6 Inferential Statistics

Patricia Haden

6.1 Introduction

Descriptive statistical techniques (see Chapter 5) provide succinct and illuminating pictures of the data we record *from our own subjects.* They do not, on their own, tell us if we could expect to see similar patterns were we to apply our experimental manipulations to any other people. To make that logical leap – from our specific subjects to the larger world – we need a second group of statistical techniques, called **inferential statistics**. The inferential techniques allow us to make inferences beyond our particular **sample** to large (possibly infinite) **populations**. For example, we may find that our classroom of students learn complex algorithms more quickly when shown an animation. Inferential statistics will tell us if other instructors, working with different groups of students, should expect to see the same benefit.

The theory underlying inferential statistics is huge and complex. Obviously, we are only going to be able to cover a very small portion of that material in this chapter. Our focus will be on those elements that are essential for understanding and interpreting the existing literature and for supporting our own research practice. We will skip anything that is not essential to that focus. Most noticeably, we are going to ignore quite a lot of the elegant mathematics that underlie inferential statistical techniques. However, if you are interested in learning more about the (quite beautiful) math behind it all, see Gravetter and Wallnau (2012, chapters 7 and 8) and Miller and Haden (2006).

In this chapter, we present a specific inferential approach, called null hypothesis significance testing (NHST; Pearce, 1992). In the computing education literature, NHST is by far the most common inferential paradigm. However, you should be aware that there are other approaches to making inferences from samples to populations. Some authors, for example, advocate an alternative called **Bayesian analysis.** Simplifying greatly, NHST methods assume a single inferential hypothesis and determine, to some degree of certainty, whether the evidence is sufficient to falsify it. Bayesian methods assume multiple inferential hypotheses and determine the relative credibility of each. Details of Bayesian techniques can be found in, for example, Edwards, Lindman, and Savage (1963), Van de Shoot et al. (2014), and Wagenmakers (2007). For guidance on when to use each inferential framework, see Francis (2017). For an overview of the

https://doi.org/10.1017/9781108654555.007

sometimes spirited arguments surrounding these different approaches, see Nickerson (2000).

6.2 Inference

Statistical inference is the process of using what you can observe to make an educated guess about something you cannot observe. This type of inference is something we do all the time. For example, when you look out of the window in the morning to decide whether it's safe to hang out the laundry, you are making an inference. Specifically, you are using the appearance of the sky in the morning (what you can observe) to make an educated guess (an inference) about whether it will rain in the afternoon (something you cannot observe at the moment).

In research, we need to be able to infer beyond what we can see, because it is not possible to observe directly every person we are interested in. For example, imagine that a medical researcher is interested in comparing two treatments for asthma – one that is entirely medical and one that uses breath training, yoga, and other physical/behavioral interventions. She can give two groups of people these treatments and measure each subject's peak flow before and after treatment. She can compute the group means and see which treatment, on average, improves peak flow most. She will have learned which treatment is better, but only for those specific people. Of course, the researcher is not interested in just those people she has measured; she is interested in finding out which treatment works better for asthma patients in general – all asthma patients, all over the world. In fact, since a good asthma treatment will be used for years, she is interested in assessing the effectiveness of each treatment for patients who haven't even been diagnosed yet – for patients *who haven't even been born yet*. Obviously, one can never test all of those people directly.

Thus, in an experimental situation, there are actually two sets of people: The **sample** – the people we take measurements on – and the **population**, this infinite, unknowable, unmeasurable set of people that we want to learn something about. What inferential statistics lets us do is look at our sample, and make inferences about how our experimental manipulations would work on the population, if we were able to test everyone. And, as we shall see, we can make these inferences in a very precise way.

6.3 Hypothesis Testing

6.3.1 The Logic of Hypothesis Testing

The inferential statistical tests covered in this chapter all use the same pattern of inference. This technique is called **hypothesis testing.** When we talk about hypothesis testing in its formal, statistical context, it can seem confusing, but

https://doi.org/10.1017/9781108654555.007

as with predicting the weather from the morning sky, it is a chain of reasoning that we frequently follow in real life. Let's look first at a real-life example, and then see how it equates to the formal process used in a quantitative statistical analysis.

Meeting for Coffee

Imagine that your friend, Bob, was supposed to meet you for coffee at Starbucks at 10.00. It is now 10.20 and Bob has not arrived. Should you be worried? This is an inferential question. You are going to use something you know to make an inference about something you don't know.

What you know:

> Bob is 20 minutes late

What you want to infer:

> Is Bob just running late and will soon arrive, or has something happened to Bob that will prevent him from showing up?

To make this decision, you start with information you have about Bob. Specifically, how does Bob usually behave? Is Bob often late, or is he usually right on time? If Bob is often late, you will probably assume nothing unusual has happened and just continue to wait. But if Bob is usually on time, being 20 minutes late today is very unusual, and you might be concerned about Bob. This chain of reasoning is the basis of all inferential statistics. We can summarize the process as:

Observed data	Bob is 20 minutes late
Desired inference	Is everything normal, or has something unusual happened to Bob?
Required information	How punctual is Bob ordinarily, when nothing unusual has happened?
Decision	If being 20 minutes late is fairly typical for Bob, infer that probably nothing unusual has happened to Bob If being 20 minutes is late is not typical for Bob, infer that probably something unusual has happened to Bob

To see the logic even more clearly, imagine that Bob is always late and Zoe is always on time. If you have to wait 20 minutes for Bob, you wouldn't worry. You'd just say, "That Bob, he's never on time". But if Zoe was 20 minutes late, you would be concerned, and you would be texting her to find out if she was sick, had a flat battery in her car, etc. In each case, you are making an inference from what you know to what you don't know, based on *what is typical*.

The General Case

We can extend our "What's up with Bob?" pattern to any similar problem. In a quantitative study, the logic of an inferential hypothesis looks like this:

https://doi.org/10.1017/9781108654555.007

Observed data	What you know happened; the data you collected from your sample
Desired inference	Is something going on here? Specifically, is your treatment having an effect?
Required information	What would your data look like *if nothing was going on*? That is, what would you expect to observe if your treatment has no effect?
Decision	If your data look like what you would expect if nothing is going on, decide that nothing is going on If your data do not look like what you would expect if nothing is going on, decide that something is up

What Kind of Inference Can We Make?

Before we try to apply this logical pattern to a data analysis situation, we must first look at exactly what kinds of inferences the techniques allow us to draw.

In hypothesis testing situations, we make inferences about population averages.[1] When we say "men are taller than women," we really mean that the population average height of men is larger than the population average height of women. When we say "smoking causes cancer," we really mean that, on average, smoking raises the risk of getting cancer. When we say "the use of a visual programming language (VPL) improves CS1 retention rates," we really mean that the average retention rate for classes that use a VPL is higher than the average retention rate for classes that do not. This is the best you can do, and all inferential statistics deals with aggregated behavior.

Imagine that you wish to study the value of using pair programming (PP) with CS1 students. You wish to compare programming skill after completion of a CS1 course when PP is used to that when it is not. For your dependent variable (DV) you decide to use a standardized CS1 final exam. You want to answer the question: "Do students score better on the exam if they have been taught using PP?" What you are asking statistically is, "Is the *population average* score on the exam when PP is used greater than the *population average* score when it is not?"

We can illustrate the situation we are exploring as follows. Imagine that you were somehow magically able to test every student in the infinite, unknowable population of CS1 students, both with PP and without (and because this is magic, you don't have to worry about order effects). You could plot the two frequency distributions – one for scores with PP and one for scores without.[2] The result would look like one of the two examples in Figure 6.1.[3]

[1] Note there are also inferential tests to look at population proportions, correlations, etc., but the logic is the same, and it is easiest to start by concentrating on averages. The important thing is that we can only make inferences about population *summaries*, not about *individuals*.

[2] For simplicity, we will assume that the shape (i.e., variability and frequency) of the population distributions is the same. The major statistical tests are robust in the face of moderate violations of this assumption.

[3] Technically, there is a third option – students who use PP could actually do worse on the exam than those who do not. Inferential techniques work correctly in this case as well. To simplify this discussion, we will omit that possibility.

https://doi.org/10.1017/9781108654555.007

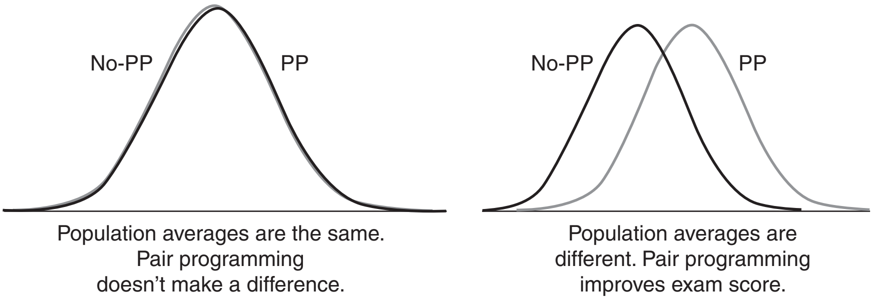


Figure 6.1 *Population frequency distributions showing no effect and a real effect of the independent variable.*

Reality can look one of these two ways: if using PP really doesn't make any difference, student performance will be the same with or without it, and the two distributions will lie on top of each other, as shown on the left. If using PP really does enhance learning, students will score better on the exam when PP is used, and the distributions will be separated, as shown on the right. The distance that the two distributions are separated is how much PP helps. When we do a hypothesis test, we are using our sample data to try to infer which of the two pictures represents the true state of the infinite, unknowable populations.

The Logic of Hypothesis Testing in an Experiment

We apply the same logic that helped us decide whether we needed to call Bob to the question of comparing these two population averages.

Observed data	Difference between average exam scores for PP and no-PP students, observed for our sample
Desired inference	Are the population averages really equal, or really different?
Required information	What would your data look like if nothing was going on? That is, PP had no impact and therefore the population averages were really equal (the image on the left in Figure 6.1)
Decision	If our observed average difference is something that *could easily happen* (due to sampling error) even if the population averages are equal, decide that the population averages probably are equal – or, more correctly, that you have no reason to assume otherwise If our observed average difference is something that *would be rare* when the population averages are really equal, decide that the population averages are probably not equal

Inferential statistical tests quantify very precisely those notions of "could easily happen" and "would be rare."

https://doi.org/10.1017/9781108654555.007

"Could Easily Happen" or "Would be Rare?"

To understand how inferential tests quantify the likelihood of our observed data, consider the following:

Assume that we have performed our PP vs. no-PP study using a between-subjects design (i.e., two separate groups of students; two separate CS1 classes). We have computed the average exam score for each of the two groups. Compare these two possible outcome scenarios:

Scenario 1:

Average no-PP exam score = 64
Average PP exam score = 65

Scenario 2:

Average no-PP exam score = 64
Average PP exam score = 83

In both scenarios, the observed sample means are different, and the mean score for students who used PP is higher. But which scenario would convince you, intuitively, that there is a real advantage to using PP? Presumably, Scenario 2, with its 19-point observed difference. In Scenario 1, the PP students did do better overall on the exam, but only by a very small amount. It is easy to imagine that, even if the population averages are truly equal (i.e., PP has no impact), you might observe a one-point difference between your group averages simply by chance. Your PP sample might have just contained a few more high-scoring students than did your non-PP sample. But a 19-point difference, as in the second scenario, seems much less likely to have occurred unless the population averages truly are different. Mathematically, the 19-point difference would be possible if PP had no effect and you just happened to have the bad luck to get two very non-representative samples, but you certainly wouldn't expect it to happen very often.

Inferential statistical tests use math to quantify this intuition exactly. They compute, to a very high degree of accuracy, how likely your observed average difference would be, *if the population averages were the same*.

The computation uses your observed means, your observed standard deviation, and your sample size. It determines the frequency distribution of all possible observed mean differences when the population means are really equal. It then computes the exact probability of a difference at least as large as your observed one in this distribution. If that probability is small (i.e., it wouldn't happen very often if the population means were equal), you will assume the population means are not equal. Conventionally in scientific disciplines, "small" is defined as 5 percent.[4] That is, if your observed data would occur 5 percent of

[4] Although it is still extremely common, there is a growing controversy around the rigid use of "5 percent = a small chance" in inferential tests. Many of the arguments are mathematically very complex, but a tractable overview of the issues can be found at www.nature.com/news/scientific-method-statistical-errors-1.14700

https://doi.org/10.1017/9781108654555.007

the time (or less) when the population averages are equal, you can assume that they are not.

Meeting for Coffee, Again

Now that we have considered the formalisms of hypothesis testing, let us once again see how it equates to our problem with Bob. We should be able to see that, while the context has become more complex, the important components are the same.

