Chapter 4

Hypothesis Testing and Inferential Statistics

Research is rarely undertaken without some question or idea that was the genesis for the project. Individuals become curious about phenomena, and this curiosity leads to research. Think for a moment about your own curiosities. Have you ever asked, "What if I...", or "I wonder what would happen if..."? For example, one phenomenon that your textbook authors see at their University is text messaging by college students. Text messages are short messages written to friends and family members via cellphone or smartphone. We might ask, "Why do individuals text message?" "What are some factors that predict text messaging?" "Is there something about an individuals' personality that might lead them to prefer text messaging instead of calling someone?"

Within this chapter, we address a number of features that are important considerations in how research is planned. How questions are framed for research will be addressed, concentrating on research and null hypotheses. This is followed by an introduction to probability and sampling distributions. Next, issues of hypothesis testing will be addressed, including Type I and Type II errors, choosing one- or two-tailed significance tests, and interpreting nonsignificant results.

Inferential Statistics, Samples and Populations

Inferential statistics are used to make inferences and generalizations from a single set of sample participants to a larger population of individuals. This is done because directly studying an entire population is impractical. Can you imagine trying to study the entire population of college students – meaning all college students everywhere? Or studying the entire population of individuals who text message? The task would be impossible! True, there are times when a

researcher can analyze an entire population. Our example using data from Nathan's famous hotdog eating contest *was* the entire population of finalists in the contest. And in the late 1980's one of your textbook authors worked with data from the entire population of those diagnosed with AIDS in California (at that time, about 25,000 cases). *However, in most instances, we rely on sample data which is representative of the population.*

Using a single set of sample participants, we draw certain conclusions about the larger population. For example, if an experiment were conducted over and over again using different samples, would we find the same result? What is the probability that our sample findings are due to chance or random error? If our goal is to compare group means from an experiment, certainly groups will differ due to many factors. We want to make sure that the group differences are due to the effects of an independent variable, not just random chance fluctuations in responses called random error. If the amount of random error is small, we can be confident that the group differences reflect a true population difference. Inferential statistics provide us with a way to assess population differences using sample data.

Murphy, Marelich, Rappaport, Hoffman, and Farthing (2007) performed an experiment to increase adherence to medications used by individuals who have HIV/AIDS, and used a sample of 141 HIV-positive individuals. One group of individuals was placed in an intervention condition where they received specific training on taking their medications, and another group of individuals was placed in a standard care condition. Thus, the independent variable is group membership (intervention or standard care). Medication adherence was the dependent variable, measured as a percentage -- if individuals took their medication 8 out of 10 days, they would be 80% adherent. Mean adherence percentages were obtained for both the intervention and standard care conditions. The means were found to be different; those in the intervention

condition showed greater adherence to their medication than those in the standard care condition.

Because inferential statistics were used to evaluate the experiment, we know that if the study were conducted over and over again with different samples of HIV-positive individuals, the mean difference would be found almost every time. Inferential statistics provide us the exact probability that the sample mean differences are due to random error, and as we noted earlier, we want the amount of random error to be small. If we ask, "What is the probability that the sample mean differences are due to chance?", inferential statistics give us that probability. If the probability of random error is small enough, we can be confident that the sample mean difference reflects a true population mean difference.

Null and Research Hypotheses

Before applying inferential statistics to a sample, hypotheses are written. *Hypotheses are tentative ideas waiting to be confirmed or refuted.* A researcher might start with general research questions about phenomena, and those questions are turned into hypotheses for scientific evaluation. In research, we write two types of hypotheses: a null hypothesis and a research hypothesis (or "alternative" hypothesis). The null hypothesis usually states that the group means in the population are equal, or that there is no association between the variables in the population. The research hypothesis states that the population means do differ, or that there is an association between variables.

Although we generate both the null and research hypotheses, it is only the null hypothesis that is statistically assessed. If the null hypothesis is found to be incorrect, we thus accept as correct the research hypothesis. Therefore we only need to assess the null hypothesis. Another reason we only assess the null hypothesis is because it is written in a very exact and testable manner – i.e., the group means will be equal, or there is no association between two variables.

