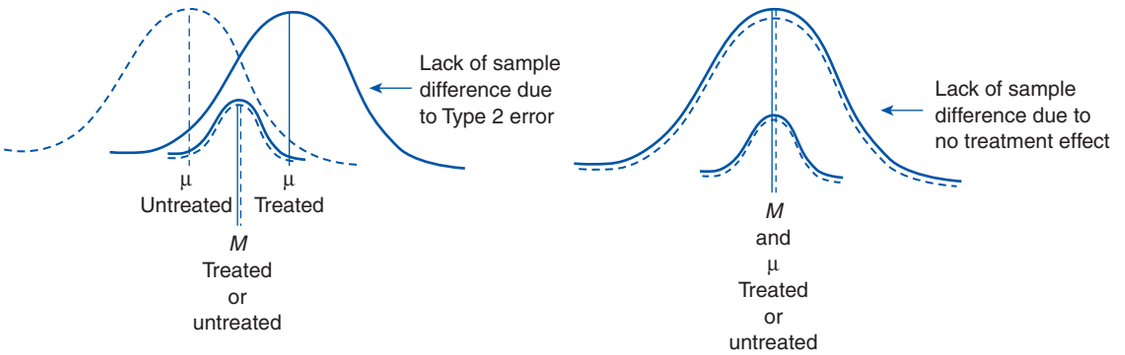

Figure 33.2 Retaining the Null Hypothesis: Type 2 Error Versus No Treatment Effect

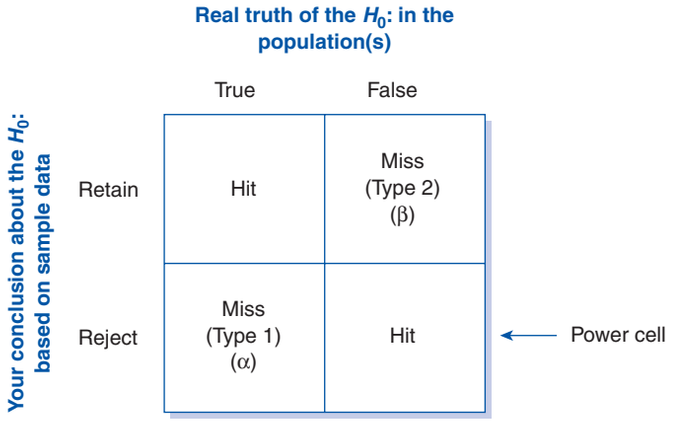

Figure 33.3 Decision Table With Power Cell

> When you have eliminated the impossible, whatever remains, however improbable, must be the truth.
>
> —Sir Arthur Conan Doyle

Recall, also, that Module 13 depicted both types of errors in a decision table (Figure 33.3).

The left column of the decision table refers to the null hypothesis. When the difference between treated and untreated groups is not big enough to reject the null hypothesis, we either correctly retain the null hypothesis (hit) or make a Type 1 error (α). These are the only two possibilities with a true null hypothesis. Because we have previously tested only the null hypothesis, α is the only error we have measured.

Now we will work within the right column of the decision table. Thus, we will be concerned about β. Note the new notation in the decision table—the power cell. **Power** occurs when the null hypothesis is really false, and we correctly reject the null hypothesis. To put it another way, power is when there really is a difference between groups due to the treatment, and we do find it.

Now look again at the decision table. When the null hypothesis is really false (right column), either we correctly find it to be false (power—which is a hit) or we fail to find it to be false (β). These are the only two possibilities with a false null hypothesis. Because either one or the other must occur, the two possibilities taken together add to 1. It follows that we can calculate power by knowing the value of Type 2 error. That is, power = 1 – β. Thus, if β is .36, then power is .64. Conversely, we can calculate Type 2 error by knowing the power of our study. That is, β = 1 – power. Thus, if power is .82, then β is .18.

CHECK YOURSELF!

State the relationship between power and β.

Until fairly recently, many studies had very little power. That is, the researcher found no statistically significant result and, hence, retained the null hypothesis, but the researcher did not know if the lack of observed difference between the means was because there really was no treatment effect or because of a Type 2 error. That is, it is quite possible that the treatment really was effective but the study simply lacked power to detect it.