Hypothesis Testing Context	Meeting for Coffee Context
The frequency distribution of all possible observed mean differences when the population means are really equal	How late is Bob usually (i.e., what is his distribution of arrival times)?
The exact probability of a difference as large as your observed one in this frequency distribution	Exactly how often is Bob at least 20 minutes late?
If that probability is small, assume the population means are not equal	"Even Bob is rarely as much as 20 minutes late. I better call and make sure he's OK"

The Two Hypotheses

In both our PP study and the meeting-for-coffee scenario, there are two competing hypotheses, and we try to use our observed data to decide which hypothesis is correct. When waiting for Bob, the two hypotheses are (1) "Something has happened to Bob, and he is not going to show up," and (2) "Bob's just running late as usual, he'll be here soon." In the PP study, the two hypotheses are (1) "Use of PP in CS1 facilitates learning," and (2) "Use of PP in CS1 has no impact." In hypothesis testing, there are always two hypotheses. Informally, one hypothesis says "something interesting is going on here" and the other hypothesis says "nothing interesting is going on here." The "something interesting" is usually an effect of our independent variable (IV; i.e., the population means are different). The "nothing interesting" hypothesis states that our IV has no effect (i.e., the population means are not different). The "something interesting" hypothesis is called the **experimental hypothesis** and is denoted **H_1**. The "nothing interesting" hypothesis is called the **null hypothesis** and is denoted **H_0**.

H_1	H_0
The experimental hypothesis	The null hypothesis
The population means are different	The population means are not different
The population average exam score is higher with PP than without	The population average exam score is the same with PP as without

https://doi.org/10.1017/9781108654555.007

In the examples above, we asked ourselves, “How likely is it that our observed data would occur if nothing bad has happened to Bob?” and, “How likely is it that our observed data would occur if the PP and no-PP groups are really the same?” That is, in both cases, we asked ourselves, “How likely are our observed data *if the null hypothesis is true*?” This is the pattern for all inferential tests.[5] Based on the probability our test computes for us, we consider whether there is evidence that the null hypothesis is true. If we decide it is not (because our observed data would happen rarely when H_0 was true), we say that we **reject the null hypothesis**. Otherwise, we say that we **fail to reject the null hypothesis**.

There is a subtle, yet very important detail in the two conclusions above: when our observed means are very different, we conclude that H_0 is false. But, note very carefully that when our observed means are close to each other, *we do not conclude that H_0 is true*. (To do so is known as “accepting the null hypothesis,” and it is a serious statistical faux pas.) We may only *fail to conclude that it is false*. More specifically, we would say that “we have no evidence that H_0 is false.” We must take this conservative position because of sampling error. The population means could actually be different (i.e., H_0 could be false) and our observed means might have been close together simply because of the particular samples we drew. If we repeated the experiment, we might obtain very different results, with observed means that were far apart. Thus, we can only say that, based on our current data, we cannot reject the null hypothesis.

Often, we can think of H_0 as the opposite of H_1. If H_1 is “two conditions are different,” H_0 will be “two conditions are not different.” We must be careful, however, as null hypotheses are a special kind of opposite. They always describe the situation where nothing is happening. Consider the following examples, with special attention to the last two:

If H_1 is …	**… then H_0 is …**
The population means are different	The population means are the same
The population correlation is different from 0 (i.e., the two DVs are correlated)	The population correlation is 0 (i.e., the two DVs are not correlated)
In the population, Group A occurs more frequently than Group B	In the population, Groups A and B have equal frequency *NOT “Group B occurs more frequently than Group A.”* H_0 always says “nothing is happening” – no difference, no effect
The population mean for Condition 1 is larger than the population mean for Condition 2	The population means are the same *NOT “The population mean for Condition 1 is smaller than the population mean for Condition 2.”* H_0 always says “nothing is happening” – no difference, no effect

[5] More specifically, it is the pattern for all inferential tests under the NHST paradigm.

https://doi.org/10.1017/9781108654555.007

6.3.2 The Hypothesis Testing Procedure

With all the elements now defined, we can consider the formal description of a hypothesis testing situation:

1. State H_1 and H_0.
2. Collect your data.
3. Let your test determine how likely it is that you would have gotten your observed data if H_0 were true.
4. If it is very unlikely (usually defined as occurring 5 percent or less of the time), infer that H_0 is not true. Reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

Then What Happens …?

If You Reject H_0:

1. Your result is "statistically significant." Note that this is a formal statistical term that means only that your observed data would occur less than 5 percent of the time when H_0 is true. It *does not mean* that your result is important or interesting or any of the other meanings we give to the word "significant" in colloquial usage.
2. You say you reject H_0 "with 95 percent confidence" or "at the 95 percent confidence level" (see below for further explanation).
3. Since H_1 is the (special) opposite of H_0, if you reject H_0, you may infer that H_1 is true.[6]

If You Fail to Reject H_0:

1. Your result is not statistically significant.
2. Do not infer that H_0 is false. As discussed above, also *do not infer that H_0 is true*. Failing to reject H_0 simply means that you did not find evidence that H_0 was false. H_0 could still be false, and you didn't find it because of flaws in your experiment, unfortunate sampling error, a small population effect that is difficult to detect, or any of a number of other possible reasons. If you fail to reject H_0, *conclude only that there is not sufficient evidence to make any inference*.

6.3.3 The Big Problem with Hypothesis Testing

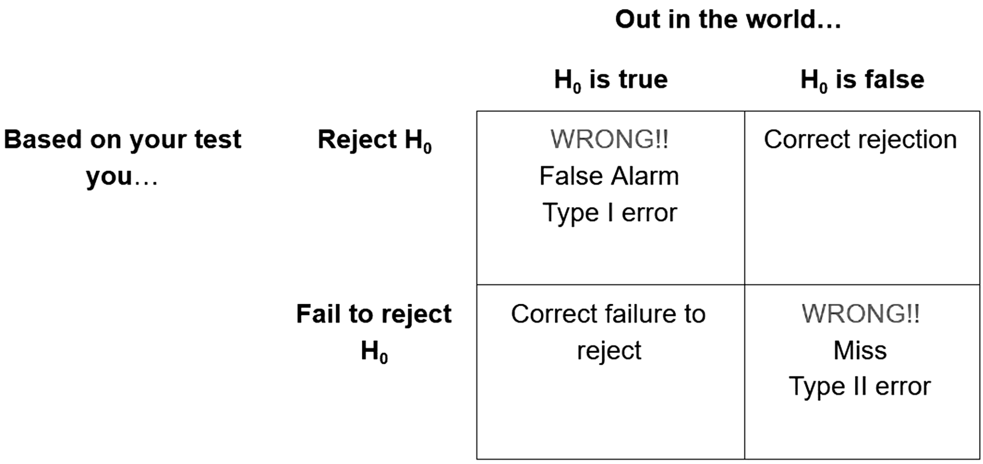
There is one serious problem with hypothesis testing: you could be wrong.

Recall that the logic is: If your observed result would occur less than 5 percent of the time when H_0 is true, infer that H_0 is false. This means:

1. Five percent of the observed outcomes that occur when H_0 is true will be identified as "unlikely."
2. When one of those outcomes occurs, H_0 will be rejected *even though it is true*.

[6] *Probably* true – technically, you have found evidence in support of H_1. See further discussion below.

https://doi.org/10.1017/9781108654555.007



		Out in the world…	
		H_0 is true	H_0 is false
Based on your test you…	Reject H_0	WRONG!! False Alarm Type I error	Correct rejection
	Fail to reject H_0	Correct failure to reject	WRONG!! Miss Type II error

Figure 6.2 *The hypothesis testing decision matrix.*

3. Therefore, 5 percent of the occasions when a true H_0 is tested, it will be rejected, and the wrong decision will have been made.

In fact, there are two ways that you can be wrong when you do a hypothesis test. Let's consider all the possible outcomes:

- In the real world, H_0 is either true or false.
- Based on the results of your hypothesis test, you will either reject H_0 or fail to reject H_0.
- Two of these decisions are correct, and two are incorrect, as shown in Figure 6.2.

This uncertainty is unavoidable. It is a consequence of not being able to measure everyone in the population. To understand the scientific literature as a whole, it is critical that you understand that there is always a chance of landing in one of the error boxes, so a proportion of published results are actually wrong. For some of the four possible outcomes shown in Figure 6.2, we can determine the odds very precisely; for others, we must estimate, but that is usually adequate to understand the results.

Consider the left-hand column of the matrix in Figure 6.2, which shows the two possibilities when H_0 is true. We know that we reject when we obtain a result that occurs no more than 5 percent of the time when H_0 is true. Therefore, decisions to reject will be wrong exactly 5 percent of the time (when the researcher happens to get one of those extreme rare scores that occur exactly 5 percent of the time). This error (rejecting when H_0 is true) is formally called a **Type I** error.[7] Informally it is called a **false alarm** because it causes the researcher to claim an effect of the IV when no such effect actually exists. This is an important thing

[7] This outcome is also called a **false positive**, but some people initially find this nomenclature confusing, in that you are claiming a *positive* outcome when you have *rejected*, which seems like a negative sort of thing to do.

https://doi.org/10.1017/9781108654555.007

to think about. For example, in the 1970s, there were hundreds of experiments performed looking for evidence of extrasensory perception (ESP). About 5 percent of them found such evidence (i.e., they rejected the null hypothesis "H_0: ESP does not exist"; they concluded that ESP does, in fact, exist). Those were the 5 percent of rare extreme observed scores that occur even when H_0 is true.

The probability of a Type I error (5 percent in our discussion so far) is also called **alpha**. Therefore, when you reject at 5 percent, you can say any of the following:

- Your data are significant at alpha = 0.05.
- Your Type I error probability is 5 percent.
- Your alpha is 5 percent.
- You have rejected with 95 percent certainty.
- Etc.

Next, consider the right-hand column of the matrix, where H_0 is really false out in the world. Failing to reject H_0 when it is actually false is formally called a **Type II** error.[8] Informally, it is called a **miss**, because there really is an effect of the IV, and your study missed it. The probability of a Type II error – formally known as **beta** – is complicated to compute, and depends on the sample size, the population variabilities, and the true population effect size. If you have very few subjects, you are more likely to miss a difference between population means; if your data are very noisy, you are more likely to miss any effect due to increased sampling error; if the effect of a treatment is small, you're more likely to miss it. The probability of a Type II error (beta) can also depend on the experimental setup. For example, if you use an insensitive instrument, you are more likely to miss a false H_0. The ability to detect an effect when present is called **power**. You will sometimes hear experiments criticized for having "insufficient power." This means that they have a low chance of correctly detecting a false H_0. Calculations can be done before you begin collecting data to estimate how much power your study will have. These computations, and their interpretation, are best done with the support of a statistician. For an overview, see Gravetter and Wallnau (2012, chapter 8).

The important thing to remember is that, after you do an inferential test, you can never tell with 100 percent certainty whether you have made an error or not. You can only know (or estimate) the probabilities. Understanding this will help you to be an informed statistical consumer and to interpret accurately the results of your own research.

6.4 Hypothesis Tests In Action

6.4.1 Basics

6.4.1.1 Which Test to Use?

There are literally hundreds of different inferential tests, and there are new ones being developed all the time for new experimental situations. It is extremely

[8] Also called a **false negative.**

https://doi.org/10.1017/9781108654555.007

important that the correct test is used for each inferential analysis. Later in this chapter, we will look in detail at the tests you are most likely to use and encounter in computer science education research.

6.4.1.2 Reporting Results Numerically

In a scientific paper, inferential test outcomes are generally described with two elements: the computed test result and the **p-value**. Some examples are:

$$\begin{aligned} t &= 6.5; \quad p < 0.05 \\ F_{3,11} &= 18.6; \quad p < 0.001 \\ \chi^2 &= 1.2; \quad p > 0.05 \\ r &= 0.86; \quad p < 0.01 \end{aligned}$$

Each test has its own name or symbol. The four results shown above are, respectively, a t-test, an analysis of variance (ANOVA; these tests both compare group means), a chi-squared (a test for population frequencies), and a test for correlation. We will discuss each of these techniques below.

In each case, the computed statistic (e.g., $t = 6.5$) is given first. Unless you have done a lot of statistics, these numbers are somewhat meaningless, as their interpretation depends on the test, the number of levels of the IV, and the sample size. However, the second value, the **p-value**, is extremely informative. The p-value for any inferential result tells you *exactly how frequently the observed data would occur if H_0 were true*. As discussed above, in much of the scientific literature, the cutoff for "rare enough to reject" is conventionally 5 percent. Thus, if the p-value is less than or equal to 5 percent, the computed statistic is considered unlikely to occur when H_0 is true and thus constitutes evidence that H_0 is not true. If $p < 0.05$, reject H_0. For the four results shown above, the t-test, the ANOVA, and the correlation test results are significant; the chi-squared is not.

6.4.1.3 Knowing the H_0 Distribution

Recall that all inferential tests follow the same logical pattern:

1. Collect your observed data.
2. Figure out how likely your observed data are, if H_0 is true.
3. If they are sufficiently unlikely, reject H_0.

The mathematical key to inferential testing is step 2 – being able to figure out how likely various outcomes are when H_0 is true. The distribution of all possible outcomes when H_0 is true must be calculable (by statisticians). Your statistical software looks at where your particular observed score falls in this distribution to determine the p-value.