Making such exact statements allow us to assess the exact probability that the null hypothesis is correct. The probability of the null hypothesis being correct is reflected in the term *statistical significance*. A result that is statistically significant -- for example, stating two group means are statistically different from each other -- indicates there is a very low probability of the means being different if the null hypothesis is correct.

We start with a detailed overview of the research hypothesis and how it is formed, followed by the null hypothesis. Even though it is the null hypothesis that is directly assessed with inferential statistics, the research hypothesis is usually included in the published study. Therefore, understanding how to write both the research and null hypotheses is important.

Research Hypothesis

The Research Hypothesis (H_1) is a statement made by a researcher about a behavior or phenomena, stated as an explanation for the behavior or phenomena. In other words, it is a statement that says directly what we think is affecting or influencing the behavior or phenomena. For example, if the behavior we are interested in is sexual jealousy, and we believe males will be more jealous than females when sexual fidelity is threatened, we could write a statement – a research hypothesis – expressing this belief, *H_1: Males will express higher levels of sexual jealousy than females.*" For text messaging behavior, we might surmise that younger adults are more likely to text message than older adults. We would write this research hypothesis as, *H_1: Younger adults will be more likely to utilize text messaging than older individuals.*

To write a research hypothesis, a simple three-step process is applied.

Step 1: Use a broad research question to identify the behavior or phenomena of interest. Starting with a broad research question (such as those listed at the beginning of the chapter)

allows us to identify the behavior or phenomena of interest. In asking "Why do individuals text message?" we can easily identify the behavior of interest is text messaging. If we ask, "What are some reasons people experience sexual jealousy?" the phenomena of interest is sexual jealousy.

Step 2: List a number of factors that would predict the behavior or phenomena. Make a list of factors – called explanative factors -- that might predict, influence, or cause change in the behavior or phenomena. For text messaging, we would make a list of issues that might influence text messaging behaviors. One explanative factor might be an extroverted personality style. Individuals who are extroverted seek social connection and excitement, and are assertive. Therefore, they would be more likely to text message than those who are introverted. Age might be another factor influencing text messaging. Cellphones and smartphones are newer forms of technology, and early adopters of new technologies tend to be younger adults. Therefore, some factors that might influence text messaging would include extroverted personality style and age. For sexual jealousy, we might believe that whether someone is male or female is an explanative factor when sexual infidelity is suspected.

Step 3: Write the research hypothesis by using the explanative factors from Step 2 to "answer" your broad research question from Step 1. The research hypothesis is written using the factors listed from Step 2 in conjunction with the research question from Step 1. For our text messaging example, we write the research question and list the explanative factors:

Research Question What are some factors that predict someone text messaging?

Explanative Factors 1) Extroverted personality style; 2) age

Next, we answer the research question in a statement form using the explanative factors. Because we have two explanative factors, we can write two research hypotheses.

Research Hypothesis 1 H_1: Individuals with an extroverted personality style will be

more likely to utilize text messaging than those individuals who are introverted.

Research Hypothesis 2 H_1: Younger adults will be more likely to utilize text messaging than older individuals.

Notice above how the factors were integrated to answer the research question in statement form. A research project can now be designed to assess whether these statements are viable. For sexual jealousy, if our research question is *"What factors influence sexual jealousy?"* and an explanative factor is whether someone is male or female, we can now answer that research question by integrating sex into a statement, "*H_1: Males will experience higher levels of sexual jealousy than females*."

Null Hypothesis

The null hypothesis (H_0) is a hypothesis a researcher makes regarding the outcome of a statistical test, and focuses on population characteristics (such as means, correlations, counts, etc.) The null hypothesis states that there is nothing systematic happening in the data, meaning any mean differences or associations in the data are due to random error. Unlike research hypotheses, the null hypothesis uses population-based notation instead of descriptive statements. For comparing the mean values of two groups, the null hypothesis would read H_0: $\mu_1 = \mu_2$, indicating the population means in each group are equal. For assessing the association between two variables, the null hypothesis would read *H_0: Rho = 0*, indicating that in the population the covariation between two variables is equal to zero. Statistical tests (e.g., chi-square, correlation, t-test, ANOVA, etc.) are used to test the null hypothesis.