Because of this unfortunate past situation, researchers now design studies that can detect not only when the null hypothesis is really true but also when it is really false, and the results are misleading the researcher to think that it is true. All journals published by the American Psychological Association (APA), as well as many other journals of rigor, now require that authors submit power estimates for their studies. In this module, we will look at factors that increase the power of a study as well as methods for estimating a study's actual power.

CHECK YOURSELF!

The results from a rigorous study must meet three requirements. One is statistical significance. Based on this module and the previous module, what are the other two requirements?

PRACTICE

1. Calculate the following.
   a. When power is 80%, what is the Type 2 error?
   b. When Type 2 error is 12%, what is the power?
2. Calculate the following.
   a. When power is 84%, what is the Type 2 error?
   b. When Type 2 error is 25%, what is the power?

Factors Affecting Power

Because power occurs when there really is a treatment effect and we detect it, power is a very good thing. We want lots of power in our studies. So what factors affect power? The factors

include size of the Type 1 error, directionality of the alternative (research) hypothesis, size of the actual difference between the means of treatment groups (effect size), amount of error variance, and sample size. We will look at each of these factors in a moment. But first, let me explain the diagrams that accompany each factor.

In each of the diagrams to follow, Type 1 error is indicated by the small area beyond the dotted line in the tail of the H_0 distribution. This is where Type 1 error has fallen all along. What's new is the addition of a second overlapping distribution. The overlapping distribution represents scores in a treatment group when the treatment does have an effect—that is, when the alternative hypothesis is true (the null hypothesis is false). Therefore, the overlapping distribution is labeled H_A. Power is indicated by the portion of the overlapping H_A distribution falling to the right of the dotted line in the H_0 distribution. For clarity, the area of power is shaded. Now let's look at the factors that influence power.

> If you want to test a man's character, give him power.
>
> —Abraham Lincoln

Size of Type 1 Error

As Type 1 error increases, power increases. Figure 33.4 can help us understand this principle.

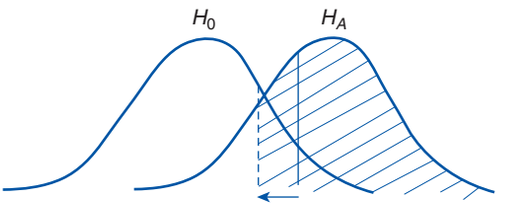

Figure 33.4 Power and the Amount of Type 1 Error

Increasing Type 1 error means accepting a greater chance of making a Type 1 error. For example, we could accept a 10% chance of Type 1 error rather than a 5% chance. Figure 33.4 shows that, as Type 1 error increases under the H_0 (indicated by the direction of the arrow), power (the shaded area under H_A) also increases.

Directionality of the Alternative Hypothesis

For any given Type 1 error level, directionality increases power. Figure 33.5 can help us understand this principle.

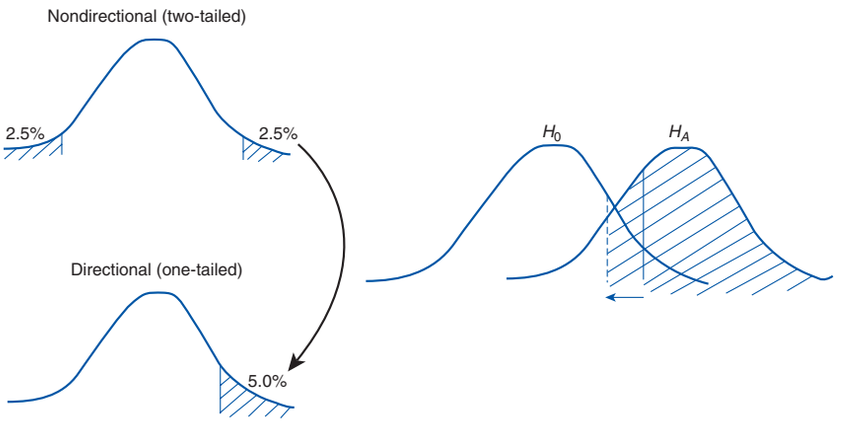

Figure 33.5 Power and Directionality of the Alternative Hypothesis

Recall that the alternative hypotheses can be either directional (one-tailed) or nondirectional (two-tailed). Let's pick a Type 1 error level—say, 5%. In a nondirectional hypothesis, that 5% is split so that 2.5% is in each tail. A nondirectional hypothesis tests whether the experimental group scores higher than the control group or the control group scores higher than the experimental group—either outcome. In contrast, in a directional hypothesis, the whole 5% is placed in one tail. It tests only whether the experimental group does better than the control group or only whether the control group does better than the experimental group—one or the other but not both.