For some inferential tests, this "distribution when H_0 is true" can be calculated exactly using the formula for the normal distribution and some fancy math. An alternative technique in modern statistics is to generate the H_0 distribution by computer simulation, running millions of artificial trials where values are sampled from the distribution described by H_0.

https://doi.org/10.1017/9781108654555.007

The different inferential tests have different computational formulae and therefore have different H_0 distributions. For example,[9] when performing a t-test, 4.00 would be a fairly large observed result, but when performing a multi-factor χ^2 test, 4.00 would be a fairly small result. It is not the absolute size of the computed statistic that matters; it is how likely it would be to occur, for the test you are doing, when H_0 is true.

Conveniently, when performing an inferential test for data analysis purposes, you don't need to worry about the whole H_0 distribution. Modern statistical analysis software will give you the p-value, which is the probability of your observed score in the H_0 distribution. (More formally, it is the proportion of the H_0 distribution composed of scores equal to, or more extreme than, your observed result.) If that probability is less than 5 percent, you reject. Thus, we can understand the output of the common inferential tests without having to worry about the derivations of their H_0 distributions.

6.4.2 The Common Inferential Tests

Although there are dozens of different inferential tests, you rarely see most of them in practice. In typical research studies in education, you are most likely to see one of the following:

Test	**Symbol for computed statistic**	**Used for …**	**H_0**
Two-sample t-test	*t*	Comparing exactly two groups with a between-subjects IV	The population means[10] are the same ($\mu_1 = \mu_2$)
Paired t-test (sometimes called the repeated measures t-test)	*t*	Testing difference scores (e.g., in a pre-test/post-test design). Used when you have a single, within-subjects IV with exactly two levels	The average population difference is 0 ($\mu_{\text{difference}} = 0$)
ANOVA[11]	*F*	Testing designs with more than two groups. This can be a single IV with three or more levels or multiple IVs in a factorial design	For each IV, the population averages for all levels are equal For all combinations of IVs, there are no interaction effects (see below for an explanation of interaction effects)

(*continued*)

[9] We will look at these computations in detail in just a minute.

[10] μ, the Greek letter "mu," is the symbol for a population mean. It is pronounced "myoo."

[11] The symbol *F* is used in honor of the inventor of ANOVA, Sir Ronald Fisher.

https://doi.org/10.1017/9781108654555.007

Test	Symbol for computed statistic	Used for …	H_0
Tests for correlation	r	Testing the correlation between two DVs For interval or ratio data, compute the Pearson product moment correlation. For ranks, compute the Spearman rank-order correlation	The population correlation[12] is 0 (ρ= 0)
Linear regression	F	Testing whether an outcome variable can be predicted from one or more measurement variables	The accuracy of the prediction is no better than chance
Chi-squared[13]	χ^2	Test for frequency (proportion) data	Proportions in the different conditions are equal and/or independent (discussed in detail below)

Note that in each case, H_0 says "there is no effect in the population." H_1, as we know, says that there is an effect. In each case, if you reject H_0, you can conclude, with known confidence, that H_1 is correct. For example, if you reject H_0 in a two-sample t-test, you have evidence that the population means of your two conditions are different. If you reject H_0 in a multiple regression experiment, you have evidence that the accuracy of a prediction made from your measurement variables is better than chance. If you reject H_0 in a correlation study, you have evidence that your two DVs really have some non-zero correlation in the population – and so on.

In the following sections, we will discuss each of these tests in detail. We will not focus on the underlying mathematics, which these days is handled by statistical software. Instead, we will concentrate on (1) understanding for which research context each test should be used and (2) how to interpret the results of each test. Throughout the discussion, note that although the different tests appear computationally very disparate, they all rest on the common logical framework of hypothesis testing described above.

[12] ρ, the Greek letter "rho," is the symbol for a population correlation. It is pronounced "row" – rhymes with "know."

[13] χ, the Greek letter "chi," in the name of this test is pronounced "kie" – rhymes with "pie."

https://doi.org/10.1017/9781108654555.007

6.4.3 Inferential Tests for Means

6.4.3.1 Comparing Groups

In this section, we will look in detail at the two t-tests and the ANOVA. These tests all make inferences about population means. They are all very common in the scientific literature. We will discuss how to choose which test is appropriate for your data set, how they are computed conceptually, and how the computed values are used to make precise decisions about H_0. As discussed above, these logical principles are essentially identical for all hypothesis tests. If you understand them in the context of the t-tests (where they are comparatively straightforward), you understand the fundamental concepts of all inferential analyses.

t-tests

The t-tests are used when you have *exactly* two groups or conditions – that is, one IV with two levels. Statistical software will accept the raw data values from the two groups and compute a t-observed and associated p-value. The t-observed is, essentially, the ratio of the difference between your groups to the total variability in the data.[14] It thus compares the difference you observed to the difference that could be expected by chance.[15] If that ratio is large, the difference between your groups is unlikely to be caused simply by data noise – it is more likely to reflect a difference in the underlying population means.

When you have a single *between-subjects* IV with exactly two levels, you should perform a **two-sample t-test**. This test compares the two group means. However, when you have a single *within-subjects* IV with exactly two levels, you should perform a **paired t-test**. When using a within-subjects IV, each subject is tested in both levels of the DV, so you can observe the effect of your manipulation by considering not the absolute performance at each level, but the difference between the two levels for each subject. In a paired t-test, the computation of t-observed is based on the subjects' difference scores. By using difference scores, you reduce, to a degree, the noise introduced into your data by variation from one subject to the next.

The H_0 Decision

Modern statistical analysis software such as SPSS will perform both types of t-test, as will mathematical scripting languages/environments such as R and MATLAB. It is only necessary then to determine if your computed t is "large." More formally, under the logic of hypothesis testing, we wish to determine *the probability that our observed t would have occurred if the population means are equal.* As discussed above, if that probability is small (usually under 5 percent), we reject H_0, the null hypothesis that states there is no difference between

[14] This should remind you of Cohen's *d*, and the two techniques are, in fact, closely related.

[15] For full computational details, see any introductory statistics text or the many available online resources, such as www.statisticshowto.com/probability-and-statistics/t-test/

https://doi.org/10.1017/9781108654555.007

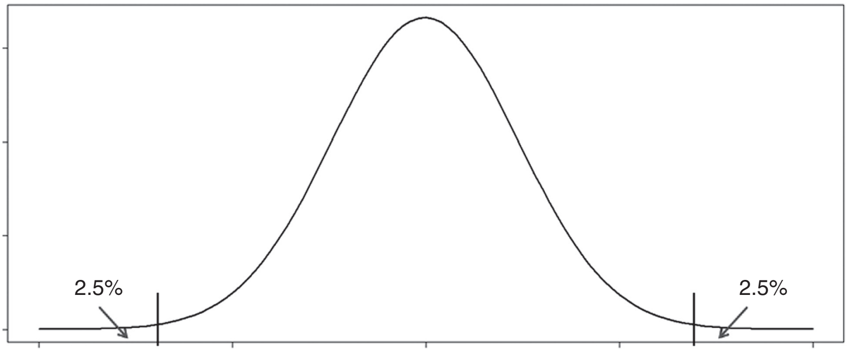


Figure 6.3 *Distribution of values of t for a given sample size and population variability when H_0 is true.*

population means, and we conclude, with 95 percent confidence, that the difference observed between our two sample groups is present in the population.

It is possible, using elegant mathematics, to determine the exact distribution of t-values that would be obtained when two population means are equal, for a specified sample size and population standard deviation. As discussed above, this distribution is used to determine exactly how big a t-value is needed to reject H_0. That value, known as **t-critical**, is the value that would be exceeded only 5 percent of the time when H_0 is true. This is illustrated in Figure 6.3, where the mean is 0 and the vertical lines indicate the locations of t-critical, both positive and negative. Values more extreme than t-critical occur only a total of 5 percent of the time.

Before analysis software became common, to determine whether a t-test was significant, you had to look up the t-critical value for your given sample size in a large table included in the back of every statistics textbook. If your t-observed exceeded[16] the value in the table, you rejected; if it did not, you failed to reject. Modern analysis software typically gives us more precise information than whether our observed *t* has exceeded t-critical. It usually tells us the exact p-value, the probability that our observed group difference would occur when H_0 is true. If, for example, we do a two-sample t-test and we see "*p* =0 .172," this means that our observed group difference would occur 17.2 percent of the time when the two population means are actually equal. We cannot be 95 percent confident that the population means are not equal (we can only be 82.8 percent confident), so we fail to reject H_0. If we see "$p < 0.01$," this means that our observed group mean difference would occur less than 1 percent of the time were the population means actually equal. We thus conclude that they are not the same. We conclude that the difference we observed in our samples would also be found if we were able to test every member of the infinite and unknowable population. We reject H_0 and we do so, in this case, with 99 percent confidence. Note that, in published papers, the *t* and *F* statistics are often

[16] "Exceeded" in this context means "is more extreme than." That is, it refers to the absolute value of your observed *t*.

https://doi.org/10.1017/9781108654555.007

given with numbers in round brackets. For example, you might see "$t(26) = 1.03$; $p = 0.312$" or "$F(1,34) = 8.53$; $p < 0.01$." The numbers in the brackets are related to the sample size and numbers of levels in the IVs. If you were doing the tests by hand, you would need those values to look up your p-values in the tables.

One-Tailed vs. Two-Tailed Tests

In Figure 6.3, we identify t-critical as the value which "cuts off" 5 percent of the t-distribution when H_0 is true. This 5 percent is accumulated by cutting off 2.5 percent from each end of the distribution. If, for a given sample size, t-critical is 1.98, that means that 2.5 percent of the t-values obtained when H_0 is true are greater than 1.98 and 2.5 percent are less than –1.98. If you obtain a t-observed of either 3.5 or –3.5, you reject. This makes sense if your experimental hypothesis H_1 is that there is *some difference* between the population means, but does not specify the direction of that difference. However, often H_1 is implicitly **directional**. For example, if you have introduced a classroom intervention that you anticipate will improve scores on the final exam, you expect the treatment group's mean to be higher than the control group's mean if the intervention works and to be no different from the control group's mean if the intervention is not effective. You have no theoretical motivation to expect that the treatment group's mean might be *worse* than the control group's. In the case of such a directional hypothesis, you want the t-test to compute the difference between your groups as *treatment group mean – control group mean*, and you are only interested in positive values of *t*. You might argue that your t-critical should, therefore, be the value that cuts 5 percent off the *positive* end of the t-distribution. This value will always be smaller than the t-critical that cuts 2.5 percent off each end, making it "easier" to reject. This procedure is called a **one-tailed test** and it is, in fact, a legitimate protocol in some cases. However, it is possible to get results that are significant by a one-tailed test and *not significant* by a two-tailed test (when your observed *t* is more extreme than the 5 percent cutoff value, but less extreme than the 2.5 percent cutoff value), which leads one-tailed tests to be viewed with some suspicion. For a one-tailed test to be mathematically and logically appropriate, a number of specific conditions must be met. Before deciding to do one-tailed tests, you may wish to seek statistical advice.

Interpretation

Modern analysis software has relieved us of the need to perform manual computations in most data analysis situations, but it is still our responsibility to correctly interpret the values the software gives us. As discussed above, if your observed *t* exceeds t-critical (or, equivalently, if your observed p-value is less than 0.05), you reject H_0. That is, you conclude with 95 percent confidence that the two population means are not equal. It is acceptable to assume that the direction of that difference (i.e., which population mean is the larger of the two) corresponds to the direction seen in your observed data. If your observed *t* is less extreme than t-critical (or, equivalently, your p-value is greater than 0.05), you fail to reject H_0. You do not, of course, conclude that H_0 is true. You

https://doi.org/10.1017/9781108654555.007

simply conclude that, in this particular situation, you have not found sufficient evidence to assume that the two population means are different.

ANOVA

ANOVA is one of the most common, and arguably the most important, of the inferential statistical techniques. It can be used to make inferences about population means for any number of groups and any number of IVs. It is also mathematically extremely complex. There are many college and university courses where one can spend an entire semester learning how to understand and perform the different kinds of ANOVA. We will concentrate on basic interpretation of the results of an ANOVA – identifying the null hypotheses and determining whether they can be rejected.

One IV: The One-Way ANOVA

When there is only one IV, the test performed is called a **one-way ANOVA**. If you have one IV and only two levels, you may perform a t-test, as above. However, if you have three or more levels, you *must* use the ANOVA. In the one-way ANOVA, the null hypothesis is that *the population averages for all levels of the IV are the same***.** If you reject this null hypothesis, you can infer only that there is some difference somewhere between the population averages. Further tests are required to formally compare specific pairs of conditions (see Gravetter & Wallnau, 2012, Chapter 6). However, in practice, it is common to assume that the pattern observed in the sample would also occur in the population, unless the observed effects are very small.

Multiple IVs: The Multi-Way ANOVA

Recall from the previous chapter that when we have multiple IVs, we usually test subjects in all possible combinations of levels, in a **factorial design**. In this situation, you must use an ANOVA – a t-test will not work. If you have two IVs, you perform a two-way ANOVA; if you have three IVs, you perform a three-way ANOVA, and so on.

A multi-way ANOVA simultaneously tests *multiple null hypotheses*. First, it tests for the effect of each IV (i.e., H_0: the population means of all levels are equal) on its own, collapsing across all levels of any other IVs. Imagine that you wish to determine the relative efficacy of lecture-based and laboratory practical-based instructions in CS1. You might also be interested in whether male and female students were differentially affected by this pedagogical difference. One approach would be a factorial design, as shown below, with a number of students in each of the four cells of the table. For each student, you could measure course grade or other appropriate performance metrics.

