Statistics Minus The Math: An Introduction for the Social Sciences

Nathan Favero

2024-02-11

Table of contents

Introduction

This version (1.2) was updated 1/28/2023.

Version 1.1 available at https://nathanfavero.com/teaching-tools/. The only change from 1.1 is that the discussion of transforming variables now appears in Ch. 2 (rather than Ch. 3).

This book was largely adapted from the public domain resource *Online Statistics Education: A Multimedia Course of Study* (http://onlinestatbook.com/ Project Leader: David M. Lane, Rice University).

1 Hypothesis Testing

1.1 Introduction to Hypothesis Testing[1]

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who insisted that martinis should be shaken rather than stirred. Let's consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. This probability can be computed from the binomial distribution, and a binomial distribution calculator[2] shows it to be 0.0106. This is a pretty low probability, and therefore someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Let's consider another example. The case study Physicians' Reactions[3] sought to determine whether physicians spend less time with obese patients. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of average weight. The chart a particular physician viewed was determined randomly. Thirty-three physicians viewed charts of average-weight patients and 38 physicians viewed charts of obese patients.

The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for average-weight patients. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the obese patients tend to see patients for less time than the other physicians. Random assignment of charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the two groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. With this in mind, is it plausible that these chance differences are responsible for the difference in times?

To assess the plausibility of the hypothesis that the difference in mean times is due to chance, we compute the probability of getting a difference as large or larger than the observed difference (31.4 - 24.7 = 6.7 minutes) if the difference were, in fact, due solely to chance. Using methods presented in a later chapter, this probability can be computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight and is not due to chance.

1.1.1 The Probability Value

It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of 0.0106 is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing.

It is easy to mistake this probability of 0.0106 as the probability he cannot tell the difference. This is not at all what it means.

The probability of 0.0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on 9/16 choices. Using the binomial calculator, we can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than 0.0001.

To reiterate, the **probability value (p value)** is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird was only guessing). In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. Using this terminology, the probability value is the probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference. A branch of statistics called Bayesian statistics provides methods for computing the probabilities of hypotheses. These computations require that one specify the probability of the hypothesis before the data are considered and, therefore, are difficult to apply in some contexts.

1.1.2 The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the **null hypothesis**. In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as:

$$\mu_{obese} = \mu_{average}$$

or as

$$\mu_{obese} - \mu_{average} = 0.$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as

$$\rho = 0$$

where ρ is the population correlation (not to be confused with r, the correlation in the sample).

Although the null hypothesis is usually that the value of a *population parameter* is 0, there are occasions in which the null hypothesis is a value other than 0. For example, if one were testing whether a subject differed from chance in their ability to determine whether a flipped coin would come up heads or tails, the null hypothesis would be that $\pi = 0.5$.

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. The **alternative hypothesis** is simply the reverse of the null hypothesis. If the null hypothesis

$$\mu_{obese} = \mu_{average}$$

is rejected, then there are two alternatives:

$$\mu_{obese} \leq \mu_{average}$$

$$\mu_{obese} \geq \mu_{average}$$

Naturally, the direction of the sample means determines which alternative is adopted. Some textbooks have incorrectly argued that rejecting the null hypothesis that two population means are equal does not justify a conclusion about which population mean is larger. Kaiser (1960)[4] showed how it is justified to draw a conclusion about the direction of the difference.

1.2 Steps in Hypothesis Testing[5]

There's much to learn about hypothesis testing, but before going any further, here's an overview of the four basic steps of any hypothesis test. Some of the details won't make sense yet, but we'll explain them in more detail in the following sections.

1. The first step is to ***specify the null hypothesis***. For a two-tailed test, the null hypothesis is typically that a parameter equals zero although there are exceptions. A typical null hypothesis is μ_1 - μ_2 = 0 which is equivalent to μ_1 = μ_2. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that μ_1 is larger than μ_2, then the null hypothesis (the reverse of the prediction) is μ_2 - μ_1 0. This is equivalent to μ_1 μ_2.
2. The second step is to ***specify the*** α level which is also known as the significance level. Typical values are 0.05 and 0.01.

3. The third step is to ***compute the probability value*** (also known as the p value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.
4. Finally, ***compare the probability value with the*** α level. If the probability value is lower then you reject the null hypothesis. Keep in mind that rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

1.3 One- and Two-Tailed Tests[6]

In the James Bond case study,[7] Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred. He was correct on 13 of the trials. From the binomial distribution, we know that the probability of being correct 13 or more times out of 16 if one is only guessing is 0.0106. Figure 1.1 shows a graph of the binomial distribution. The red bars show the values greater than or equal to 13. As you can see in the figure, the probabilities are calculated for the upper tail of the distribution. A probability calculated in only one tail of the distribution is called a "one-tailed probability."

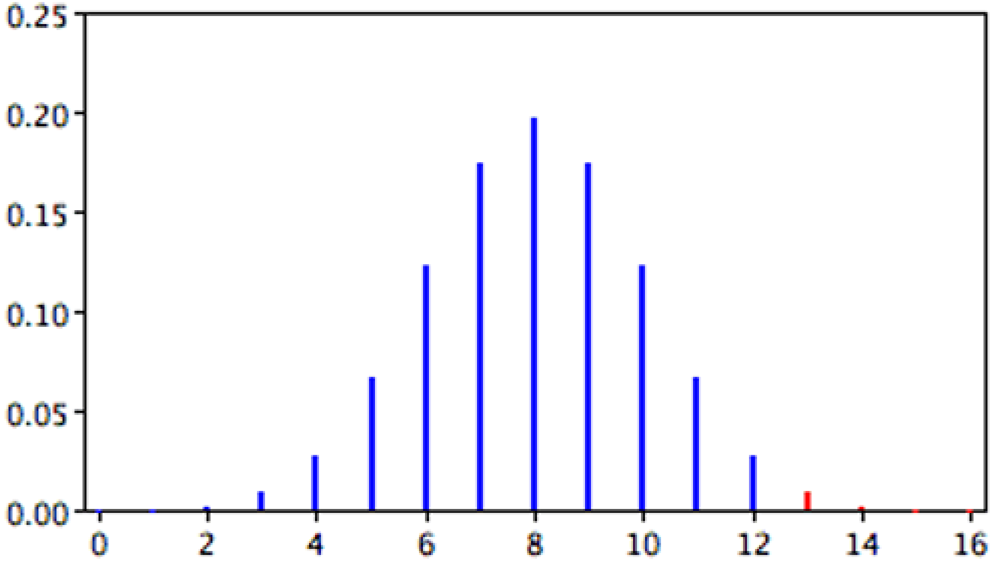


Figure 1.1: The binomial distribution. The upper (right-hand) tail is red.