The two diagrams on the left show that Type 1 error in a single tail increases (indicated by the arrow) as the hypothesis switches from nondirectional to directional. In Figure 33.5, it goes from 2.5% to 5%. Because that is the tail of the H_0 distribution that overlaps with the H_A distribution, power (shaded area in the H_A) increases. This is shown by the arrow in Figure 33.5 on the right.

Size of the Actual Difference Between the Means

As the actual difference between the means (you might recognize this as the effect size) increases, power increases. Figure 33.6 helps us understand this principle.

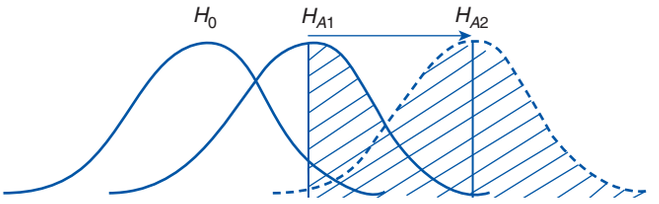

Figure 33.6 Power and the Actual Difference Between the Means

If the experimental and control groups differ in their response to treatment by only a little (from H_0 to H_{A1} in Figure 33.6), we might or might not detect the difference in our samples. But as the difference between the experimental and control groups increases (from H_0 through H_{A1} and on to H_{A2} in Figure 33.6), we are more likely to detect the difference between our H_0 and H_A samples. In Figure 33.6, power (shaded area under H_A) increases as the difference between the means increases. In the extreme case, we could depict yet another alternative hypothesis, H_{A3}, so far away from H_0 that the two distributions do not overlap at all. In that case, we would have 100% power: No matter which cases ended up in our samples, we would definitely find the difference, because even the lowest members of H_A would outscore the highest members of H_0.

The principle we have just examined deals with the size of the difference between the means. Now recall the formula for a two-sample t test:

$$t_{\text{2-samp}} = \frac{M_1 - M_2}{\sigma_{M_1 - M_2}}$$

Where does this difference between the means lie? Does it lie in the numerator or in the denominator?

It resides in the numerator. Thus, anything that increases the difference between the means increases our ability to find treatment differences. Anything that decreases the difference between the means decreases our ability to find treatment differences.

Amount of Error Variance

The greater the error variance (or the standard deviation), the less the power. Figure 33.7 helps us understand this principle.

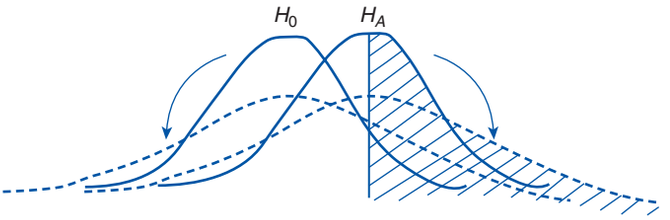

Figure 33.7 Power and the Amount of Error Variance

> A man is like a fraction whose numerator is who he is and whose denominator is what he thinks of himself. The larger the denominator, the smaller the fraction.
>
> — Leo Tolstoy

In Figure 33.7, as the distributions become more variable within their own groups (indicated by the arrow), there is less overlap between the H_0 and H_A groups. Hence, power (shaded area in H_A) decreases. H_A, H_0, or both can increase in variability. The rule remains the same: Within-group variability decreases power, regardless of the group in which it occurs.

The principle we have just examined deals with the size of the standard deviation (variability within groups). Consider again the formula for a two-sample *t* test:

$$t_{2\text{-samp}} = \frac{M_1 - M_2}{\sigma_{M_1 - M_2}}$$

Where does this within-group variability reside in the *t* statistic? Does it lie in the numerator or in the denominator? To answer that, recall that $\sigma_{M_1 - M_2}$ is calculated from the σ_M for each group and that one of the variables in the σ_M is the size of the standard deviation.

So it resides in the denominator. Thus, anything that increases the variability within groups will decrease our ability to find any treatment difference that does exist. That is, too much within-group error variance masks the treatment effect between groups in the numerator. Conversely, anything that decreases the variability within groups will increase our ability to find any treatment differences that do exist. That is, less within-group variance will allow the treatment effect between groups in the numerator to be more apparent.