	Lecture-based course	**Practical-based course**
Women		
Men		

https://doi.org/10.1017/9781108654555.007

Because you have more than one IV and more than two groups (you have four), you must not use a t-test (or multiple t-tests[17]) on these data. You should perform an ANOVA. Because it has two IVs, this is a two-way ANOVA. The ANOVA tests multiple null hypotheses simultaneously while maintaining the correct overall probability of Type I error.

The two-way ANOVA for this design would test for a difference between pedagogies, collapsed across gender. It would also test for a difference between men and women, collapsed across pedagogy. These two null hypotheses would be formally stated as:

$$H_0\text{: } \mu_{\text{Lecture}} = \mu_{\text{Practical}}$$
$$H_0\text{: } \mu_{\text{Women}} = \mu_{\text{Men}}$$

Imagine that you were able to reject $H_0\text{: } \mu_{\text{Lecture}} = \mu_{\text{Practical}}$ (we will see in a moment how you decide when to reject). This would mean that you had evidence that the two teaching methods were different, *ignoring the effect of gender.* This would be logically equivalent to having simply compared the two teaching methods and not kept track of gender at all.

Imagine that, at the same time, you *failed* to reject $H_0\text{: } \mu_{\text{Women}} = \mu_{\text{Men}}$. This would mean that you found no evidence of a difference in the population averages for men and women *ignoring the effect of teaching method*. This would be logically equivalent to having compared men and women and not kept track of which pedagogical method they had experienced. These tests that consider a single IV while ignoring the other IV are called tests for **main effects**. If the results were as described above, you would say that you found a main effect of pedagogy, but no main effect of gender.

In addition, the two-way ANOVA tests for a new kind of effect called an **interaction**. An interaction is present when the effect of one IV is different for the different levels of the other IV. Continuing with our imaginary CS1 teaching experiment, we could collect our data and make a **factorial plot** – a graph showing the average performance for each of the four groups in the design (men in lectures, men in practicals, women in lectures, and women in practicals). The graph might look like that shown in Figure 6.4.

Figure 6.4 shows us that (in our sample) lectures are more effective for men while practicals are more effective for women. That is, the effect of the pedagogy IV is different for the different levels of the gender IV. This is an interaction.

Figure 6.5 shows another, more subtle possibility.

Figure 6.5 shows us that, for our participants, practical-based teaching is better than lecture-based teaching for both men and women, but the difference between the two treatments is much larger for women. For the men, it hardly matters which treatment you use; for the women, it matters a great deal. Again, the effect of the pedagogy IV is different for the different levels of the gender IV. There is an interaction.

[17] See the Section 6.5 for a discussion of why you must not do multiple t-tests.

https://doi.org/10.1017/9781108654555.007

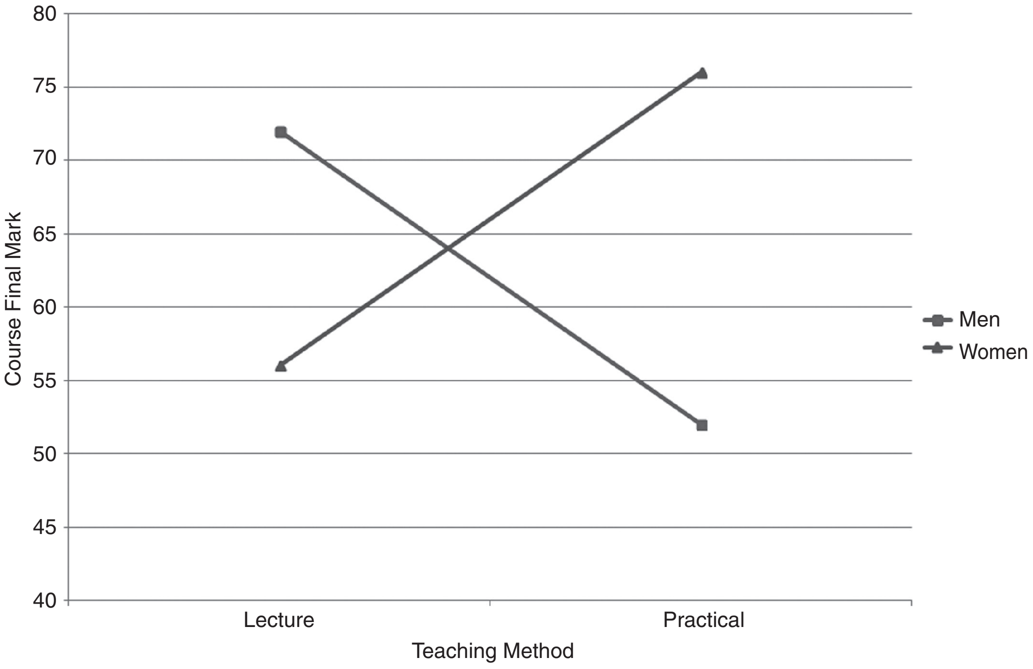


Figure 6.4 *Crossover interaction.*

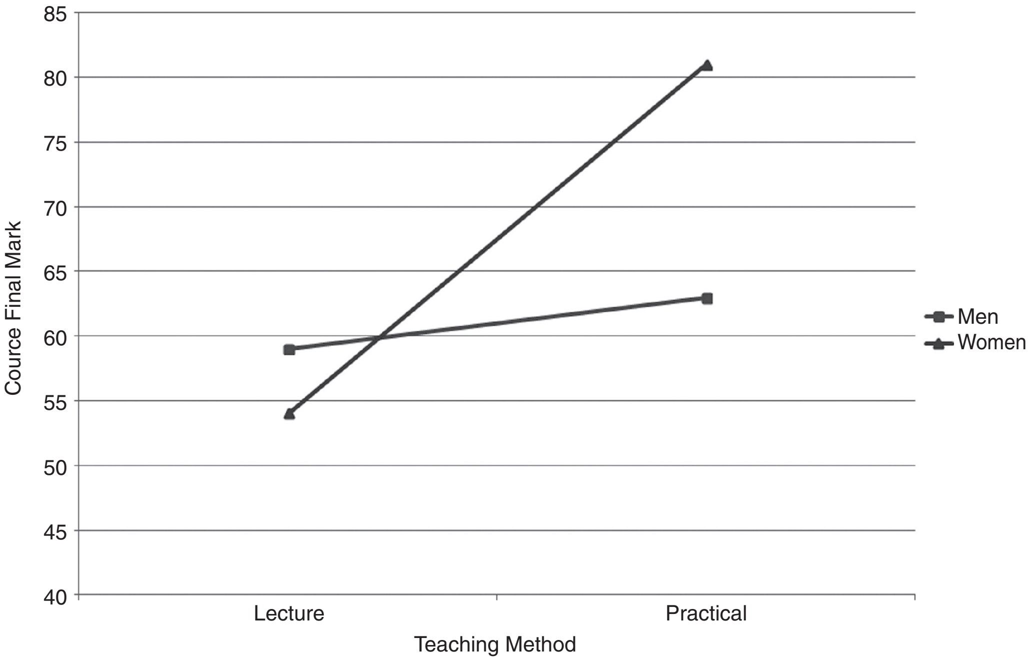


Figure 6.5 *Moderate interaction.*

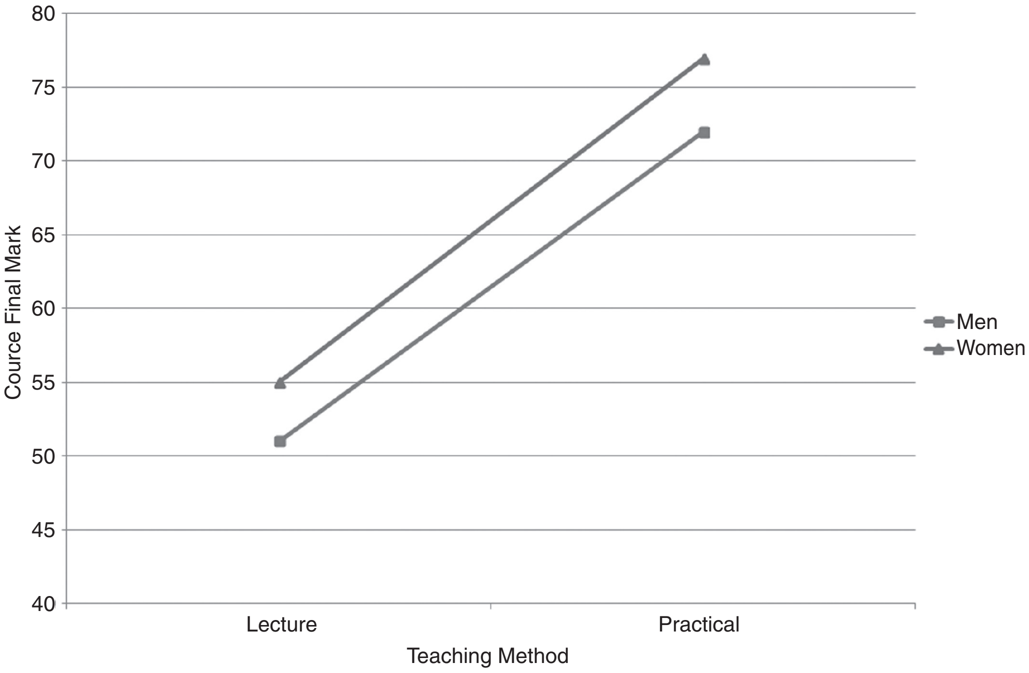


Figure 6.6 *No interaction.*

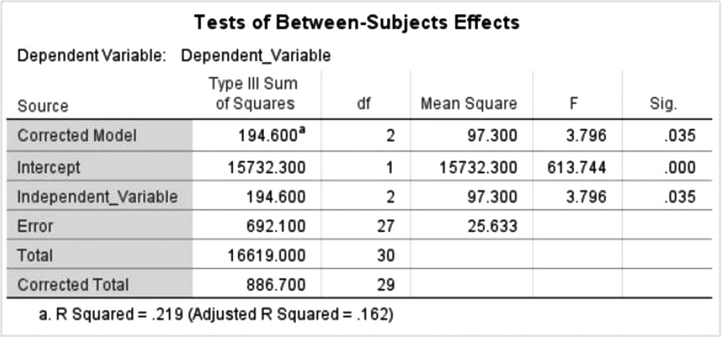
Finally, the factorial plot might have looked like Figure 6.6.

In this case, the mean for practical-based teaching is approximately 20 units larger than the mean for lecture-based teaching for both men and women. That is, the size and direction of the effect of the pedagogy IV is the same for both levels of the gender IV. There is therefore no interaction.

Comparing the three graphs, you can see that the more the two lines of a factorial plot deviate from parallel, the more extreme is the interaction. This makes it easy to get a sense of the presence of an interaction in your data simply by visual inspection of the factorial plot. But, as always, we need to do an inferential test to establish whether we would expect the pattern in our observed data to also be present in the population. When testing for interactions, the null hypothesis is "H_0: there is no interaction in the population." If we reject, we infer that there is an interaction in the population, and the pattern in the population is the same as the pattern seen in the observed group means.

With a two-way ANOVA, you test for two main effects and one interaction. By extension, if you have three IVs in a factorial design, you test for three main effects (IV_1, IV_2, and IV_3), three two-way interactions (each pairwise interaction of IV_1 by IV_2, IV_1 by IV_3, and IV_2 by IV_3, collapsed across the other IV), and a single three-way interaction that tests to see if all of the two-way interactions are the same at all levels of the third IV – and so on, for as many IVs as you have in your design. Unfortunately, these higher-order interactions are very delicate to interpret. So, while the ANOVA is quite happy to take dozens of IVs and test

https://doi.org/10.1017/9781108654555.007



Tests of Between-Subjects Effects

Dependent Variable: Dependent_Variable

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	194.600^a	2	97.300	3.796	.035
Intercept	15732.300	1	15732.300	613.744	.000
Independent_Variable	194.600	2	97.300	3.796	.035
Error	692.100	27	25.633		
Total	16619.000	30			
Corrected Total	886.700	29			

a. R Squared = .219 (Adjusted R Squared = .162)

Figure 6.7 *Generic one-way ANOVA table in SPSS 24.*

hundreds of simultaneous interactions, in practice, one rarely sees more than three IVs and a three-way ANOVA.

The ANOVA Table

To compute an ANOVA, you need special statistical software. (As of this writing, Excel can only do simple ANOVAs without a special statistics plug-in module.) The output of an ANOVA is not a single value as with a t-test, but a table of values. Different lines in the table correspond to different null hypotheses. As with the t-test, all you have to do is find the p-value for each null hypothesis to decide whether you can reject.

The One-Way ANOVA Table

The results of a one-way ANOVA will look something like Figure 6.7 (this is output from SPSS 24).[18]

The numbers of interest in this table are the *F* of 3.796 and the Sig. of 0.035 in the row headed "Independent_Variable." These are, respectively, the computed statistic *F* and the p-value for the main effect of your IV. Since the p-value is less than 0.05, you would reject the null hypothesis that the population means of all your levels are equal. You can look at your sample means to get an idea of which conditions are different, but as stated above, additional tests are required to make precise inferences about pairwise comparisons. The ANOVA formally only tells you that there is some difference somewhere.