A slightly different question can be asked of the data: “What is the probability of getting a result as extreme or more extreme than the one observed?” Since the chance expectation is 8/16, a result of 3/16 is equally as extreme as 13/16. Thus, to calculate this probability, we would consider both tails of the distribution. Since the binomial distribution is symmetric when $\pi = 0.5$, this probability is exactly double the probability of 0.0106 computed previously. Therefore, p = 0.0212. A probability calculated in both tails of a distribution is called a “two-tailed probability” (see Figure 1.2).

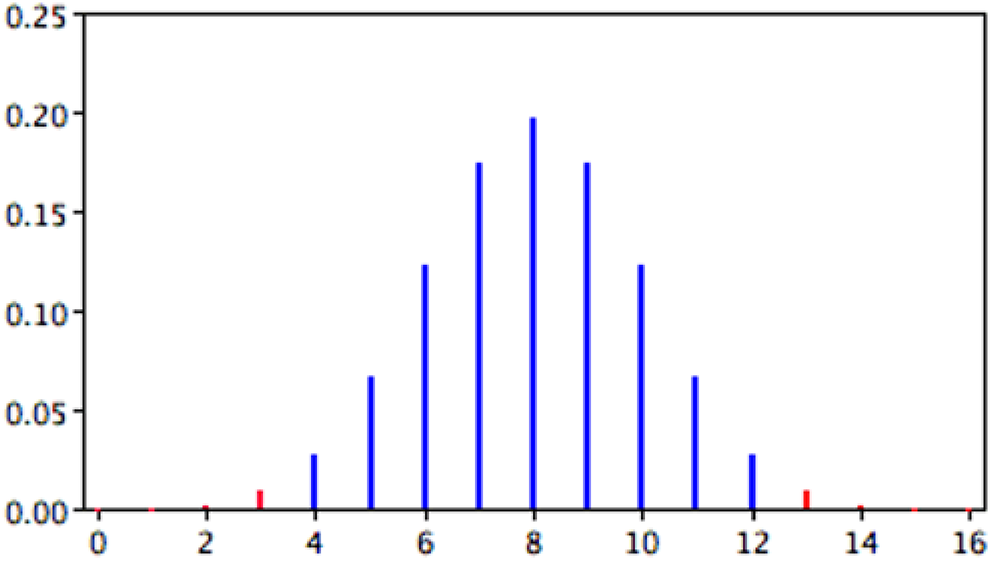


Figure 1.2: The binomial distribution. Both tails are red.

Should the one-tailed or the two-tailed probability be used to assess Mr. Bond’s performance? That depends on the way the question is posed. If we are asking whether Mr. Bond can tell the difference between shaken or stirred martinis, then we would conclude he could if he performed either much better than chance or much worse than chance. If he performed much worse than chance, we would conclude that he can tell the difference, but he does not know which is which. Therefore, since we are going to reject the null hypothesis if Mr. Bond does either very well or very poorly, we will use a two-tailed probability.

On the other hand, if our question is whether Mr. Bond is better than chance at determining whether a martini is shaken or stirred, we would use a one-tailed probability. What would the one-tailed probability be if Mr. Bond were correct on only 3 of the 16 trials? Since the one-tailed probability is the probability of the right-hand tail, it would be the probability of getting 3 or more correct out of 16. This is a very high probability and the null hypothesis would not be rejected.

The null hypothesis for the two-tailed test is $\pi = 0.5$. By contrast, the null hypothesis for the one-tailed test is $\pi \quad 0.5$. Accordingly, we reject the two-tailed hypothesis if the sample proportion deviates greatly from 0.5 in either direction. The one-tailed hypothesis is rejected only if the sample proportion is much greater than 0.5. The alternative hypothesis in the two-tailed test is $\pi \neq 0.5$. In the one-tailed test it is $\pi > 0.5$.

You should always decide whether you are going to use a one-tailed or a two-tailed probability before looking at the data. Statistical tests that compute one-tailed probabilities are called one-tailed tests; those that compute two-tailed probabilities are called two-tailed tests. Two-tailed tests are much more common than one-tailed tests in scientific research because an outcome signifying that something other than chance is operating is usually worth noting. One-tailed tests are appropriate when it is not important to distinguish between no effect and an effect in the unexpected direction. For example, consider an experiment designed to test the efficacy of a treatment for the common cold. The researcher would only be interested in whether the treatment was better than a placebo control. It would not be worth distinguishing between the case in which the treatment was worse than a placebo and the case in which it was the same because in both cases the drug would be worthless.

Some have argued that a one-tailed test is justified whenever the researcher predicts the direction of an effect. The problem with this argument is that if the effect comes out strongly in the non-predicted direction, the researcher is not justified in concluding that the effect is not zero. Since this is unrealistic, one-tailed tests are usually viewed skeptically if justified on this basis alone.

1.4 Significance Testing[8]

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α (alpha) level or simply α. It is also called the significance level.

When the null hypothesis is rejected, the effect is said to be **statistically significant**. For example, in the Physicians' Reactions case study,[9] the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what "significant" usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is.

Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough.

Why does the word "significant" in the phrase "statistically significant" mean something so different from other uses of the word? Interestingly, this is because the meaning of "significant" in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was "significant" if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of "significant" changed, leading to the potential misinterpretation.

There are two approaches (at least) to conducting significance tests. In one (favored by R. Fisher), a significance test is conducted and the probability value reflects the strength of the evidence against the null hypothesis. If the probability is below 0.01, the data provide strong evidence that the null hypothesis is false. If the probability value is below 0.05 but larger than 0.01, then the null hypothesis is typically rejected, but not with as much confidence as it would be if the probability value were below 0.01. Probability values between 0.05 and 0.10 provide weak evidence against the null hypothesis and, by convention, are not considered low enough to justify rejecting it. Higher probabilities provide less evidence that the null hypothesis is false.