Returning to our example on text messaging behavior, let's retain a sample of individuals and give them a measure of extroversion, and also assess how many text messages they sent in

the past 24-hours. We then create two groups – those who are extroverted, and those who are introverted. Our research hypothesis is *H_1: Individuals with an extroverted personality style will be more likely to utilize text messaging than those individuals who are introverted*. The null hypothesis, written using the population means (μ) will then be H_0: $\mu_1 = \mu_2$, with μ_1 being the population mean of text messaging for the extroverted group, and μ_2 being the population mean of introverted group. The null hypothesis states the group means are equal, and that any difference between the means is due to chance occurrence. A statistical test, such as an independent samples t-test (covered in Chapter 9), is then applied to assess H_0: $\mu_1 = \mu_2$.

Once a null hypothesis is formed and tested, formal language is used to interpret the findings. A null hypothesis is either (a) rejected, or (b) accepted. If we reject the null hypothesis, this indicates the group means are different beyond chance occurrence. We can then evaluate the direction of the mean differences and interpret the findings.

Applying the Research and Null Hypotheses

Let's walk through another example to illustrate how to generate the null hypothesis. Earlier we derived the research hypotheses "*H_1: Males will experience higher levels of sexual jealousy than females*." The null hypothesis is generated next. The research hypothesis states that males will experience higher levels of sexual jealousy than females, so we write the null hypothesis as H_0: $\mu_1 = \mu_2$. According to the null hypothesis, sexual jealousy mean values will be equal between each group.

To assess the null hypothesis, we conduct a study to compare the sexual jealousy of males and females. A sample of 100 individuals (50 men and 50 women) is taken from a college dormitory. Individuals are asked to report their sex (male or female), and are given a standardized measure of sexual jealousy to complete. Therefore, we have two groups of

individuals (males and females), and we have a score for each individual regarding their sexual jealousy.

To test the null hypothesis (H_0: $\mu_1 = \mu_2$) a test statistic and probability level are used to confirm or refute H_0. In the example above, an independent samples t-test statistic could be used to assess whether the group mean values are equal or different. A mean difference between the sample groups is calculated, and the probability of that mean difference occurring is derived. If the probability is small (say less than .05 or 5%), we conclude there is less than a 5% chance that this mean difference occurred by chance. In other words, it is doubtful that the observed mean differences occurred by chance – one group reports a higher level of sexual jealousy compared to the other.

In terms of the statistical hypotheses, formal language is used when concluding which of the hypotheses are viable.

- If we find $\mu_1 \neq \mu_2$, we reject the null hypothesis (H_0). We would next look at the direction of the mean difference to see if our research hypothesis is supported.
- If we find $\mu_1 = \mu_2$, then we accept the null hypothesis (H_0). Our sample means did not differ.

Probability and Sampling Distributions

Once hypotheses have been written, we can move next into how to formally evaluate them through the use of probability. We first introduced issues of probability in Chapter 2 when presenting the normal distribution. Probability is the likelihood of an event occurring, or the likelihood of getting a particular outcome. Probabilities are everywhere if you take a moment and look around. During baseball season we know the probability that Derek Jeter of the New York Yankees will get a hit – his batting average is a probability. People who play roulette

evaluate the probability of the ball landing on red or black and make their bets accordingly. And every day the U.S. Geological Survey provides the probability of an earthquake occurring in California.

We use probability in the same manner for research. We ask, "What is the probability that the outcome from our sample data is due to chance occurrence?" In other words, if we had two group means and used a statistical test to assess the difference, what is the probability that the mean difference is due to chance? If the resulting probability is very low, we reject the notion that the mean difference is solely due to random or chance error. We can be confident that the sample mean difference reflects a true population difference

Probability and the Three-Card Monte

A simple way to understand probability is to see how it can be applied to an example. One of your book authors has recently begun exploring the world of card tricks, and one classic card trick is the 3-card Monte. It's a simple trick, but most people think it's a card game. There are a lot of different variations of the trick, but we will present one of the typical setups.