The Two-Way ANOVA Table

A two-way ANOVA tests three null hypotheses simultaneously (the two main effects and the interaction). Therefore, the two-way ANOVA table has more rows – one row for each null hypothesis.

Consider the following (made-up) experiment: a researcher wishes to know if a person's interest in computer science depends on gender and current

[18] In an actual analysis, the name "Independent_Variable" would be changed to something descriptive.

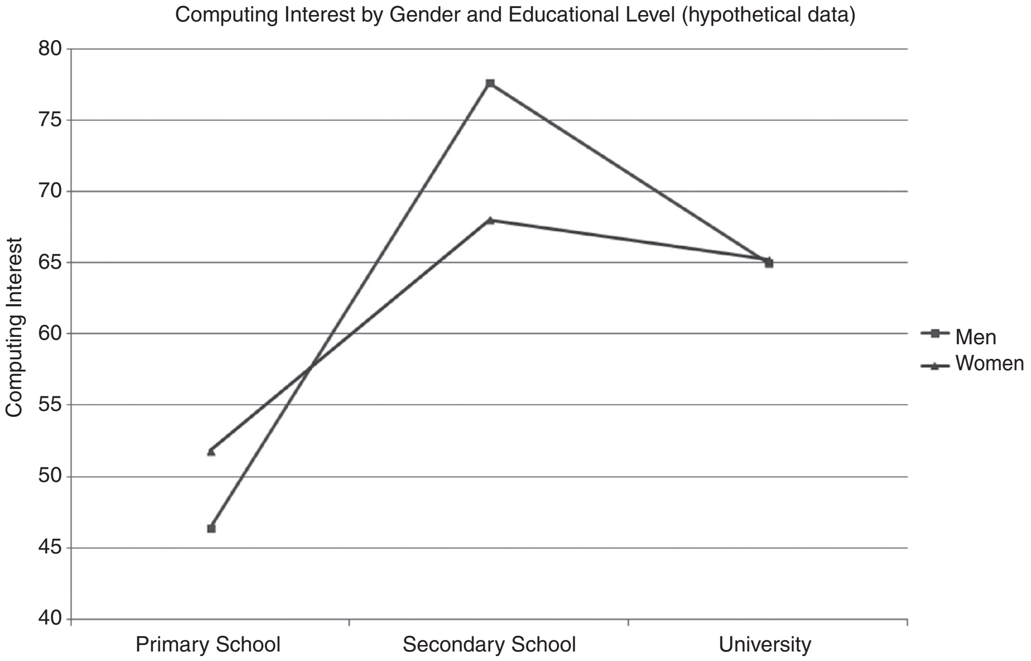


Figure 6.8 *Factorial plot for computing interest study (hypothetical).*

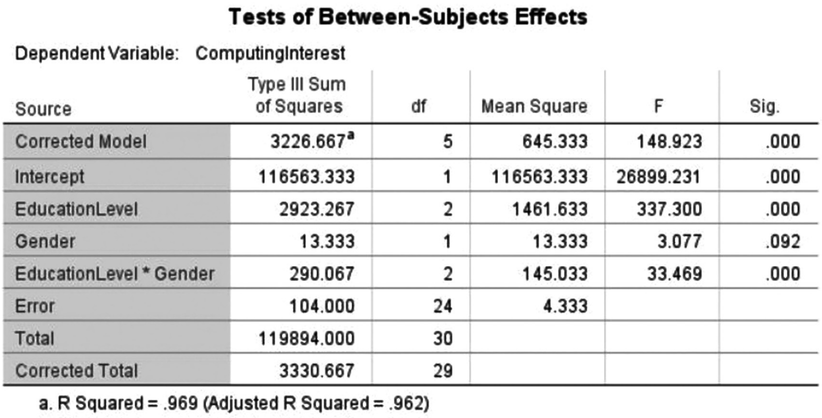
educational level. The study has two IVs: gender (levels = male and female) and educational level (levels = primary school, secondary school, and university). In a factorial design, there are thus six groups of subjects: male–primary, male–secondary, male–university, female–primary, female–secondary, and female–university. Assume that you have some good DV for measuring interest in computer science. Data are collected in all six groups and the group means are as shown in Figure 6.8.

In our observed data, it appears that interest in computing rises between primary and secondary school and then falls in university. This effect appears to be more extreme for men than for women. As always, however, we need to use inferential statistics to tell us how likely it is that we would observe such a pattern just by chance. Since we have two IVs, we perform a two-way ANOVA. The output from a recent version of SPSS for this data set is shown in Figure 6.9.

As with the one-way ANOVA, we can ignore much of this table, and simply look for the F and p-values for each null hypothesis. We see here one line for each of the three null hypotheses we test: the two main effects and the interaction.

The F for the main effect of EducationLevel is 337.3. The p-value is 0.000 (for $p < 0.0001$, SPSS gives up and calls it zero). Thus, we reject H_0: *the population mean computing interest scores for all education levels are the same*. We infer that interest in computer science is different for primary school, secondary school, and university students. We make this inference at very high confidence (99.999 percent). Informally, we can conclude that the pattern in the population

https://doi.org/10.1017/9781108654555.007



Tests of Between-Subjects Effects

Dependent Variable: ComputingInterest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3226.667[a]	5	645.333	148.923	.000
Intercept	116563.333	1	116563.333	26899.231	.000
EducationLevel	2923.267	2	1461.633	337.300	.000
Gender	13.333	1	13.333	3.077	.092
EducationLevel * Gender	290.067	2	145.033	33.469	.000
Error	104.000	24	4.333		
Total	119894.000	30			
Corrected Total	3330.667	29			

a. R Squared = .969 (Adjusted R Squared = .962)

Figure 6.9 *SPSS 24 output for two-way ANOVA.*

means is the same as the pattern in our sample means – interest rises from primary to secondary school, and then dips in university.

The *F* for the main effect of Gender is 3.077. The p-value is 0.092. Since this is greater than 0.05, we fail to reject H_0: *the population means for men and women are equal*. We have not found evidence that interest in computer science is different, on average, for men and women. This result is, at first glance, surprising. Statistically, we have concluded that there is no difference between males and females, even though it is obvious just by inspection that the line for males and the line for females in the factorial plot are different. In this situation, it is essential to remember that the null hypothesis for a main effect *collapses across all levels of other IVs.* To get a more complete picture, one must consider the role of the interaction.

The *F* for the interaction (EducationLevel × Gender) is 33.469. The p-value is 0.000 (i.e., $p < 0.001$). Thus, we reject H_0: *the effect of educational level on interest in computing is the same for men and for women*. Looking at the means, we can informally conclude that the jump in interest between primary school and secondary school is more extreme for men than it is for women. Or, considering the whole pattern, we could say that, in primary school, girls are more interested in computing than boys, this effect is reversed in secondary school, and by university, the interest levels of both genders are, on average, about the same. Note that, had we simply lumped all our subjects of all ages together and looked only at the effect of gender, we would have concluded there was no difference between males and females on this measure. The extended experimental design (taking education level into account) and analysis that includes the statistical interaction provide a much more nuanced, and more informative, insight into our research question.

Note that while we could probably have made these interpretations informally just by looking at the factorial plot, the ANOVA tells us that we can be confident that this pattern is not simply an accident in our sample, is not just

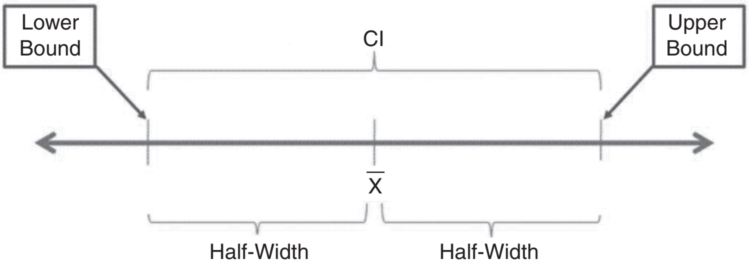


Figure 6.10 *CI for a population mean.*

due to random data variability, but would most likely be found if we could test everyone in the infinite, unknowable population.

6.4.3.2 *Inference via Parameter Estimation – Confidence Interval for μ*

The t-tests and ANOVA allow us to make inferences about population means based on observed data from samples, each a subset of their population. When making these inferences, we are distinguishing between a measure (in this case, an average) taken from a sample and the same measure taken from an infinite population. The former is formally called a ***statistic***; the latter is called a ***parameter***. The t-tests and ANOVA allow us to explore differences between population group means (parameters). There is also an alternative approach – the **confidence interval** (CI) – that allows us to estimate the value of a population parameter directly. CI analyses produce a range of values into which you can be 95 percent certain (assuming alpha is set to 0.05) the true population parameter value falls. A CI is centered around your observed sample statistic and extends symmetrically above and below that value, as shown in Figure 6.10.

The logic is that your sample statistic is your best single-value estimate of the true population parameter, and you then need to give yourself some room on either side of that estimate. The formula for the half-width of the CI incorporates the variability and size of your sample and the confidence level you wish to achieve. CIs are simple to compute[19] and can be performed in most statistical software.

CIs offer a useful descriptive analysis on their own, but they can also be used for formal hypothesis testing. Assume that you are implementing a nationalized computing exam for high school seniors and that your education department states that such exams must take, on average, no more than 60 minutes to complete. You fear that the exam you have written will take longer than that. You have H_1: $\mu > 60$ and H_0: $\mu = 60$ (or equivalently $\mu \leq 60$). If you give the exam to a large sample of high school seniors, you can record their completion times and compute a CI for μ. Assume that the resulting CI is (65, 71). You can therefore

[19] CIs are so easy you can compute them by hand, if necessary. See http://onlinestatbook.com/2/estimation/mean.html for an example.

https://doi.org/10.1017/9781108654555.007

be 95 percent confident that the true population mean (i.e., what you would get if you could test all of the high school seniors in your educational system) is between 65 and 71. Hence, you are 95 percent certain that it is *above* 60. You reject H_0 and conclude with 95 percent confidence that the exam is too long.

6.4.4 An Inferential Test for Frequencies

The two-sample *t*, the paired *t*, the ANOVA, and the CI for μ all allow inferences about population means. All are concerned with numerical data sets, where we compute group averages as measures of central tendency and make inferences from the observed group means to the corresponding population means. But not all quantitative research is interested in measures of central tendency – often we are interested in **frequency**. That is, we wish to know how often (or what proportion of the time) some particular condition or response occurs in our data. In these cases, we wish to infer not a population average, but a *population frequency***.** We wish to infer how often (or what proportion of the time) a condition or response would occur in our infinite, unknowable population.

6.4.4.1 The Frequency Fallacy

Caution must be used when interpreting frequencies, as it is easy to make misleading, yet accurate, statistical statements. For example, it is well known that more people die each year from beestings than from skydiving accidents. Certain people (primarily skydivers) cite this fact as evidence that going skydiving is safer than being stung by a bee. But while the bare statistical fact is true (it is – if you just count them up, more people die each year from beestings), the interpretation is not. The problem is that the *total number of beestings that occur each year is much higher than the total number of skydives*. Thus, even if an individual beesting and an individual skydive are equally risky, you will get a larger number of beesting deaths, simply because there are more beestings overall. In fact, you can get a larger number of beesting deaths even if skydiving is actually more dangerous per event, as long as there are sufficiently more beestings.

As another example, if you look at combat injuries among military personnel, there are invariably more men injured than women. Does this mean that men are more foolhardy and take greater risks? No, it simply reflects the fact that there are more men in combat, so naturally more men are injured. To quantify this, assume that among the military personnel you are studying, there are 90 percent men and 10 percent women. If you observe 100 injuries, how many women would you expect to be injured and how many men, assuming that men and women actually have the same chance of getting hurt? Because you have 90 percent men, you would expect 90 percent of your injuries to be men (i.e., 90 out of 100) and 10 percent of your injuries to be women (i.e., 10 out of 100). In absolute frequencies, you have many more injuries to men than to women (90 vs. 10), but that pattern is exactly what you would expect by chance when the real injury risk for men and women is the same.

https://doi.org/10.1017/9781108654555.007

Assume that your actual distribution of injuries was 80 men and 20 women. You still have many more injured men in absolute terms, but now it appears that it is actually women who are at greater risk of injury than men, because you have many more injured women (20) than you would expect by chance (10). You can see that it takes some statistical nous to be able to explain why, in this case, *20 is actually larger than 80.*

To avoid this confusion, frequency data are usually presented as percentages or proportions, rather than as absolute counts.[20] In our combat injury example above, 10 percent of our sample were women, but 20 percent of our injuries were to women. This causes us to suspect some relationship between risk and gender. As always, our next step is to determine whether this observed pattern in our sample is due to chance or if it would be found in the unmeasured population of all military personnel.