Sample Size

The bigger the sample size, the greater the power. This principle is shown in Figure 33.8.

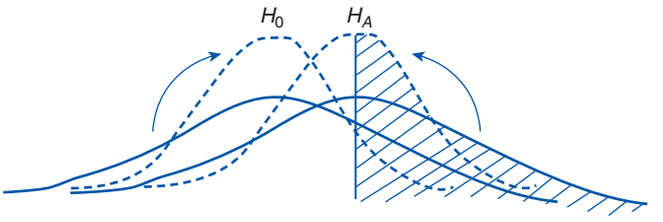

Figure 33.8 Power and Sample Size

Figure 33.8 shows that, as sample size increases, within-group variability (error variance) decreases. Thus, power (the shaded portion of H_A) increases.

The principle we have just examined deals with sample size. Let's again consider the formula for a two-sample *t* test:

$$t_{2\text{-samp}} = \frac{M_1 - M_2}{\sigma_{M_1 - M_2}}$$

Where does sample size reside in the *t* statistic? Does it lie in the numerator or in the denominator? To answer that, recall that $\sigma_{M_1 - M_2}$ is calculated from the σ_M for each group and that one of the variables in the σ_M is the sample size.

It resides in the denominator. Now recall the number-drawing exercise (Exercise 7) you performed in Module 15 when we studied the standard error of the mean. From that exercise, you learned that the larger the sample size, the smaller the σ_M (error variance). Thus, sample size is one variable that affects the value of the σ_M and, hence, the $\sigma_{M_1 - M_2}$.

We also learned in the preceding section that anything that increases the value of the denominator will decrease the ability to find any treatment difference that does exist in the numerator. And conversely, anything that decreases the value of the denominator will increase the ability to find any treatment difference that does exist in the numerator. Because larger sample sizes decrease the value of the denominator of our test statistic (*t*, *F*), larger sample sizes increase our ability to find treatment differences in the numerator.

CHECK YOURSELF!

List five factors affecting power. For each factor, tell the direction of the effect.

Factor	*Direction of Effect*
1.	
2.	
3.	
4.	
5.	

PRACTICE

Directions for this set of practice exercises: A researcher is conducting studies under the following conditions. For each study,

a. State whether the change in condition will tend to increase or decrease the study's power.
b. If the change in condition affects primarily the numerator or the denominator of the test statistic, state which part is primarily affected.

3. Fifty subjects are randomly selected and agree to participate in an experiment. However, on the day of the experiment, only 32 subjects show up.
4. A researcher had originally intended to select subjects for placement into treatment groups (i.e., the independent variable) based on a median split of their scores on a

placement test for the independent variable. However, now he has decided to put in one group only people scoring at least 1 standard deviation above the mean on the placement test and put in the other group only people scoring at least 1 standard deviation below the mean on the placement test. Sample size remains the same.

5. A research assistant is helping to collect observational data. However, he is careless in scoring the observations. As a result, dependent variable scores are inaccurate and prone to error.

6. A researcher had originally planned to conduct a study while allowing only 1% Type 1 error. However, now he has decided to permit 5% Type 1 error.

7. Just before conducting an experiment, the researcher switches his hypothesis from nondirectional to directional. He maintains the same total Type 1 error.

Putting It Together: Alpha, Power, Effect Size, and Sample Size

In my undergraduate course in economics, the professor graphed the effect of a change in an economic variable on an economic outcome. As I recall, he graphed the effect of a rise in loan interest rates on the purchasing power of consumers. However, he always qualified the effect with a Latin phrase: "ceteris paribus." If you have taken a course in either economics or Latin perhaps you've heard this phrase. It means "all other things remaining the same." In other words, what is the effect of a change in this variable when this variable is the *only* one to change? Unfortunately, a single variable is rarely the only one to change. For example, an increase in loan interest rates slows consumer spending in some areas while spurring compensatory spending in other areas. This, in turn, brings about changes in yet other areas and . . . well, suddenly the relationship is no longer simple.

We see this same complexity in behavioral research. Recall that one reason for conducting a factorial ANOVA rather than a series of *t* tests is that independent variables rarely affect the dependent variable in a simple fashion. Rather, the independent variables tend to *interact* to bring about a more complex outcome.