You show three cards to your subject face up on a table: a queen, a king, and an ace. Ask the subject to keep an eye on the queen and turn all the cards face down on the table. Move the cards around picking them up a number of times and placing them in different orders. You ask the subject, "Where is the queen?" The subject points to one of the face-down cards, and when flipped over if it's the queen, they win – they found the queen! But if it's the king or ace, they lose. There is a 1 in 3 chance of picking the correct card (a 33.3% chance of winning). The "trick" portion of the three-card Monte will be explained later, but for now we will work with the 33.3% chance of winning to explain probability.

Let's run an experiment where you present three cards to a friend and have them find the

queen. To make this an experiment, you perform the game 9 times with your friend. You show him where the queen is, then turn the cards face down and move the cards around, and have him try to find the queen. Do this 9 times (therefore, you have 9 trials of the game). How successful is your friend across the 9 trials in finding the queen? Did his "success" in finding the queen across the trials reflect random error, or did he do better than guessing?

Earlier we noted a 33.3% chance of picking the queen (1 out of 3). If we write a null hypothesis, it would reflect this random chance of the queen being chosen, H_0: $p = .333$, where p is the expected probability of picking a queen. But let's say you are optimistic that your friend will do better than this 33.3% chance. We can say for H_1 that your friend will choose the queen at a higher rate than 33.3%. In other words, he will do better than guessing at finding the queen.

Based on the null hypothesis, we can expect that someone will find the queen in 3 of the 9 trials (we expect 1 out of 3, so with 9 trials we expect 3 correct choices). In performing this experiment, however, if your friend were to guess the queen more than 3 times, does this reflect something beyond random guessing? If he finds the queen 4 times out of the 9 trials, this isn't that much different than finding the queen 3 times (what we expect under the null hypothesis), so you might conclude that he is doing no better than randomly guessing.

But suppose your friend finds the queen 8 times out of the 9 trials. That's really impressive (almost 89% correct)! Getting 8 out of 9 seems quite beyond the null hypothesis, and probably indicates he is doing better than random guessing. If we apply the research hypothesis to his result, we would say H_1: *The probability of picking the queen in at least 8 out of 9 trials is greater than .333 ($p > .333$).* If this is true, we would reject the null hypothesis and state that getting 8 of 9 trials correct to be significant. When we use the term "significant" we mean that the result is unlikely if the null hypothesis is true.

However, we haven't formally tested this result. Finding the queen almost 89% of the time is very impressive, but is it significant? We need a decision rule or cutoff to evaluate whether this is a significant finding. The cutoff we adopt is called the *alpha level*. Usually the alpha level adopted is .05. If the probability of our study outcome is less than .05, then our results are deemed significant; there is less than a 5% chance in the broader population that our study outcome was due to random error, and the null hypothesis is rejected.

Sampling Distributions

Your friend got 8 of the 9 trials correct, and we may conclude finding the queen that many times is very unlikely. A table of probabilities (Table 4.1) can be used to ascertain the exact probability of picking the queen 8 out of 9 trials *given* the expectation of 33% correct from the null hypothesis. We can ask, "Assuming the null hypothesis is true (i.e., $p = .333$), what is the probability of picking the queen in at least 8 out of 9 trials?" If the resulting probability is less than .05, we conclude getting 8 out of 9 trials correct is statistically significant and reject the null hypothesis ($p = .333$).

We can see from Table 4.1 that finding the queen exactly 3 out of 9 times results is a probability of 0.2731 (which is the highest probability in the table); there is a 27% chance of finding the queen 3 times out of 9 trials. And it is very probable that one could guess 4 out of 9 (a probability of 0.2045, or about 20%). An outcome of 8 out of 9 trials correct, however, is very unlikely (.0009). Our research hypothesis was H_1*: The probability of picking the queen in at least 8 out of 9 trials is greater than .333 (p > .333)*. The probability of getting 8 out of 9 trials correct is .0009. Since our research hypothesis states picking the queen *in at least 8 out of 9 trials*, we also need the probability of 9 out of 9 trials, which is .00005. The probability then of picking the queen in at least 8 out of 9 trials is *.00095*: .0009 + .00005 = .00095. Is this

probability smaller than our alpha level of .05? It is, and therefore the null hypothesis is rejected.