6.4.4.2 Inferential Test for Frequency Dependence – χ^2

If there is no patterned relationship between two factors (like injury risk and gender in our combat example), they are said to be **independent**. To make an inference about dependency from sample frequency data, you use a test called the **chi-squared**,[21] symbolized as χ^2, available in most statistical software. The χ^2 computes a measure based on the total difference between your expected frequencies and your observed frequencies. In the combat injuries example, since we have 90 percent men and 10 percent women in our sample, we *expect* 90 percent of the injuries to be to men and 10 percent to be to women. If we actually had 80 injuries to men and 20 to women, our observed frequencies (80 and 20 percent) would be different from our expected frequencies (90 and 10 percent), and the computed χ^2 would be greater than 0. Like all inferential tests, χ^2 gives us a p-value that tells us how likely our observed data are to occur when H_0 is true – that is, if our categories are truly independent in the population. If our data are unlikely to occur when H_0 is true ($p < 0.05$), we reject, and infer that the categories are not independent.

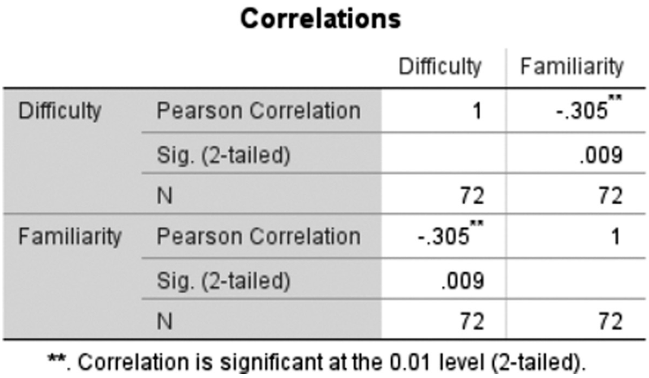
6.4.5 Inferential Tests for Correlation

In the previous chapter on descriptive statistics, we described the use of Pearson product moment correlations to quantify the relationship between two DVs. As with means and frequencies, we can use hypothesis testing to make inferences from our observed correlation r to the population correlation ρ. In a correlational context, the null hypothesis is that the two variables are uncorrelated, H_0: $\rho = 0$. Symmetrically, the experimental hypothesis is H_1: $\rho \neq 0$ (i.e., there is, in the population, some non-zero correlation between the two DVs). Following

[20] This assumes sufficiently large sample sizes. With small numbers of participants, percentages can be misleading. For clarity, authors may wish to present both absolute counts and percentages.

[21] Alternatively, “chi-square.” Both are acceptable.

https://doi.org/10.1017/9781108654555.007



Correlations

		Difficulty	Familiarity
Difficulty	Pearson Correlation	1	-.305**
	Sig. (2-tailed)		.009
	N	72	72
Familiarity	Pearson Correlation	-.305**	1
	Sig. (2-tailed)	.009	
	N	72	72

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 6.11 *SPSS 24 output for a significant Pearson product moment correlation.*

the now familiar logic of hypothesis testing, we need to determine the probability of obtaining our observed sample correlation *r* when H_0 is true. If that probability is small (once again, less than 0.05), we reject H_0 and conclude that we have evidence in support of H_1. When hypothesis testing for *r*, there is no additional value to compute (like a *t* or an *F*) – you just need r-critical, the value that cuts off 5 percent of the *r* distribution for your sample size when H_0 is true. You can obtain r-critical from tables in statistics books, and modern analysis software usually provides a p-value when you compute a sample correlation. An example from SPSS is shown in Figure 6.11, where the p-value is 0.009, as given in the row labeled "Sig. (2-tailed)." If the provided *p* is less than 0.05, you may reject the null hypothesis that your two DVs are uncorrelated in the population.

6.4.6 Inferential Tests for Prediction

We have looked at various inferential tests for differences between groups. We can use these techniques to see how a particular teaching method works *on average*. However, in some situations, we may want to know not only how a method works on average, but *for whom it will work best*.

Imagine that you had found a teaching technique that improved performance on average, but there was a great deal of variability in outcome. That is, it worked for some students, but not for others. Studies have found, for example, that PP raises lab completion rates for some students, but lowers them for others (e.g., Wood et al., 2013). You would like to know exactly what properties of the students or their situation determined whether the technique had worked for them. You could then use this knowledge to decide in advance if the method was likely to work for a given student. Especially if the method was costly or difficult to use, you would like to be able to *predict the outcome* based on your knowledge of individual students.

To allow prediction, we use a multivariable experimental methodology. As always, we have a primary outcome measure – our DV – whose behavior

https://doi.org/10.1017/9781108654555.007

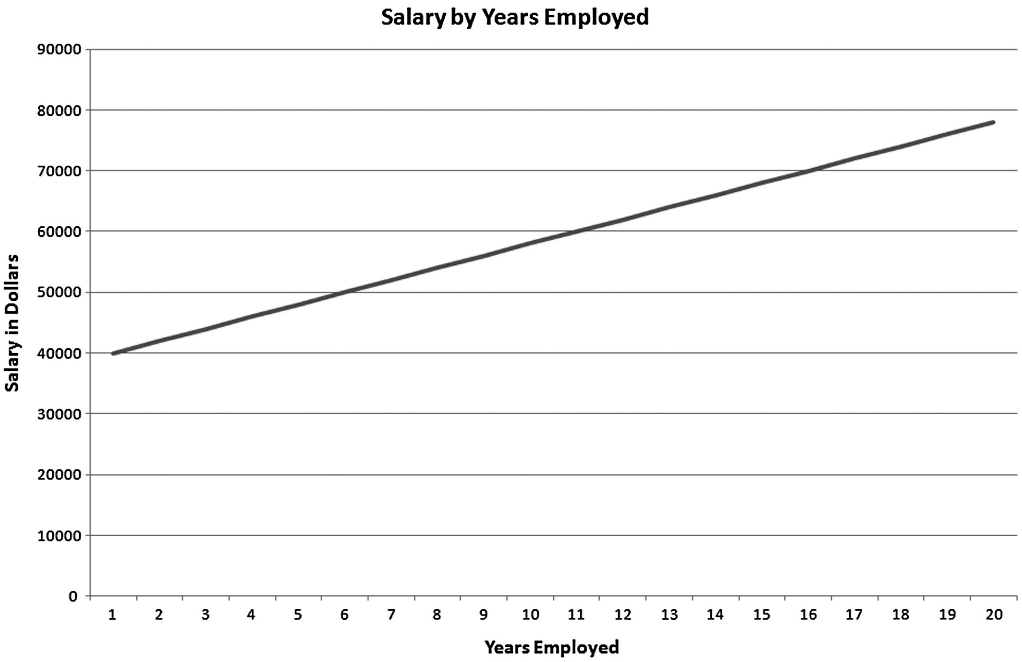


Figure 6.12 *Salary by years employed (hypothetical).*

we wish to understand. But instead of grouping our subjects according to a categorical IV, we take one or more other measures for each subject (as when exploring the correlation between two DVs). These measures are things that we have reason to believe may be related to how well our experimental treatment works. They are called **predictor variables**. We then use statistical techniques to try to predict, from the values of the predictors, the value of the outcome variable. The statistical techniques compute an equation we can use in the future. We measure the predictor variables on a new subject, plug the values into our equation, and compute an expected value on our primary outcome measure.

This predictive statistical technique is called **regression**. There are various flavors of regression for different data situations, but they all work to predict an outcome variable from a set of one or more predictor variables. The most common form of regression is **linear regression**, in which outcome values are predicted to fall approximately on a line. We will first consider this notion of "falling on a line," then we will look at how to use linear regression.

Imagine that you have taken a job teaching programming at a local polytechnic. At this school, teachers earn a starting annual salary of $40,000. They get a raise of $2,000 for each year they are employed. Thus, after one year, you would be earning $42,000, after two years, $44,000, after 5 years, $50,000, and so on. We could make a graph showing the relationship between salary and years employed. It would look like Figure 6.12.

https://doi.org/10.1017/9781108654555.007

We would describe the way salary is determined as “\$40,000 plus \$2,000 for every year you have been employed.” If we wrote this in mathematical notation, it would be:

$$\text{Salary} = \$40{,}000 + (\$2{,}000 \times \text{Years Employed})$$

The above expression is an example of the algebraic formula for a straight line $y = a + bx$. The simplest linear regression, where you use one numeric predictor variable to predict one numeric outcome variable, produces an equation of this form; x is the predictor, y is the variable to be predicted, and a and b are the intercept and slope of the line, respectively. The slope and intercept values define the single line *that best fits our observed data*. We give the analysis program all of our x and y pairs and it produces the values for a and b. This equation is called a **regression model** because it describes (models) the mathematical relationship between predictor and predicted.

6.4.6.1 Simple Linear Regression

In the example above, we could predict total salary *exactly* knowing only the number of years of employment. Life is rarely ever this tidy. Usually, predictor and predicted show a trend, not a perfect straight-line relationship. Linear regression uses math to find the straight line that is closest to our real data. It also tells us how close our data are to that line. When we use the equation to make a prediction, we can then gain a sense of how accurate the prediction is likely to be.

In the previous chapter, in the context of normal scores, we discussed the need to discover determinants (predictors) of eventual success in programming courses. This type of research situation – where you wish to predict one DV from another – is an appropriate problem for a regression approach. Imagine you were continuing to explore the relationship between math ability and programming ability. One approach would be to collect, from a number of students, their final mark in a math course (your predictor variable) and their final mark in CS1 (what you wish to be able to predict). Since you have two scores for each subject (math mark and CS1 mark), you can make a scatterplot. Your results might look like those shown in Figure 6.13.

While the data points don’t all lie perfectly on a straight line, you can see that there is a definite trend in the upward direction (i.e., a positive correlation). Students who performed well in math tended to perform well in CS1. A linear regression on these data (which can be done in SPSS, MATLAB, etc.) would give you the following linear formula:

$$\text{Predicted CS1 Mark} = 11.16 + (0.86 \times \text{Math Mark})$$

Thus, for a student who got a mark of 57 in math, the predicted CS1 mark is $11.16 + (0.86 \times 57) = 60.12$. For a student with a math mark of 94, the predicted CS1 mark is $11.16 + (0.86 \times 94) = 91.9$. You could compute the predicted CS1 for any student for whom you had a math mark simply by plugging that value into the equation.

The linear model (the equation) produced by simple linear regression is known as the **line of best fit**. It describes the linear set of points for which the difference

https://doi.org/10.1017/9781108654555.007

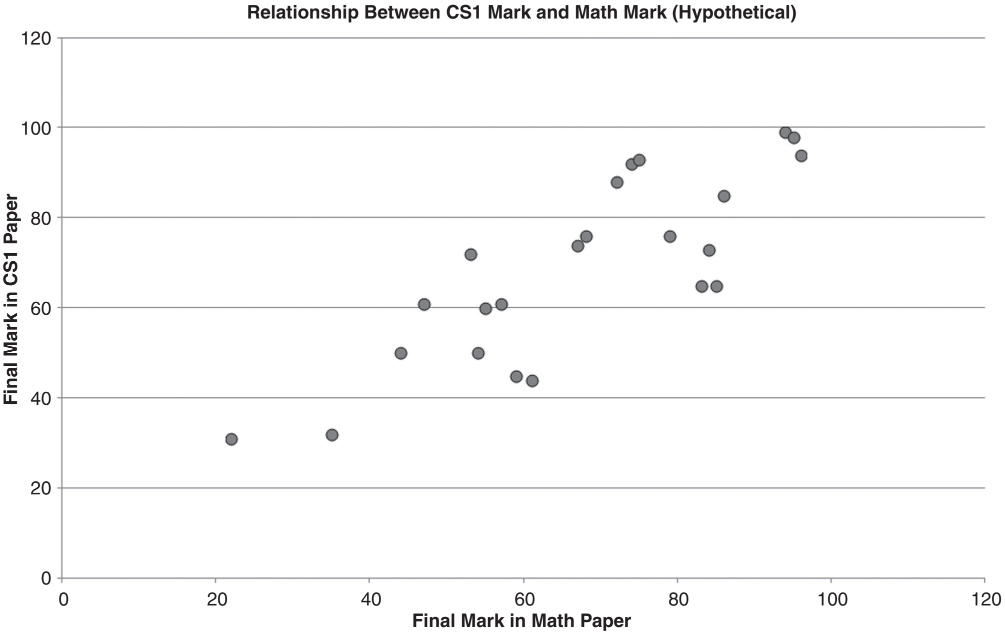


Figure 6.13 *Outcome of a regression study predicting CS1 mark from math mark (hypothetical).*

between the actual dependent scores in our data set (the variable we are trying to predict) and those values produced by the equation is minimized. Most analysis programs can display the line of best fit on the scatter plot. For our hypothetical data set, Figure 6.14 shows the line of best fit as produced in Excel.

The real data points don't fall exactly on the line, but the line does run through the center of the points. The math behind regression ensures that this is precisely the line that is closest *to all of the points as a group*. It is thus the most accurate linear formula to use for prediction.