The alternative approach (favored by the statisticians Neyman and Pearson) is to specify an α level before analyzing the data. If the data analysis results in a probability value below the α level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected. According to this perspective, if a result is significant, then it does not matter how significant it is. Moreover, if it is not significant, then it does not matter how close to being significant it is. Therefore, if the 0.05 level is being used, then probability values of 0.049 and 0.001 are treated identically. Similarly, probability values of 0.06 and 0.34 are treated identically.

The former approach (preferred by Fisher) is more suitable for scientific research and will be adopted here. The latter is more suitable for applications in which a yes/no decision must be made. For example, if a statistical analysis were undertaken to determine whether a machine in a manufacturing plant were malfunctioning, the statistical analysis would be used to determine whether or not the machine should be shut down for repair. The plant manager would be less interested in assessing the weight of the evidence than knowing what action should be taken. There is no need for an immediate decision in scientific research where a researcher may conclude that there is some evidence against the null hypothesis, but that more research is needed before a definitive conclusion can be drawn.

1.5 Testing a Single Mean[10]

The way we calculate the probability (p) value for a hypothesis test depends on what type of statement is made in our null hypothesis. Normally, statistical software will automatically compute a p value behind the scenes, but we still want to learn a bit about how the software comes up with this value. To illustrate what these calculations can look like, this section will focus on what to do if we want to test a null hypothesis stating that the population mean is equal to some hypothesized value. For example, suppose an experimenter wanted to know if people are influenced by a subliminal message and performed the following experiment. Each of nine subjects is presented with a series of 100 pairs of pictures. As a pair of pictures is presented, a subliminal message is presented suggesting the picture that the subject should choose. The question is whether the (population) mean number of times the suggested picture is chosen is equal to 50. In other words, the null hypothesis is that the population mean (μ) is 50. The (hypothetical) data are shown in Table 1.1. The data in Table 1.1 have a sample mean (M) of 51. Thus the sample mean differs from the hypothesized population mean by 1.

Table 1.1: Distribution of scores.

Frequency
45
48
49
49
51
52
53
55
57

The significance test consists of computing the probability of a sample mean differing from μ by one (the difference between the hypothesized population mean and the sample mean) or more. The first step is to determine the sampling distribution of the mean. As we learned in the prior chapter, the mean and standard deviation of the sampling distribution of the mean are

$$\mu_M = \mu$$

and

$$\sigma_M = \frac{\sigma}{\sqrt(N)}$$

respectively. It is clear that $\mu_M = 50$. In order to compute the standard deviation of the sampling distribution of the mean, we have to know the population standard deviation (σ).

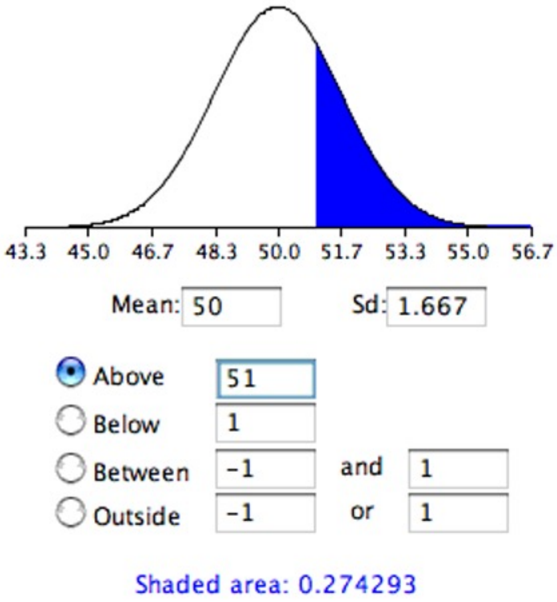
The current example was constructed to be one of the few instances in which the standard deviation is known. In practice, it is very unlikely that you would know σ and therefore you would use s, the sample estimate of σ. However, it is instructive to see how the probability is computed if σ is known before proceeding to see how it is calculated when σ is estimated.

For the current example, if the null hypothesis is true, then based on the binomial distribution, one can compute that variance of the number correct is

$$\sigma^2 = N\pi(1-\pi) = 100(0.5)(1-0.5) = 25.$$

Therefore, $\sigma = 5$. For a σ of 5 and an N of 9, the standard deviation of the sampling distribution of the mean is $5/3 = 1.667$. Recall that the standard deviation of a sampling distribution is called the standard error.

To recap, we wish to know the probability of obtaining a sample mean of 51 or more when the sampling distribution of the mean has a mean of 50 and a standard deviation of 1.667. To compute this probability, we will make the assumption that the sampling distribution of the mean is normally distributed. We can then use a normal distribution calculator as shown in Figure 1.3.



Mean: 50 Sd: 1.667

● Above 51

○ Below 1

○ Between -1 and 1

○ Outside -1 or 1

Shaded area: 0.274293

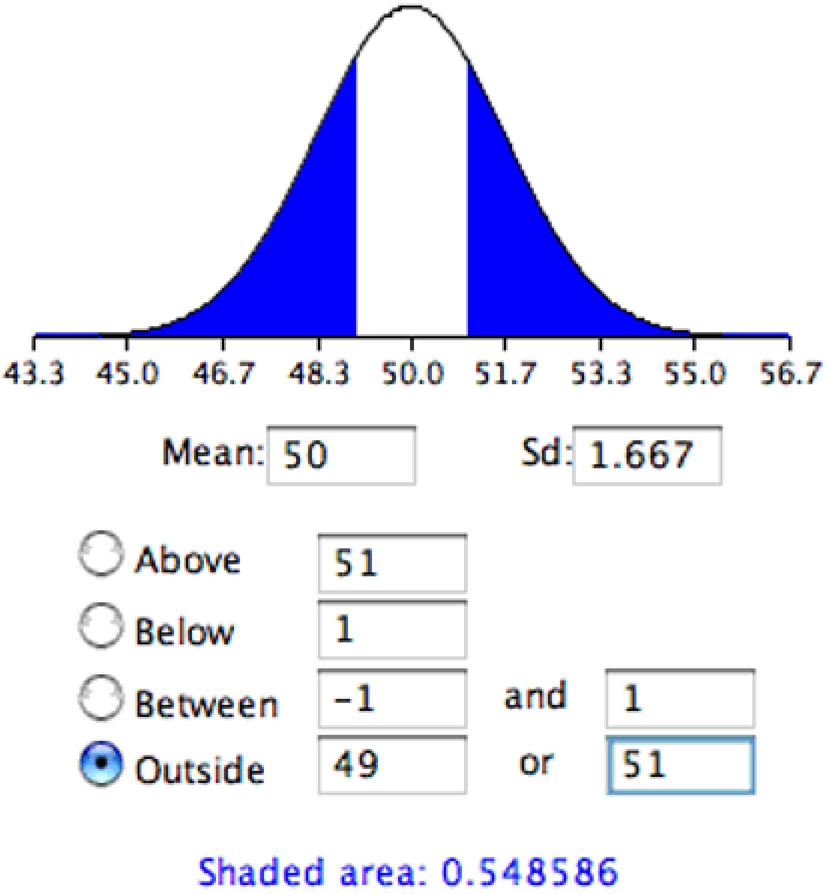
Figure 1.3: Probability of a sample mean being 51 or greater.