The probability values in Table 4.1 come from the *binomial distribution*. As we saw in Chapter 2, various distributions are used when applying inferential statistics, in particular the normal distribution (the binomial distribution was used here because we had a success/failure outcomes and a limited number of trials). The distributions we use for inferential statistics are called *sampling distributions*, and reflect the assumption that the null hypothesis is true. They allow us to assess the likelihood of an event or outcome. Sampling distributions are the foundation on which statistical tests are built, and allow us to evaluate the probability of our sample data results given the null hypothesis is correct. If the resulting probability of an event or outcome is very small (usually less than .05), the null hypothesis is rejected and we can accept the research hypothesis.

By the way, getting 8 out of 9 trials correct is so rare, one should become suspect that shenanigans are afoot. Indeed, in the three-card Monte trick, having a person get so many trials correct is part of the trick. Typically, the trick is performed in front of a small crowd of people, with the goal to have individuals bet their money when it is their turn to play the game. To entice people to play, the dealer uses a "mark" (a person who is in on the trick), who wins over and over again to make the game look easy. Here, your friend is the mark and he gets 8 out of 9 trials correct -- what an easy game! He wins so much because you manipulate where the queen lands. This is done using a card throw technique, where you retain the queen in your hand as you move the cards around, eventually placing it in a predefined location that the "mark" knows. The next person to play assumes the game is easy and is eager to play. As you move the cards around, you again use the card throw technique, this time to trick the person into the other two cards.

Type I and Type II errors

In the previous section, we discussed hypothesis testing and probability of the null hypothesis being true. Since the evaluation of the null hypothesis is probabilistic, *we do not know for certain* if the decision to reject the null hypothesis, or accept the null hypothesis, is correct. Errors associated with making the wrong decision in these situations are called Type I and Type II errors. A decision matrix is shown in Figure 4.1.

A Type I error is when we decide to reject the null hypothesis, when in fact we should have <u>accepted</u> the null hypothesis. Type I error is equal to the research study alpha or probability level. For example, say in our example above, with H_0: $\mu_1 = \mu_2$, we use a t-test and conclude at a $p < .05$ level that the two groups of individuals differ in terms of sending text messages – we reject the null hypothesis. Here, we are concluding there is at least a 95% probability that if we repeated this study, the resulting mean differences will be replicated. In other words, if we drew 100 samples from the same population, we would get these same results in at least 95 of the samples. Notice, however, that in 5% of these samples, the means will not be different. Therefore, the Type I error rate for this study is 5%.

A Type II error is when we accept the null hypothesis, but we should have <u>rejected</u> it. Type II error is associated with how sensitive your research design is in detecting effects (called power, which is labeled Beta or B). When research is designed, the probability of Type II error (i.e., Beta or B) should be low. Type II error is influenced by three issues:

(1) significance (alpha) level adopted for the study

(2) study sample size

(3) effect size

Setting a very low alpha levels will decrease Type I error, but will increase Type II error. Larger

samples will be more likely to reveal true population differences, thus reducing Type II error. If an effect is small, a small sample will increase Type II error since finding true differences will be harder. But if the effect is large, Type II error is doubtful regardless of sample size since the effect is so evident.

As we noted earlier, Type I and II errors are typically placed into a decision matrix (Figure 4.1). We illustrate how this matrix works by looking at a non-research example; the results of a home pregnancy test. The decision to be made is whether the individual is really pregnant. In Figure 4.2, we present the pregnancy test decision matrix. The null hypothesis is that the individual is not pregnant. If the null hypothesis is rejected, the individual is pregnant. Just like when making statistical decisions, we can see the two types of possible errors. Making a Type I error is when we conclude that the individual is pregnant, when in fact they are not. Making a Type II error is when we conclude that the individual is not pregnant, when in fact they really are.