6.4.6.2 Regression and Inference

The regression equation is computed based on your observed data – the predictor and predicted values you have collected. As such, like any descriptive statistic, it can tell you only about the predictive relationship *in your sample*. Fortunately, regression analysis also performs an inferential test and computes a p-value. The null hypothesis is that the true slope of the regression line (if you could gather *x* and *y* values from every member of the population) is 0. A slope of 0 would mean that the line of best fit was flat, indicating that changes on the x-axis (the predictor variable) had no consistent relationship (either increasing or decreasing) with changes on the y-axis. Intuitively, you can view H_0 for regression as stating that "this equation predicts the outcome variable no better than chance" (i.e., random guessing). If your p-value allows you to

https://doi.org/10.1017/9781108654555.007

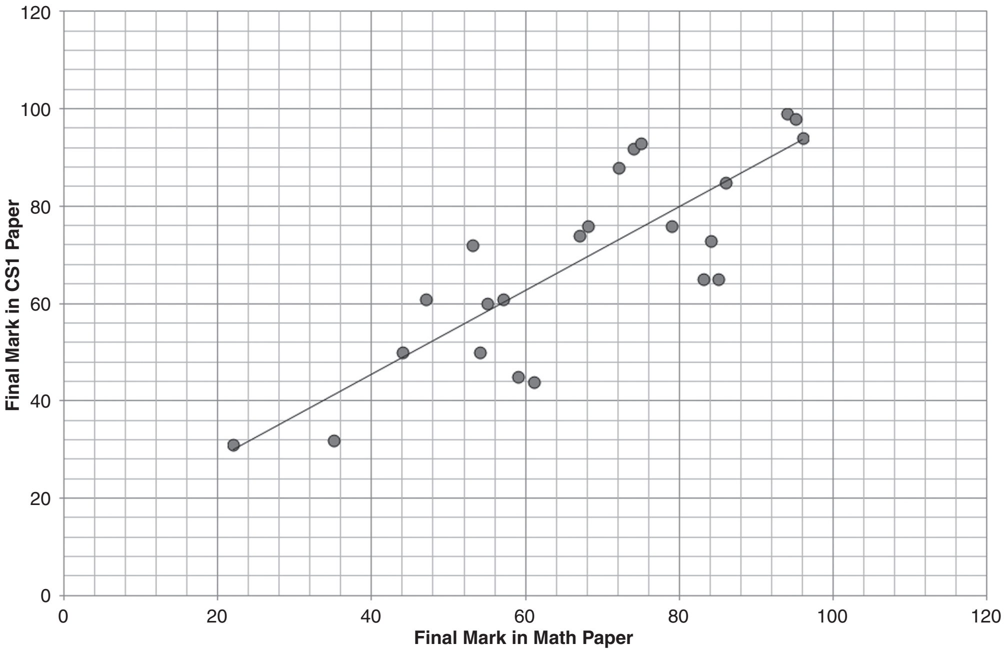


Figure 6.14 *Linear regression (hypothetical data) with line of best fit in Excel.*

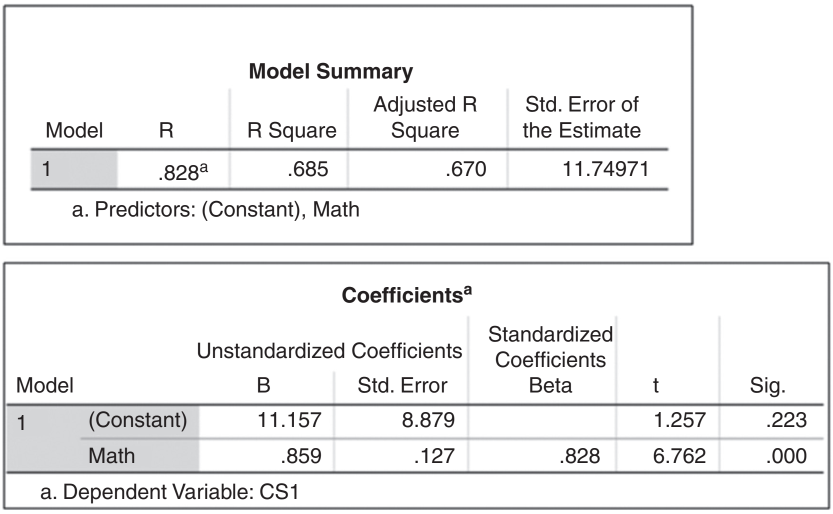
reject H_0 (i.e., it is less than 0.05), you can assume that your regression equation predicts better than chance. You can get a sense of how accurate your prediction is likely to be by looking at the scatter plot. Roughly, the closer your data points are to the line of best fit, the more accurate your prediction. For more precision, a regression analysis will return a value R^2, the **explained variance**. The underlying mathematical logic is complex, but intuitively, R^2 measures the extent to which the differences between subjects' outcome measure scores can be explained by looking at their predictor variable scores. Values range between 0 and 1, and the larger R^2 is, the more accurate are your regression equation's predictions.

Figure 6.15 shows the output of linear regression for the hypothetical "maths predicting CS1" data set shown above using SPSS 24.

The upper table in Figure 6.15 shows R^2 as discussed. The lower table shows the intercept (column B, row Constant) and slope (column B, row Math) of the regression equation. The "Sig." column displays the p-values. The p-value for the Math predictor variable is shown as 0 (it is, of course, not really 0, but it is a value that is too small to be displayed in SPSS's three significant digits), allowing us to reject H_0. We can thus conclude that math mark predicts CS1 mark better than chance.

Naturally, not all measures are useful predictors. Imagine that our math and CS1 scores were as shown in Figure 6.16 (the CS1 scores are the same as in the earlier dataset; the math scores have been changed).

https://doi.org/10.1017/9781108654555.007



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.828[a]	.685	.670	11.74971

a. Predictors: (Constant), Math

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.157	8.879		1.257	.223
	Math	.859	.127	.828	6.762	.000

a. Dependent Variable: CS1

Figure 6.15 *Output of linear regression analysis in SPSS 24.*

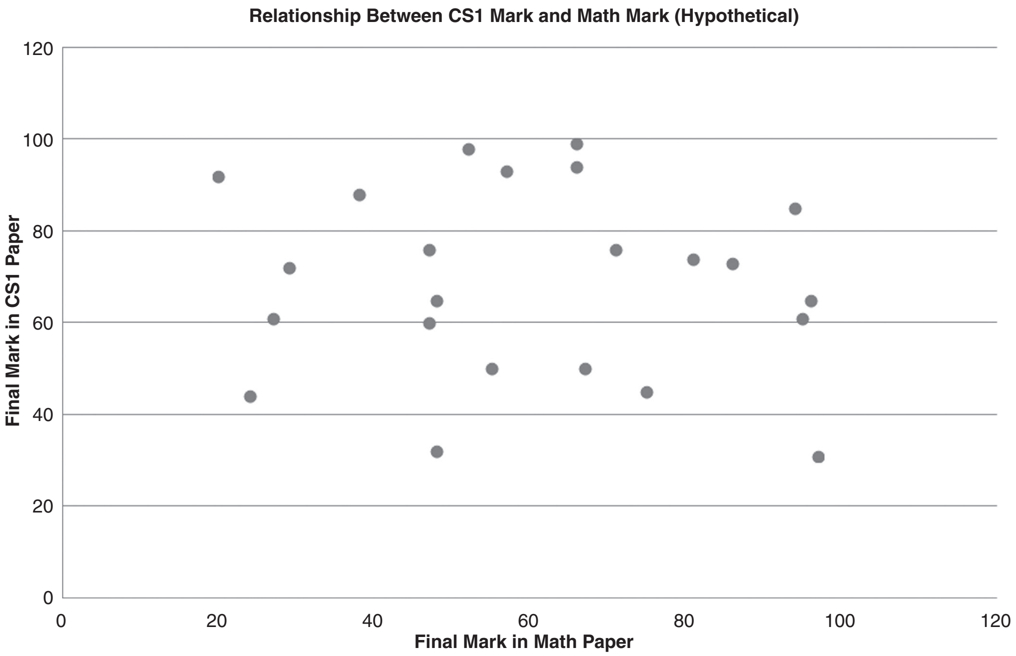


Figure 6.16 *A poor predictor.*

By eye, the relationship certainly looks less predictive than it did before. This intuition is supported by inspecting the Pearson correlations of the two versions (0.828 for the first example data set and –0.103 for the second). It is also easy to see that the slope of the line of best fit (in Figure 6.17) is much closer to 0 (i.e., a flat line) for the second set of values than it was for the first.

https://doi.org/10.1017/9781108654555.007

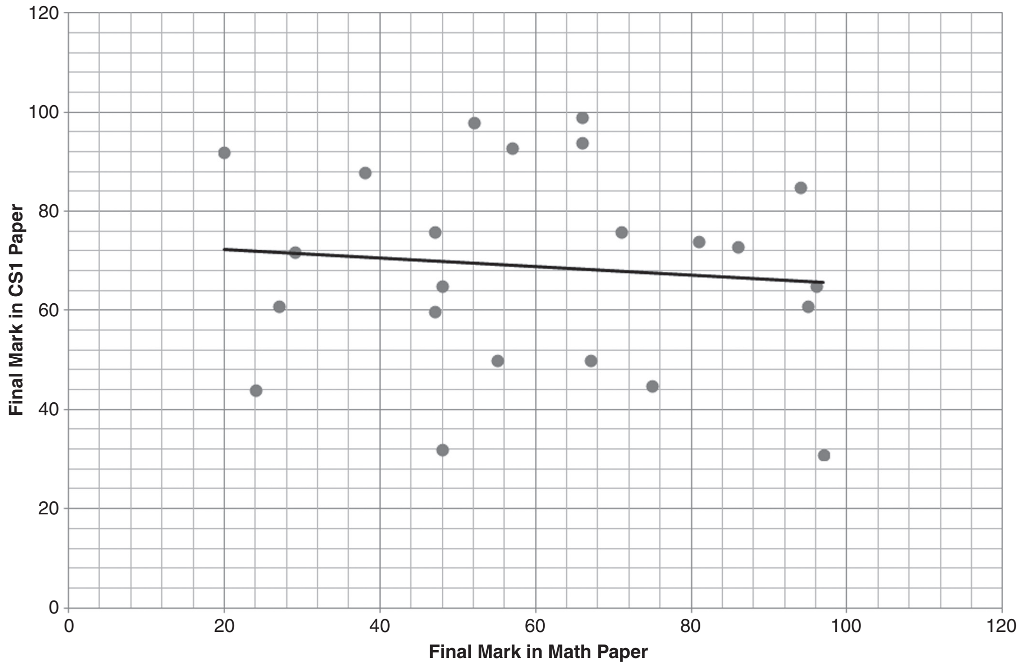
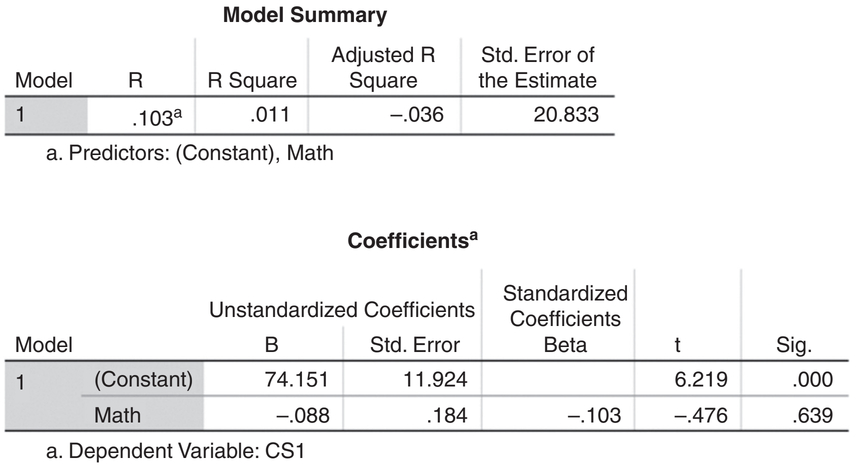


Figure 6.17 *Line of best fit for a poor predictor.*



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.103[a]	.011	–.036	20.833

a. Predictors: (Constant), Math

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	74.151	11.924		6.219	.000
	Math	–.088	.184	–.103	–.476	.639

a. Dependent Variable: CS1

Figure 6.18 *Linear regression analysis for a poor predictor in SPSS 24.*

The output of the linear regression analysis for this second data set is shown in Figure 6.18.

The R^2 is small (0.011), indicating that the values predicted by the best linear model for these data are not close to the actual data values. The p-value is large (0.639), so we fail to reject H_0. Logically, we have no evidence from these data to conclude that math mark can be used to accurately predict CS1 mark.

6.4.6.3 Other Kinds of Regression

There are three common variations on the simple linear regression discussed above.

Multiple Linear Regression

It is possible to collect multiple predictor variables and use them simultaneously to predict your outcome measure. For example, in our math and CS1 mark example above, you might also have measured each subject's verbal skill, problem-solving skill, age, etc. All of these predictors are submitted to the statistical test, and you get a single equation of the form:

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

The x_i are your different predictor variables; y is the outcome variable you wish to predict. The equation contains a coefficient (the corresponding b_i) for each predictor. If you have a new student, you can measure all of the predictors (math mark, verbal skill mark, problem-solving skill, etc.) and plug them into the equation to get a predicted CS1 mark. Multiple regression analysis usually gives a p-value for each individual predictor and a p-value for the best possible (most accurate) model that can be constructed from all the predictors you provide. The mathematics behind multiple regression is very beautiful (see Miller & Haden, 2006, for details). Not surprisingly, a multiple regression equation is often more accurate than a simple regression equation with only one predictor variable.