The same principle applies to power analysis. If all other things remained constant, 100% power would be the ideal. However, as we gain power to detect a false null hypothesis we also increase the risk of falsely rejecting a true null hypothesis. That is, as power goes up, Type 1 error also goes up. Certainly, that's not a good outcome. In addition, some methods of increasing power—for example, increasing sample size—also increase the chance of finding statistical significance when the meaningful effect size is actually quite small. That, too, is not a good outcome. So what is the best amount of power? Although no single value is best, the best trade-off among competing goals hovers somewhere around 80% power.

Now let's look at how we can estimate the actual power of a study. Power can be estimated from formulas originally found in a book by Jacob Cohen (1988). However, the formulas are cumbersome to use. In the same book, Cohen presents a series of tables derived from the formulas. Power tables show the interaction between six variables: power, alpha, effect size, directionality of the hypothesis, independence of the samples, and sample size. The tables aid in determining a study's power, given a known alpha, effect size, directionality, independence, and sample size. They also aid in determining ahead of time the necessary sample size for a study with a given alpha, desired power, projected effect size, given directionality, and given independence.

Because a power table must portray interaction among six variables, which is more than can be effectively shown in a single table, we use a series of tables to depict the interaction of only three of the variables: power, effect size, and sample size. The remaining three variables—alpha, directionality, and independence—are depicted via separate tables for each condition.

The entire set of tables is found in Cohen's book. For illustrative purposes, we will look at just three tables. Table 33.1 is for a two-sample nondirectional independent *t* test. Table 33.2 is for a one-way ANOVA. Table 33.3 is for a chi-square test of independence. Each table assumes .05 α. For the two parametric tests (*t* test and ANOVA), independent samples and equal sample sizes are also assumed.

Table 33.1 Sample Size Needed per Group to Obtain .80 or .90 Power, Given Various Cohen's *d*s or Effect Size *r*s, for a Two-Sample Independent *t* Test With Equal Sample Sizes and .05 Nondirectional α

Cohen's d	*Effect Size r*	*.80 Power*	*.90 Power*
.10	.063	1,571	2,102
.20 (small)	.100 (small)	393	526
.30	.148	175	234
.40	.196	99	132
.50 (medium)	.243 (medium)	64	85
.60	.287	45	59
.70	.330	33	44
.80 (large)	.371 (large)	26	34
.90	.447	17	22
1.00	.514	12	16

SOURCE: Adapted from Cohen (1988).

Table 33.2 Sample Size Needed per Group to Obtain .80 or .90 Power, Given Various Effect Size ηs, for an Independent Sample's One-Way ANOVA With Equal Sample Sizes and .05 α

	.80 Power			*.90 Power*		
	Number of Groups			*Number of Groups*		
Effect Size η	*3*	*4*	*5*	*3*	*4*	*5*
.100 (small)	322	271	240	417	350	310
.196	80	68	60	106	88	78
.243 (medium)	52	44	39	68	58	50
.287	36	31	27	48	40	35
.371 (large)	21	18	16	27	23	20
.447	14	12	11	18	15	13
.573	8	7	6	10	9	8

SOURCE: Adapted from Cohen (1988).

Table 33.3 Total Sample Size Needed to Obtain .80 or .90 Power, Given Various Effect Size ϕs or *V*s, for a Two-Variable Chi-Square Test of Independence at .05 α and 1 to 5 *df*

	.80 Power					*.90 Power*				
Effect Size or V	*1 df*	*2 df*	*3 df*	*4 df*	*5 df*	*1 df*	*2 df*	*3 df*	*4 df*	*5 df*
.100 (small)	783	960	—	—	—	—	—	—	—	—
.200	196	240	275	300	322	262	320	350	388	417
.300 (medium)	88	108	120	134	143	117	140	160	170	185
.400	49	60	68	75	80	66	80	90	97	104
.500 (large)	32	38	44	48	51	43	50	57	62	66

SOURCE: Adapted from Cohen (1988).

Examine the three tables. It is apparent from each table that, as effect size (left column) increases, necessary sample size (table entries) decreases. This makes sense because, as you learned in Module 32, the bigger the actual treatment effect (i.e., the bigger the difference between the means), the greater the probability of finding that effect. When the actual difference between means is very large, we will find it even with a small number of subjects.