Notice that the mean is set to 50, the standard deviation to 1.667, and the area above 51 is requested and shown to be 0.274.

Therefore, the probability of obtaining a sample mean of 51 or larger is 0.274. Since a mean of 51 or higher is not unlikely under the assumption that the subliminal message has no effect, the effect is not significant and the null hypothesis is not rejected.

The test conducted above was a one-tailed test because it computed the probability of a sample mean being one or more points higher than the hypothesized mean of 50 and the area computed was the area above 51. To test the two-tailed hypothesis, you would compute the probability of a sample mean differing by one or more in either direction from the hypothesized mean of 50. You would do so by computing the probability of a mean being less than or equal to 49 or greater than or equal to 51.

The results from a normal distribution calculator are shown in Figure 1.4.



Mean: 50 Sd: 1.667

○ Above 51

○ Below 1

○ Between -1 and 1

● Outside 49 or 51

Shaded area: 0.548586

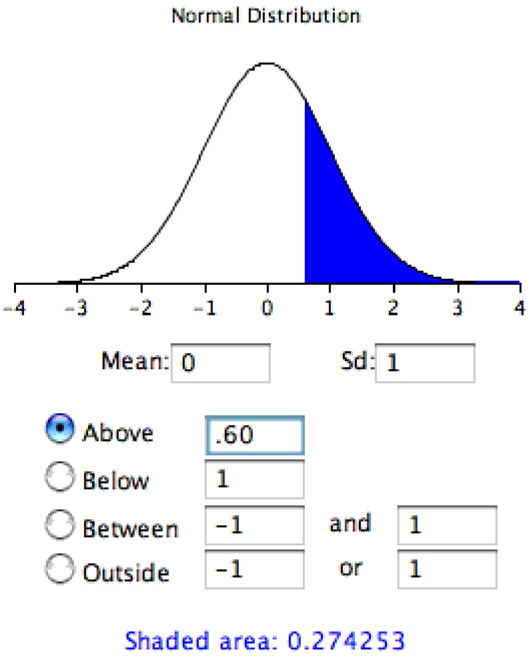
Figure 1.4: Probability of a sample mean being less than or equal to 49 or greater than or equal to 51.

As you can see, the probability is 0.548 which, as expected, is twice the probability of 0.274 shown in Figure 1.3.

Before normal calculators such as the one illustrated above were widely available, probability calculations were made based on the standard normal distribution. This was done by computing Z based on the formula

$$Z = \frac{M - \mu}{\sigma_M}$$

where Z is the value on the standard normal distribution, M is the sample mean, μ is the hypothesized value of the mean, and σ_M is the standard error of the mean. For this example, $Z = (51\text{-}50)/1.667 = 0.60$. Use a normal calculator, with a mean of 0 and a standard deviation of 1, as shown below.



Mean: 0 Sd: 1

● Above .60
○ Below 1
○ Between -1 and 1
○ Outside -1 or 1

Shaded area: 0.274253

Figure 1.5: Calculation using the standardized normal distribution.

Notice that the probability (the shaded area) is the same as previously calculated (for the one-tailed test).

As noted, in real-world data analyses it is very rare that you would know σ and wish to estimate μ. Typically σ is not known and is estimated in a sample by s, and σ_M is estimated by s_M. For our next example, we will consider the data in the "ADHD Treatment" case study.[11] These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. Table 1.2 shows the data for the placebo (0 mg) and highest dosage level (0.6 mg) of methylphenidate. Of particular interest here is the column labeled "Diff" that shows the difference in performance between the 0.6 mg (D60) and the 0 mg (D0) conditions. These difference scores are positive for children who performed better in the 0.6 mg condition than in the control condition and negative for those who scored better in the control condition. If methylphenidate has a positive effect, then the mean difference score in the population will be positive. The null hypothesis is that the mean difference score in the population is 0.

Table 1.2: DOG scores as a function of dosage.

D0	**D60**	**Diff**
57	62	5
27	49	22
32	30	-2
31	34	3
34	38	4
38	36	-2
71	77	6
33	51	18
34	45	11
53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

To test this null hypothesis, we compute t using a special case of the following formula:

$$\text{t} = \frac{\text{statistic-hypothesized value}}{\text{estimated standard error of the statistic}}$$

The special case of this formula applicable to testing a single mean is

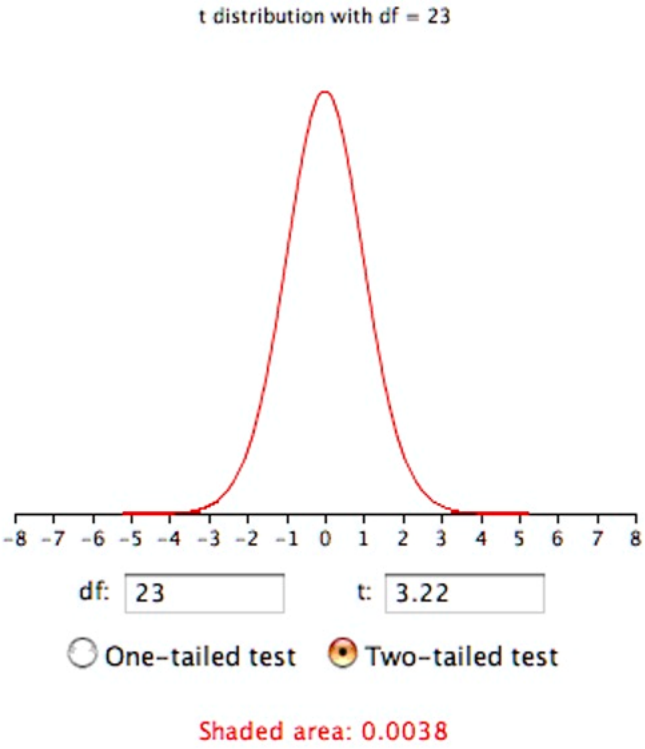
$$\text{t} = \frac{M - \mu}{S_M}$$

where t is the value we compute for the significance test, M is the sample mean, μ is the hypothesized value of the population mean, and s_M is the estimated standard error of the mean. Notice the similarity of this formula to the formula for Z we saw before.