Researchers traditionally have used either a .05 or a .01 significance level in the decision to reject the null hypothesis. If there is less than a .05 or .01 probability that the results occurred because of random error, the results are said to be significant. However, there is nothing magical about a .05 or a .01 significance level. The significance level chosen merely specifies the probability of a Type I error if the null hypothesis is rejected. The significance level chosen by the researcher usually is dependent on the consequences of making a Type I verses a Type II error. In the home pregnancy test example, which would be a more serious error, Type I or Type II? Type II error here would be more serious. The home pregnancy test would reveal no pregnancy when in fact the individual is pregnant. Type I error is less serious since individuals follow-up a positive home pregnancy test with a visit to a healthcare provider for a more specific

test (which would reveal the true negative status).

Although we just presented an example where Type II error is deemed more serious, researchers generally believe that the consequences of making a Type I error are more serious. If the null hypothesis is rejected, the researcher might publish the results in a journal, and the results might be reported by others in textbooks or in newspaper or magazine articles. Researchers don't want to mislead people or risk damaging their reputations by publishing results that aren't reliable and so cannot be replicated. In our example on text messaging, if we conclude that younger individuals text message more than older individuals, we want to make sure this result is reliable, otherwise we might ruin our reputations by making such a bold statement. Thus researchers want to guard against the possibility of making a Type I error by using a very low significance level (.05 or .01).

However, in certain circumstances a Type I error is not serious. For example, if you were engaged in exploratory research, your results would be used primarily to decide whether your research ideas were worth pursuing. In this situation, it would be a mistake to overlook potentially important data by using a very conservative significance level. In exploratory research, a probability or significance level of .15 may be more appropriate for deciding whether to do more research. Another situation is when there is little consequence in committing a Type I error. For example, home pregnancy tests are actually designed to be extremely sensitive to any hormonal evidence of pregnancy. Even though this heightens the chance of a Type I error (a false positive), the consequence of a false positive test is minimized since a more specific pregnancy test is eventually performed through a healthcare provider.

The point is that the significance level chosen and the consequences of a Type I or a Type II error are determined by what the results will be used for.

Choosing a One-tailed or Two-tailed test of significance

So far, we have introduced you to a number of different issues regarding hypotheses, probability and sampling distributions, and errors in statistical decisions. In particular, the *alpha level* was presented, which is our decision-point or cutoff used to assess the null hypothesis. Not only is the alpha level used to assess the null hypothesis, it is also used as an indicator of Type I error.

When we adopt an alpha level such as .05, we need to make one additional decision before we apply an inferential statistic – whether to use the full sampling distribution (called a two-tailed test) for our assessment, or use half of the sampling distribution (a one-tailed test). The choice of using a one-tailed or two-tailed significance test to assess the significance of a test statistic is based on how the null hypothesis (H_0) is written. Although most published research reflect reflects two-tailed significance tests, it is important to know the circumstances when a one-tailed test is appropriate.

In the earlier example addressing text messaging and extroversion, the null hypothesis was written as $\mu_1 = \mu_2$. A two-tailed test of significance would be used which allows for a statistical decision to be discovered in either direction. In other words, $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$ can be the outcomes that would allow us to reject the null hypothesis. Thus, whether extroverted individuals text message more than introverted individuals, or whether introverted individuals text message more than extroverted individuals, both sets of findings may be discovered.

In most instances, the null hypothesis will be written in such a manner as to suggest that groups do not differ, or that there is no association between variables. However, there are situations when the null hypothesis is directional. In such instances, a one-tail test would be used to assess the directional null hypothesis. If the null hypothesis for a study were written as H_0:

$\mu_1 < \mu_2$, we would use a one-tail test to assess whether $\mu_1 < \mu_2$ was true. Here, unlike H_0: $\mu_1 = \mu_2$ in the previous example, we would want to show significance for $\mu_1 < \mu_2$. If we find $\mu_1 < \mu_2$ is significant, then we will accept the null hypothesis. If we did not find $\mu_1 < \mu_2$ to be significant, then we would reject the null hypothesis.