Logistic Regression

It is also possible to predict outcome measures that are categorical rather than numerical. For example, you might want to use math mark to predict whether students pass or fail CS1. Linear regression can only be used to predict numerical DVs, not categorical ones like pass/fail. The technique for predicting categorical outcomes is called logistic regression. Logistic regression analyses allow you to predict not a numerical value, but the *likelihood* that an individual, based on the values of his or her predictor variable(s), will end up in each category of the outcome measure. The interpretation of a linear regression analysis involves principles of mathematical probability and is quite subtle. It is easy to make technical errors when interpreting logistic regression. It is recommended, therefore, that if you choose to use logistic regression, you consult a statistician for assistance.

Hierarchical Linear Modeling

In simple linear regression, all data points are treated as independent of each other. This makes it impossible to capture the impact of natural nestings and groupings in the data, such as "all the students in the same classroom" or "all the colleges in the same country." To perform regression that maintains such structures, one can use hierarchical linear modeling. Again, it is best to consult a statistician in this case.

https://doi.org/10.1017/9781108654555.007

6.4.7 Parametric vs. Nonparametric Tests

When you are reading the scientific literature, you will often see χ^2 described as a "nonparametric" test. There are two general classes of statistical test – parametric and nonparametric. The parametric tests are those that operate accurately only on ratio or interval, approximately normally distributed data.[22] If you use them on data that do not fulfill these criteria, they don't work correctly (specifically, your Type I error probability won't be accurate). The t-tests and ANOVA are parametric tests – they require numeric, approximately normally distributed data. Tests that are designed for use on categorical data, frequency data, non-normal data, etc., are called nonparametric. There are many nonparametric tests, each specific to a particular kind of non-numeric or non-normal analysis situation. There is no sense in which either category of test – parametric or nonparametric – is superior to the other; it is simply important to always use a test that is appropriate for your data set. For a detailed discussion of non-parametric methods commonly used in education and the social sciences, see Corder and Foreman (2014).

6.5 Common Mistakes (and How to Avoid Them)

Statistical analysis, especially if you own a copy of SPSS or Excel, is deceptively easy to do. Fling in a bunch of numbers, click a few menus, get a p-value. Unfortunately, the statistical process is filled with snares and traps for the unwary. Statistical errors can occur in all aspects of design, presentation, analysis, and interpretation. Detailing all of the ways that stats can be misused (both inadvertently and maliciously) is beyond the scope of this chapter. Interested readers are directed to Campbell (1974), a classic text on the subject, which is both informative and entertaining, or to Smith (2012) for a more technical presentation.

We have already discussed some of these errors. There are errors of design: choosing invalid, unreliable, or biased DVs dooms your search for knowledge from the start; omitting a control group obscures true effect sizes; failing to counterbalance can introduce subtle order effects. There are errors of logic: accepting the null hypothesis and inappropriate causal inference can both cause a researcher to draw conclusions from his or her data that are not logically sound. There are formal mechanical errors: for example, using a parametric test on extremely non-normal data. In this section, we discuss a few slightly more complex, yet common mistakes that researchers, and readers, should watch out for.

6.5.1 Alpha Bloat

Recall that when we do a hypothesis test, we choose a confidence/significance level (alpha), typically of 5 percent. We reject when $p < 0.05$; we are 95 percent certain

[22] There are formal tests for normality, but they are beyond the scope of this discussion. If your frequency distribution looks distinctly non-normal, consult a statistician to see if you need to use a nonparametric test.

https://doi.org/10.1017/9781108654555.007

that our decision to reject is correct, and so on. The underlying mathematics of the tests ensures that these probabilistic statements are true, and this certainty allows us to judge accurately the complete body of literature in our scientific area.

But there is a problem. When you do a single test on your data, and reject, you have a 5 percent chance of being wrong. If you do a second test *on those same data* and reject, you have *another* 5 percent chance of being wrong. If you do a third test and reject, you have *yet another* 5 percent chance of being wrong. Because of the complexity of probabilities, these 5 percent chances don't necessarily simply add up, but they do accumulate. If you do multiple tests on the same set of data, your overall alpha (your Type I error probability) is not really 5 percent; it is something larger (much larger, if you do a lot of tests) than 5 percent. Thus, we say that your alpha is "bloated."

Unfortunately, we often want to do multiple inferential tests on a single set of data. We have multiple DVs and we want to test all the pairwise correlations between them. We have multiple questions on a survey and we want to compare group means on each one to see on which questions our groups differ – and so on. Our goal is to be able to perform all of these inferential tests while maintaining a *total* Type I error probability, across all our tests, of 5 percent. To do this, we must reduce the critical p-value for each *individual* test. That is, we must require a smaller p-value (and hence a larger observed effect size) to reject any individual null hypothesis. The most straightforward technique for determining what this "smaller" critical value for *p* should be is the Bonferroni correction (McDonald, 2015). The Bonferroni correction says that, if you want alpha = *a* and you want to do *n* tests, you can only reject p-values smaller than *a* / *n*. For example, if you are doing ten tests on the same data set, and you want to maintain a total Type I error probability of 5 percent, your critical p-value for each individual test is 0.05 / 10 = 0.005. That is, you can only reject H_0 for tests that return a p-value less than or equal to 0.005. Obviously, you will need much larger observed effect sizes in order to reject. Equally obviously, making rejection more difficult will increase the chances of missing a genuine significant effect (i.e., making a Type II error).

The arguments about which type of error (a false alarm or a miss) is worse are complex and philosophical and depend on the research context, but simplifying greatly, Type I errors lead people to say "we know this to be true," while Type II errors only lead people to say "we didn't learn anything from this experiment," so Type I errors are probably more dangerous. We should avoid alpha bloat. The Bonferroni correction is very easy to do, but it is somewhat over-conservative (your total alpha sometimes actually falls short of 5 percent). There are other, more complicated but more elegant correction methods. See Simes (1986) for a popular alternative.

6.5.2 The File Drawer Problem

Throughout our discussion of inferential analysis, we concentrated on when we could reject H_0. Recall that we conceptualize H_0 as the "nothing is happening"

https://doi.org/10.1017/9781108654555.007

hypothesis. Naturally, it is satisfying to be able to reject it and conclude that something is, after all, happening. It is, in fact, common for researchers to prefer to publish only those studies where H_0 was rejected and H_1 was supported. Why, after all, would you want to publicize the fact that your research hypothesis was, quite possibly, wrong? It is also often claimed that it is more *difficult* to publish null results. Journals and conferences are less interested in studies that "didn't work." This pattern is known as the **file drawer problem** (Rosenthal, 1979). We publish our results when we reject H_0, but stick the rest into the file drawer.[23]

The file drawer problem is dangerous in a number of ways. First, it suppresses results that may be of practical value. For example, Maxwell and Taylor (2017) compared student performance (using grades) in two CS2 courses, one with a visual media focus and one with a scientific computation focus. They found no significant difference between the two courses. Rather than tossing this study in the file drawer, the authors used it as an opportunity to consider the extent to which CS material can be embedded into a variety of non-CS contexts.

Second, we must remember that, in NHST, there is a 5 percent (assuming alpha = 5 percent) probability that a decision to reject H_0 is *wrong*. When there is a bias toward publishing significant results, these studies, although they are wrong, actually have an extra chance of being published. In the worst case, for a given false research hypothesis (i.e., when H_0 is true and H_1 is false), the 5 percent of studies that incorrectly support it will be published, and the 95 percent of studies that correctly fail to support it will not be published, leading to a completely inaccurate picture across the literature.

There is no simple answer to the file drawer problem, unfortunately. A cultural change on the part of readers, authors, and journal editors might be required to eliminate it. Recently, there have even been suggestions that all raw research data should simply be made openly accessible on the Internet, so that the "file drawer" becomes publicly available (Yarkoni, 2011).

6.5.3 p-Hacking

The file drawer problem is one consequence of the fact that it's easier to get significant results published than non-significant ones. This **publication bias** means that researchers are under pressure to find significant results. Ideally, we would all increase our chances of getting significant results only through careful experimental design and data collection. Unfortunately, there are other ways to do it that are, essentially, cheating. For example, you could simply exclude any data that don't support your H_1. If you have multiple DVs, you can simply suppress any that don't produce significant results. You can leverage alpha bloat and keep running more and more tests until you find something significant (ignoring the fact that it may well be a Type I error). Such manipulations are collectively termed **p-hacking.** They are methods for manipulating your data

[23] Rosenthal coined this term back when people still used paper and kept it in file cabinets. A more modern digital equivalent is probably needed.

https://doi.org/10.1017/9781108654555.007

until you get p-values that you like, and they are bad statistics. If sufficient p-hacking occurs in a research area, it leads to publication of many false claims and conclusions. In some fields, there is evidence that many published results cannot be replicated, indicating a high incidence of Type I errors, possibly due to systematic p-hacking (Ioannidis, 2005).

The best protection against p-hacking is to decide exactly what data you are going to collect and exactly how you are going to analyze them *before you begin your study*. Do not give in to any later temptations to tweak or fiddle with your data in order to get significant results. There are now online repositories[24] that allow you to pre-register your research plan in advance in order to provide publicly verifiable evidence that your results are free of p-hacks.

6.5.4 Skimping on the Results Section

We have described some of the abundant opportunities to make mechanical, logical, and even philosophical errors in statistical analysis and interpretation. It is every author's responsibility to provide evidence that he or she has avoided these pitfalls by explicating all necessary statistical detail in every manuscript. It is, for example, never sufficient to offer a p-value without saying what inferential test was run to generate it.[25] How can the reader judge whether the correct test was chosen if the test is not identified? It is not sufficient to claim a difference between groups without providing actual data values (group means and, when appropriate, a measure of effect size). How can the reader judge if an effect is of practical value without knowing how large it is? It is not sufficient to omit details of who the participants were, exactly what was measured, how the data were collected, and whether any data values were excluded from the analyses. Without such information, how can the reader assess the validity, reliability, and potential bias of the DV? Assuming that your design and analysis have been performed correctly, providing full details to the reader will give him or her confidence in the conclusions you draw.

6.6 Conclusion

Most computing educators are not statisticians. It is hard enough to be both a teacher and a computer scientist – there's not necessarily time left over to become an expert in statistics. However, to do effective computing education research, we need to follow the correct statistical path. By observing the general principles and techniques described in this chapter, you can successfully find the knowledge buried in your data in the majority of cases. And you will be able to recognize those tricky situations where you need an expert statistical hand to guide you.

[24] See, for example, https://osf.io/

[25] This actually happens quite frequently in published manuscripts.

https://doi.org/10.1017/9781108654555.007

References

Campbell, S. (1974). *Flaws and Fallacies in Statistical Thinking*. Upper Saddle River, NJ: Prentice Hall.

Corder, G. W., & Foreman, D. L. (2014). *Nonparametric Statistics: A Step-by-Step Approach*, 2nd edn. Hoboken, NJ: Wiley.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193–242.

Francis, G. (2017). Equivalent statistics and data interpretation. *Behavior Research Methods*, 49, 1524–1538

Gravetter, F., & Wallnau, W. B. (2012). *Essentials of Statistics for the Behavioral Sciences*, 8th edn. Boston, MA: Wadsworth Publishing.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine,* 2, e124.

Maxwell, B. A., & Taylor, S. R. (2017). Comparing outcomes across different contexts in CS1. In *Proceedings of the 48th ACM Technical Symposium on Computer Science Education (SIGCSE '17)* (pp. 399–403). New York: ACM.

McDonald, J. H (2015). *Handbook of Biological Statistics (online edition).* Retrieved from www.biostathandbook.com/multiplecomparisons.html

Miller, J., & Haden, P. (2006). *Statistical Analysis with the General Linear Model.* Retrieved from www.otago.ac.nz/psychology/otago039309.pdf

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301.

Pearce, S. C. (1992). Introduction to Fisher (1925): Statistical methods for research workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Volume 2: Methodology and Distributions* (pp. 59–65). New York: Springer-Verlag.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3), 751–754.

Smith, M. (2012) *Common ~~Misteaks~~ Mistakes in Using Statistics: Spotting and Avoiding Them*. Retrieved from www.ma.utexas.edu/users/mks/statmistakes/StatisticsMistakes.html

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, 85, 842–860.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14*,* 779–804.

Wood, K., Parsons, D., Gasson, J., & Haden, P. (2013). It's never too early: Pair programming in CS1. In *Proceedings of the Fifteenth Australasian Computing Education Conference* (pp.13–21). Darlinghurst, Australia: Australian Computer Society.

Yarkoni, T. (2011). *Solving the file drawer problem by making the internet the drawer*. Retrieved from www.talyarkoni.org/blog/2009/11/26/solving-the-file-drawer-problem-by-making-the-internet-the-drawer/

https://doi.org/10.1017/9781108654555.007