In the previous example, we assumed that the scores were normally distributed. In this case, it is the population of difference scores that we assume to be normally distributed.

The mean (M) of the N = 24 difference scores is 4.958, the hypothesized value of μ is 0, and the standard deviation (s) is 7.538. The estimate of the standard error of the mean is computed as:

Therefore, t = 4.96/1.54 = 3.22. The probability value for t depends on the degrees of freedom. The number of degrees of freedom is equal to N - 1 = 23. As shown below, a t distribution calculator finds that the probability of a t less than -3.22 or greater than 3.22 is only 0.0038. Therefore, if the drug had no effect, the probability of finding a difference between means as large or larger (in either direction) than the difference found is very low. Therefore the null hypothesis that the population mean difference score is zero can be rejected. The conclusion is that the population mean for the drug condition is higher than the population mean for the placebo condition.



df: 23 t: 3.22

○ One-tailed test ● Two-tailed test

Shaded area: 0.0038

Figure 1.6: Calculation using the t distribution.

In order to conduct this hypothesis test, we made the following *assumptions*:

1. Each value is sampled independently from each other value.
2. The values are sampled from a normal distribution.

1.6 Type I and Type II Errors[12]

In the Physicians' Reactions case study,[13] the probability value associated with the significance test is 0.0057. Therefore, the null hypothesis was rejected, and it was concluded that physicians intend to spend less time with obese patients. Despite the low probability value, it is possible that the null hypothesis of no true difference between obese and average-weight patients is true and that the large difference between sample means occurred by chance. If this is the case, then the conclusion that physicians intend to spend less time with obese patients is in error. This type of error is called a Type I error. More generally, a **Type I error** occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05, then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α (alpha) level or simply α. It is also called the significance level. As discussed in the section on significance testing, it is better to interpret the probability value as an indication of the weight of evidence against the null hypothesis than as part of a decision rule for making a reject or do-not-reject decision. Therefore, keep in mind that rejecting the null hypothesis is not an all-or-nothing decision.

The Type I error rate is affected by the α level: the lower the α level, the lower the Type I error rate. It might seem that α is the probability of a Type I error. However, this is not correct. Instead, α is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error.

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a **Type II error**. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called (beta). The probability of correctly rejecting a false null hypothesis equals 1- and is called *statistical power*.

[1] This section is adapted from David M. Lane. “Introduction.” *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/intro.html

[2] http://onlinestatbook.com/2/calculators/binomial_dist.html

[3] http://onlinestatbook.com/2/case_studies/weight.html

[4] Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, *67*, 160-167.

[5] This section is adapted from David M. Lane. “Steps in Hypothesis Testing.” *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/steps.html

[6] This section is adapted from David M. Lane. “One- and Two-Tailed Tests.” *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/tails.html

[7] http://onlinestatbook.com/2/case_studies/bond.html

[8] This section is adapted from David M. Lane. “Significance Testing.” *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/significance.html

[9] http://onlinestatbook.com/2/case_studies/weight.html

[10] This section is adapted from David M. Lane. “Testing a Single Mean.” *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/tests_of_means/single_mean.html

[11] http://onlinestatbook.com/2/case_studies/adhd.html

[12] This section is adapted from David M. Lane. “Type I and Type II Errors.” *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/errors.html

[13] http://onlinestatbook.com/2/case_studies/weight.html

2 Comparing Means (How a Qualitative Variable Relates to a Quantitative Variable)

2.1 Difference between Two Means[1]

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. This section covers how to test for differences between means from two separate groups of subjects.

We take as an example the data from the "Animal Research" case study.[2] In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 2.1.

Table 2.1: Means and Variances in Animal Research study.

Group	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very important. What is important is whether there is a difference in the *population* means.

In order to test whether there is a difference between population means, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the scores are not independent.

One could look at these assumptions in much more detail, but suffice it to say that small-to-moderate violations of assumptions 1 and 2 do not make much difference. It is important not to violate assumption 3.

We saw the following general formula for significance testing in the section on testing a single mean:

$$\text{t} = \frac{\text{statistic-hypothesized value}}{\text{estimated standard error of the statistic}}$$

In this case, our statistic is the difference between sample means and our hypothesized value is 0. The hypothesized value is the null hypothesis that the difference between population means is 0.

We continue to use the data from the “Animal Research” case study and will compute a significance test on the difference between the mean score of the females and the mean score of the males. For this calculation, we will make the three assumptions specified above.

The first step is to compute the statistic, which is simply the difference between means.

$$M_1 - M_2 = 5.3529 - 3.8824 = 1.4705$$

Since the hypothesized value is 0, we do not need to subtract it from the statistic.

The next step is to compute the estimate of the standard error of the statistic. In this case, the statistic is the difference between means, so the estimated standard error of the statistic is $(S_{M1} - {}_{M_2})$. Recall from the relevant section in the chapter on sampling distributions that the formula for the standard error of the difference between means is:

$$\sigma_{M1-M2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

In order to estimate this quantity, we estimate ^{2} and use that estimate in place of ^{2}. Since we are assuming the two population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$\text{MSE} = \frac{s_1^2 + s_2^2}{2}$$

where MSE is our estimate of σ^2. In this example,

$$\text{MSE} = (2.743 + 2.985)/2 = 2.864.$$

Since n (the number of scores in each group) is 17,

$$S_{M1-M2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805$$

The next step is to compute t by plugging these values into the formula:

$$t = 1.4705/.5805 = 2.533.$$

Finally, we compute the probability of getting a t as large or larger than 2.533 or as small or smaller than -2.533. To do this, we need to know the degrees of freedom. The degrees of freedom is the number of independent estimates of variance on which MSE is based. This is equal to (n_1 - 1) + (n_2 - 1), where n_1 is the sample size of the first group and n_2 is the sample size of the second group. For this example, $n_1 = n_2 = 17$. When $n_1 = n_2$, it is conventional to use "n" to refer to the sample size of each group. Therefore, the degrees of freedom is 16 + 16 = 32.