For example, a directional hypothesis would be applied in an experiment which evaluates negative side effects of a new drug to treat migraine headaches. New drugs are always assessed for side effects, and we want to make sure that the new drug does not generate more side effects than the standard drug treatment. For the experiment, an intervention group receives the new migraine headache drug, and a control group receives standard drug care. The null hypothesis is that the side effects mean for the intervention group (μ_2) will be greater than the side effects mean (μ_1) for the control group: $\mu_1 < \mu_2$. A one-tailed test of significance is applied since we are only testing $\mu_1 < \mu_2$. If the null hypothesis is accepted, the new drug exhibits more side effects and the drug should not be released. If $\mu_1 < \mu_2$ is not true, then the new drug does not exhibit more side effects than the standard drug care for migraines. In this instance we only want to insure that the new drug doesn't show more side effects.

As a final note, you will find that some textbooks explain in great detail one-tailed significance tests and how they are applied to assess your research hypothesis (not null hypothesis). Those approaches are viable within statistics and research methods texts. However, in the current text, we have chosen to make the focus of hypothesis testing the null hypothesis, following R.A. Fisher's writings.

Nonsignificant Results

Although one might believe most studies yield significant results, generally that is not true. It just appears to be the case since most of the studies published in peer-review journals

report significant findings. In fact, researchers do conduct studies that fail to reach significance.

What leads to a nonsignificant finding? Certainly the nonsignificant finding could be true, indicating no effect in the population and thus we would accept the null hypothesis. But, there are a number of reasons one fails to reject the null hypothesis, thus possibly committing a Type II error. The issues listed below should be considered when a single study reaches nonsignificant results (i.e., accept the null hypothesis).

As we indicated earlier regarding Type II error, a nonsignificant finding may be due to a small effect size coupled with a small sample size. Or a significance level (alpha) that is too stringent was utilized. For example, using an alpha level of .001 would increase the Type II error rate considerably. Sometimes, procedural issues in the study are problematic. Possibly the questionnaire was poorly written, or the experimental manipulation was not applied correctly. How the dependent variable was operationalized may also be an issue. Even the statistical test chosen to evaluate the null hypothesis may be inappropriate.

Overall, we wish to emphasize that a nonsignificant finding does not necessarily indicate that the null hypothesis is true. Evaluating the points above would be the next step, and if any issues are discovered, the analyses and/or the entire study are conducted again addressing the problems.

Interesting Web Links

http://www.stat.tamu.edu/~west/applets/binomialdemo.html

Sample Exercises (to be expanded in later drafts)

1. Pick 3 research articles which you find interesting from peer-reviewed research journals. Do the following:

    a) Within these articles, find the Null and Research hypotheses. If the hypotheses are not clearly stated, derive what they are based on the content of the articles.

    b) Evaluate whether the authors use a one-tail or two-tail test of significance, and also note the probability level(s) used for the statistical tests.

    c) Create a decision matrix for Type I and Type II errors similar to the matrices presented in this chapter for the different hypotheses you discover.

2. Practice writing Null and Research hypotheses by choosing different phenomena you find interesting, following the techniques outlined in the section *Null and Research Hypotheses*. How testable do you think these hypotheses are? How might you design a research study to evaluate these hypotheses?

3. We formally challenge you to find a research article that (a) uses a one-tail significance test, and/or (b) uses a probability level greater than .05 (such as .10 or even higher). What arguments (if any) did the authors make for choosing these alternatives? What information may have been gained or loss?

Table 4.1: Exact probability of choosing a queen for the three-card Monte experiment (9 trials) assuming the probability of success is 33.3%

Number of correct answers	Probability
0	0.0261
1	0.1174
2	0.2345
3	0.2731
4	0.2045
5	0.1021
6	0.0340
7	0.0073
8	0.0009
9	0.00005

Figure 4.1: Decision matrix Type I and Type II errors

		Population	
		Null Hypothesis is True	Null Hypothesis is False
Decision	Reject Null Hypothesis	*Type I error (α)*	**Correct decision (1-** B **)**
	Accept the Null Hypothesis	**Correct decision (1-α)**	*Type II error (* B *)*

Figure 4.2: Decision matrix for pregnancy test

		True State	
		Null Hypothesis is True (NO pregnancy)	Null Hypothesis is False (YES pregnancy)
Decision	Reject Null Hypothesis (test indicates pregnancy)	*Type I error*	**Correct decision**
	Accept the Null Hypothesis (test indicates NO pregnancy)	**Correct decision**	*Type II error*