It is also apparent from each table that, as desired power (across the top) increases, necessary sample size (table entries) also increases. This, too, makes sense. Recall from the discussion earlier in this module that smaller samples contain greater error variance and that error variance lies in the denominator of the significance test statistic. Too much error variance in the denominator masks the treatment effect in the numerator. Thus, the more certain we want to be of actually finding an existing treatment effect, the more subjects we need.

Now, let's try reading one of the tables. Reading down the .80 power column in Table 33.1, when Cohen's *d* is .20 or when effect size *r* is .100, a study with .80 power requires 393 subjects per group. However, when Cohen's *d* is .50 or when effect size *r* is .243, only 64 subjects per group are needed. With a Cohen's *d* of .80 or with effect size *r* of .371, only 26 subjects per group are needed.

The other way of reading the table is to estimate actual power once a study is complete. From the same table, if we conduct a study with 26 subjects per group and find Cohen's *d* to be .80 or find effect size *r* to be .371, the power of our study is .80. But if we find Cohen's *d* to be .50 or find effect size *r* to be .243 and have only those same 26 subjects, we have fallen short of 80% power. Complex formulas and more complete tables would tell us exactly how short our power fell.

Each of the tables reads similarly. Note that sample size is per group for the parametric studies (*t* test, ANOVA), but sample size is total for chi-square. This is because chi-square has no "groups," only categories across variables.

PRACTICE

Directions for this set of practice exercises: Use Power Tables 33.1, 33.2, and 33.3 to answer the questions.

8. You are designing a study to be evaluated with a two-sample nondirectional independent *t* test at α = .05. You want .90 power. You expect Cohen's *d* to be about .40. How many subjects will you need per group?

9. You are designing a study to be evaluated with a two-sample nondirectional independent *t* test at α = .05. You want .80 power. You expect Cohen's *d* to be about .70. How many subjects will you need per group?

10. You are designing a study to be evaluated with a two-sample nondirectional independent *t* test at $\alpha = .05$. You are able to obtain 25 subjects per group. Approximately what will effect size *r* have to be for you to obtain .80 power?
11. You are designing a study to be evaluated with a two-sample nondirectional independent *t* test at $\alpha = .05$. You are able to obtain 65 subjects per group. Approximately what will Cohen's *d* have to be for you to obtain .80 power?
12. You calculate a two-sample nondirectional independent *t* test at $\alpha = .05$. Your study has 130 subjects per group and effect size *r* is .20. To the nearest 10%, what is the power of your study?
13. You calculate a two-sample nondirectional independent *t* test at $\alpha = .05$. Your study has 45 subjects per group and Cohen's *d* is .60. To the nearest 10%, what is the power of your study?
14. You are designing a three-group study to be evaluated with a one-way ANOVA at $\alpha = .05$. You expect effect size η to be about .29. How many subjects will you need per group for .80 power?
15. You are designing a four-group study to be evaluated with an ANOVA at $\alpha = .05$. You are able to obtain 40 subjects per group. Approximately what will effect size η have to be for .90 power?
16. You calculate a one-way ANOVA at $\alpha = .05$. Your study has four groups with 30 subjects per group. Effect size η is .30. To the nearest 10%, what is the power of your study?
17. You are designing a study to be evaluated with a chi-square test of independence at 3 *df* and $\alpha = .05$. You expect Cramer's *V* effect size to be about .30. How many total subjects will you need for .80 power?
18. You are designing a study to be evaluated with a chi-square test of independence at 4 *df* and $\alpha = .05$. You are able to obtain 60 total subjects. Approximately what will Cramer's *V* effect size need to be for .90 power?
19. You calculate a chi-square test of independence at $\alpha = .05$. Your study has 1 *df*, 200 total subjects, and effect size ϕ is .19. To the nearest 10%, what is the power of your study?

Looking Ahead

This module completes the instruction on hypothesis testing (Modules 10 to 33). You have learned how to word hypotheses, sample subjects, conceptualize sampling distributions, calculate tests of hypotheses, determine statistical significance, construct confidence intervals, determine effect size, and plan for and estimate power. In the remaining modules, we will turn our attention to correlation and prediction. The design and interpretation of correlational studies are quite different from the experimental studies with which we have worked thus far. With correlational studies, we are looking to establish only relationships, not causality. Instruction concludes with an introduction to multiple regression. This technique uses correlational (relational) statistics to analyze data from either correlational or experimental designs. When used for the latter, causation can again be inferred.

Visit the study site at www.sagepub.com/steinberg2e for practice quizzes and other study resources.