Once we have the degrees of freedom, we can use a t distribution calculator[3] to find the probability. Figure 2.1 shows that the probability value for a two-tailed test is 0.0164. The two-tailed test is used when the null hypothesis can be rejected regardless of the direction of the effect. As shown in Figure 2.1, it is the probability of a t < -2.533 or a t > 2.533.

The results of a one-tailed test are shown in Figure 2.2. As you can see, the probability value of 0.0082 is half the value for the two-tailed test.

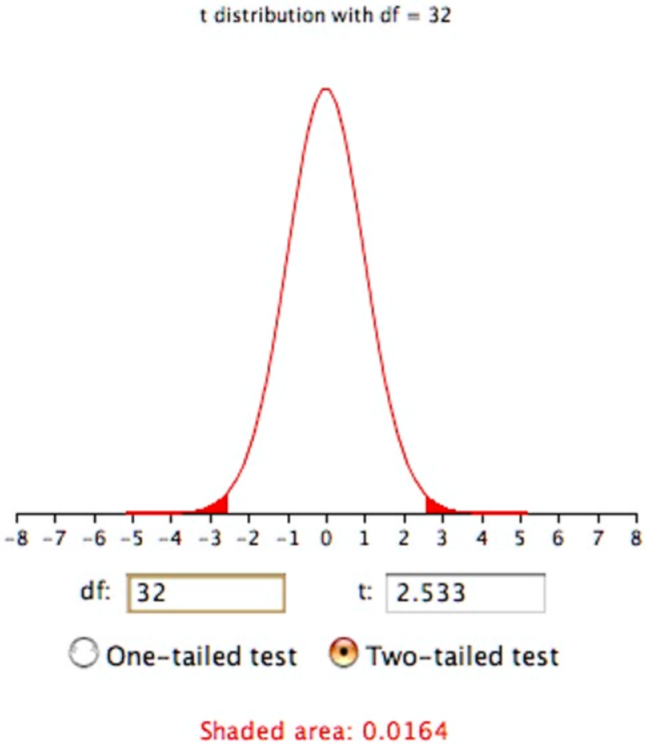
2.1.1 Formatting Data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table 2.2.

Table 2.2: Example data.

Group 1	**Group 2**
3	2
4	6
5	8

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 2.2 is shown in Table 2.3.

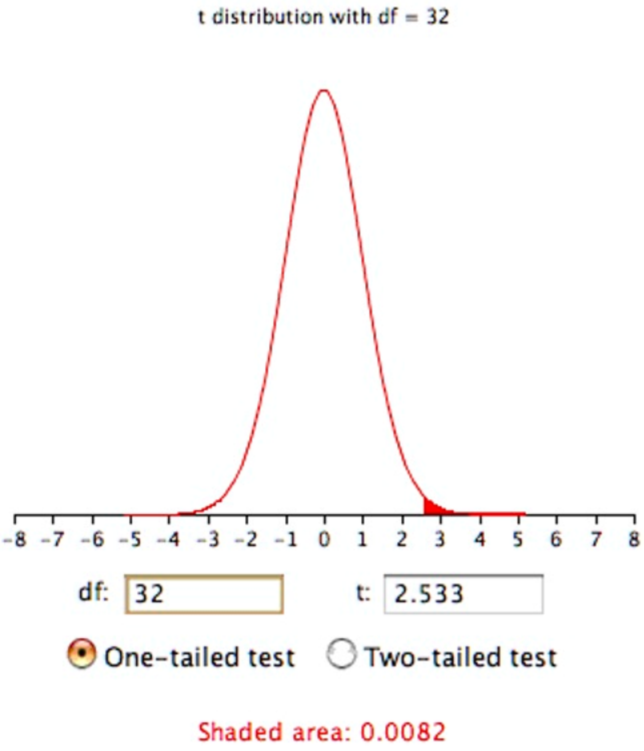


df: 32 t: 2.533

○ One-tailed test ● Two-tailed test

Shaded area: 0.0164

Figure 2.1: The two-tailed probability.



df: 32   t: 2.533

● One–tailed test ○ Two–tailed test

Shaded area: 0.0082

Figure 2.2: The one-tailed probability.

Table 2.3: Reformatted data.

G	**Y**
1	3
1	4
1	5
2	2
2	6
2	8

Using statistical software, we'd find that the t value is -0.718, the df = 4, and p = 0.512.

2.2 Pairwise Comparisons Among Multiple Means[4]

Many experiments are designed to compare more than two conditions. We will take as an example the case study "Smiles and Leniency."[5] In this study, the effect of different types of smiles on the leniency shown to a person was investigated. An obvious way to proceed would be to do a t test of the difference between each group mean and each of the other group means. This procedure would lead to the six comparisons shown in Table 2.4.

Table 2.4: Six Comparisons among Means.

felt vs. miserable		
felt vs. neutral		

miserable vs. neutral		

The problem with this approach is that if you did this analysis, you would have six chances to make a Type I error. Therefore, if you were using the 0.05 significance level, the probability that you would make a Type I error on at least one of these comparisons is greater than 0.05. The more means that are compared, the more the Type I error rate is inflated. Figure 2.3 shows the number of possible comparisons between pairs of means (pairwise comparisons) as a function of the number of means. If there are only two means, then only one comparison can be made. If there are 12 means, then there are 66 possible comparisons.

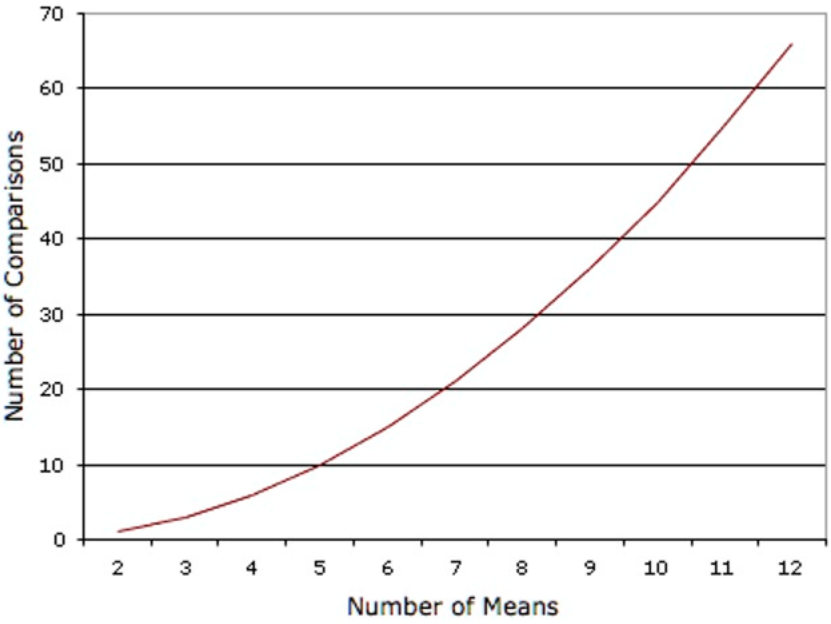


Figure 2.3: Number of pairwise comparisons as a function of the number of means.

Figure 2.4 shows the probability of a Type I error as a function of the number of means. As you can see, if you have an experiment with 12 means, the probability is about 0.70 that at least one of the 66 comparisons among means would be significant even if all 12 population means were the same.

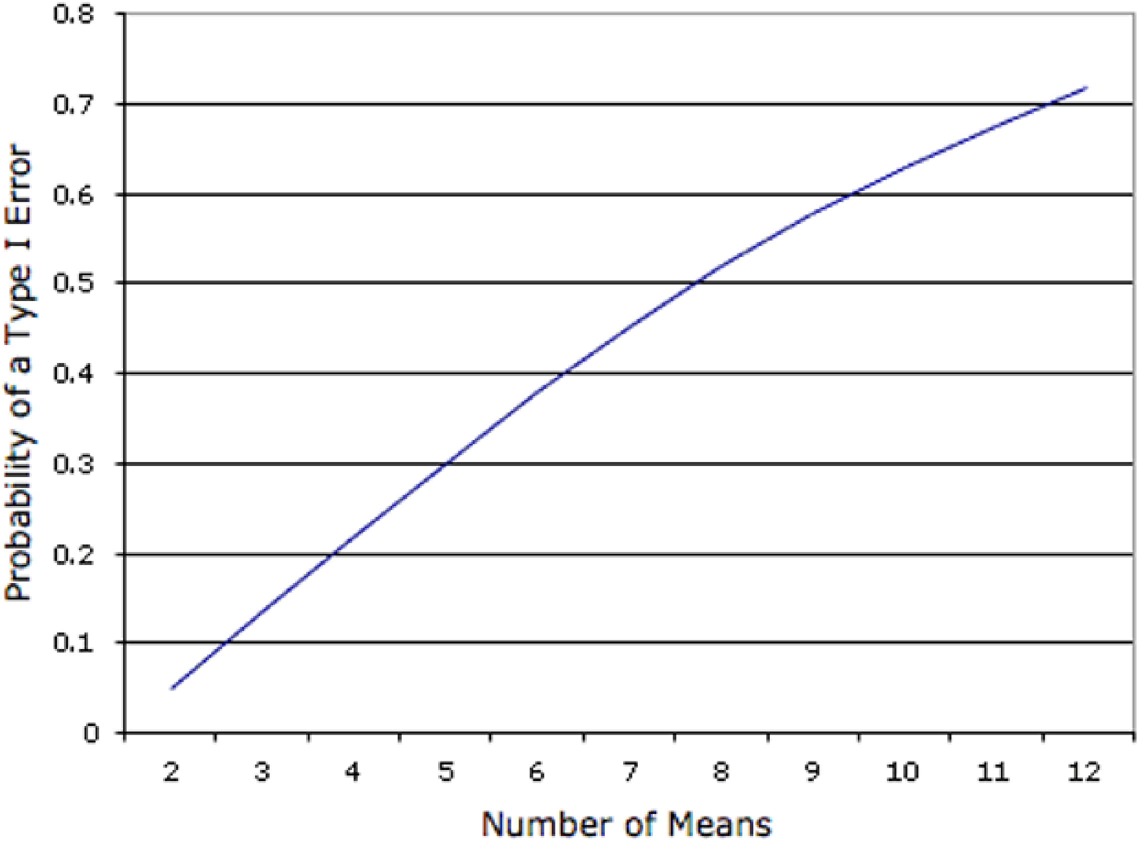


Figure 2.4: Probability of a Type I error as a function of the number of means.

The Type I error rate can be controlled using a test called the Tukey Honestly Significant Difference test or Tukey HSD for short. The Tukey HSD test is one example of a multiple comparison test, but several alternatives are frequently used, such as the Bonferroni correction. Regardless of the exact method used for a multiple comparison test, the interpretation of results is similar. The Tukey HSD is based on a variation of the t distribution that takes into account the number of means being compared. This distribution is called the studentized range distribution.

Normally, statistical software will make all the necessary calculations for you in the background. But to illustrate what sorts of calculations the software is relying on, let's return to the leniency study to see how to compute the Tukey HSD test. You will see that the computations are very similar to those of an independent-groups t test. The steps are outlined below:

1. Compute the means and variances of each group. They are shown below.

Condition	**Mean**	**Variance**
False	5.37	3.34
Felt	4.91	2.83
Miserable	4.91	2.11
Neutral	4.12	2.32

2. Compute MSE, which is simply the mean of the variances. It is equal to 2.65.
3. Compute

$$Q = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n}}}$$

for each pair of means, where M_i is one mean, M_j is the other mean, and n is the number of scores in each group. For these data, there are 34 observations per group. The value in the denominator is 0.279.

4. Compute p for each comparison using a Studentized Range Calculator.[6] The degrees of freedom is equal to the total number of observations minus the number of means. For this experiment, df = 136 - 4 = 132.

The tests for these data are shown in Table 2.6.

Table 2.6: Six Pairwise Comparisons.

Comparison	**M_i-M_j**	**Q**	**p**
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649

False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

The only significant comparison is between the false smile and the neutral smile.

It is not unusual to obtain results that on the surface appear paradoxical. For example, these results appear to indicate that (a) the false smile is the same as the miserable smile, (b) the miserable smile is the same as the neutral control, and (c) the false smile is different from the neutral control. This apparent contradiction is avoided if you are careful not to accept the null hypothesis when you fail to reject it. The finding that the false smile is not significantly different from the miserable smile does not mean that they are really the same. Rather it means that there is not convincing evidence that they are different. Similarly, the non-significant difference between the miserable smile and the control does not mean that they are the same. The proper conclusion is that the false smile is higher than the control and that the miserable smile is either (a) equal to the false smile, (b) equal to the control, or (c) somewhere in-between.

The assumptions of the Tukey test are essentially the same as for an independent-groups t test: normality, homogeneity of variance, and independent observations. The test is quite robust to violations of normality. Violating homogeneity of variance can be more problematical than in the two-sample case since the MSE is based on data from all groups. The assumption of independence of observations is important and should not be violated.

2.2.1 Computer Analysis

For most computer programs, you should format your data the same way you do for an independent-groups t test. The only difference is that if you have, say, four groups, you would code each group as 1, 2, 3, or 4 rather than just 1 or 2.

2.2.2 Tukey's Test Need Not be a Follow-Up to ANOVA

Some textbooks introduce the Tukey test only as a follow-up to an analysis of variance. There is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA (or even know what one is). If you or your instructor do not wish to take our word for this, see the excellent article on this and other issues in statistical analysis by Leland Wilkinson and the APA Board of Scientific Affairs' Task Force on Statistical Inference, published in the *American Psychologist*, August 1999, Vol. *54*, No. 8, 594–604.

2.3 Analysis of Variance (ANOVA)[7]

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called "Analysis of Variance" rather than "Analysis of Means." As you will see, the name is appropriate because inferences about means are made by analyzing variance.

ANOVA is used to test general rather than specific differences among means. This can be seen best by example. In the case study "Smiles and Leniency,"[8] the effect of different types of smiles on the leniency shown to a person was investigated. Four different types of smiles (neutral, false, felt, miserable) were investigated. In the prior section, we learned how to test differences among means. The results from the Tukey HSD test are shown in Table 2.7.

Table 2.7: Six Pairwise Comparisons.

Comparison	**M_i-M_j**	**Q**	**p**
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

Notice that the only significant difference is between the False and Neutral conditions.

ANOVA tests the non-specific null hypothesis that all four population means are equal. That is,

$$\mu_{false} = \mu_{felt} = \mu_{miserable} = \mu_{neutral}$$

This non-specific null hypothesis is sometimes called the omnibus null hypothesis. When the omnibus null hypothesis is rejected, the conclusion is that at least one population mean is different from at least one other mean. However, since the ANOVA does not reveal which means are different from which, it offers less specific information than the Tukey HSD test. The Tukey HSD is therefore preferable to ANOVA in this situation. Some textbooks introduce the Tukey test only as a follow-up to an ANOVA. However, there is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA.

You might be wondering why you should learn about ANOVA when the Tukey test is better. One reason is that there are complex types of analyses that can be done with ANOVA and not with the Tukey test. A second is that ANOVA is by far the most commonly-used technique for comparing means, and it is important to understand ANOVA in order to understand research reports.

[1] This section is adapted from David M. Lane. "Difference between Two Means (Independent Groups)." *Online Statistics Education: A Multimedia Course of Study*. http://onlinestatbook.com/2/tests_of_means/difference_means.html

[2] http://onlinestatbook.com/2/case_studies/animal_research.html

[3] http://onlinestatbook.com/2/calculators/t_dist.html

[4] This section is adapted from David M. Lane. "All Pairwise Comparisons Among Means." *Online Statistics Education: A Multimedia Course of Study*. http://onlinestatbook.com/2/tests_of_means/pairwise.html

[5] http://onlinestatbook.com/2/case_studies/leniency.html

[6] http://onlinestatbook.com/2/calculators/studentized_range_dist.html

[7] This section is adapted from David M. Lane. "Introduction." *Online Statistics Education: A Multimedia Course of Study*. http://onlinestatbook.com/2/analysis_of_variance/intro.html

[8] http://onlinestatbook.com/2/case_studies/leniency.html

3 Comparing Groups (How Two Qualitative Variables Relate to One Another)

3.1 Chi Square Distribution[1]

A standard normal deviate is a random sample from the standard normal distribution. The Chi Square distribution is the distribution of the sum of squared standard normal deviates. The degrees of freedom of the distribution is equal to the number of standard normal deviates being summed. Therefore, Chi Square with one degree of freedom, written as $^{2}(1)$, is simply the distribution of a single normal deviate squared. The area of a Chi Square distribution below 4 is the same as the area of a standard normal distribution below 2, since 4 is 2^2.

Consider the following problem: you sample two scores from a standard normal distribution, square each score, and sum the squares. What is the probability that the sum of these two squares will be six or higher? Since two scores are sampled, the answer can be found using the Chi Square distribution with two degrees of freedom. A Chi Square calculator can be used to find that the probability of a Chi Square (with 2 df) being six or higher is 0.050.

The mean of a Chi Square distribution is its degrees of freedom. Chi Square distributions are positively skewed, with the degree of skew decreasing with increasing degrees of freedom. As the degrees of freedom increases, the Chi Square distribution approaches a normal distribution. Figure 3.1 shows density functions for three Chi Square distributions. Notice how the skew decreases as the degrees of freedom increases.

The Chi Square distribution is very important because many test statistics are approximately distributed as Chi Square. Two of the more common tests using the Chi Square distribution are tests of deviations of differences between theoretically expected and observed frequencies (one-way tables) and the relationship between categorical variables (contingency tables). Numerous other tests beyond the scope of this work are based on the Chi Square distribution.

3.2 One-Way Tables[2]

The Chi Square distribution can be used to test whether observed data differ *significantly* from theoretical expectations. For example, for a fair six-sided die, the probability of any given outcome on a single roll would be 1/6. The data in Table 3.1 were obtained by rolling

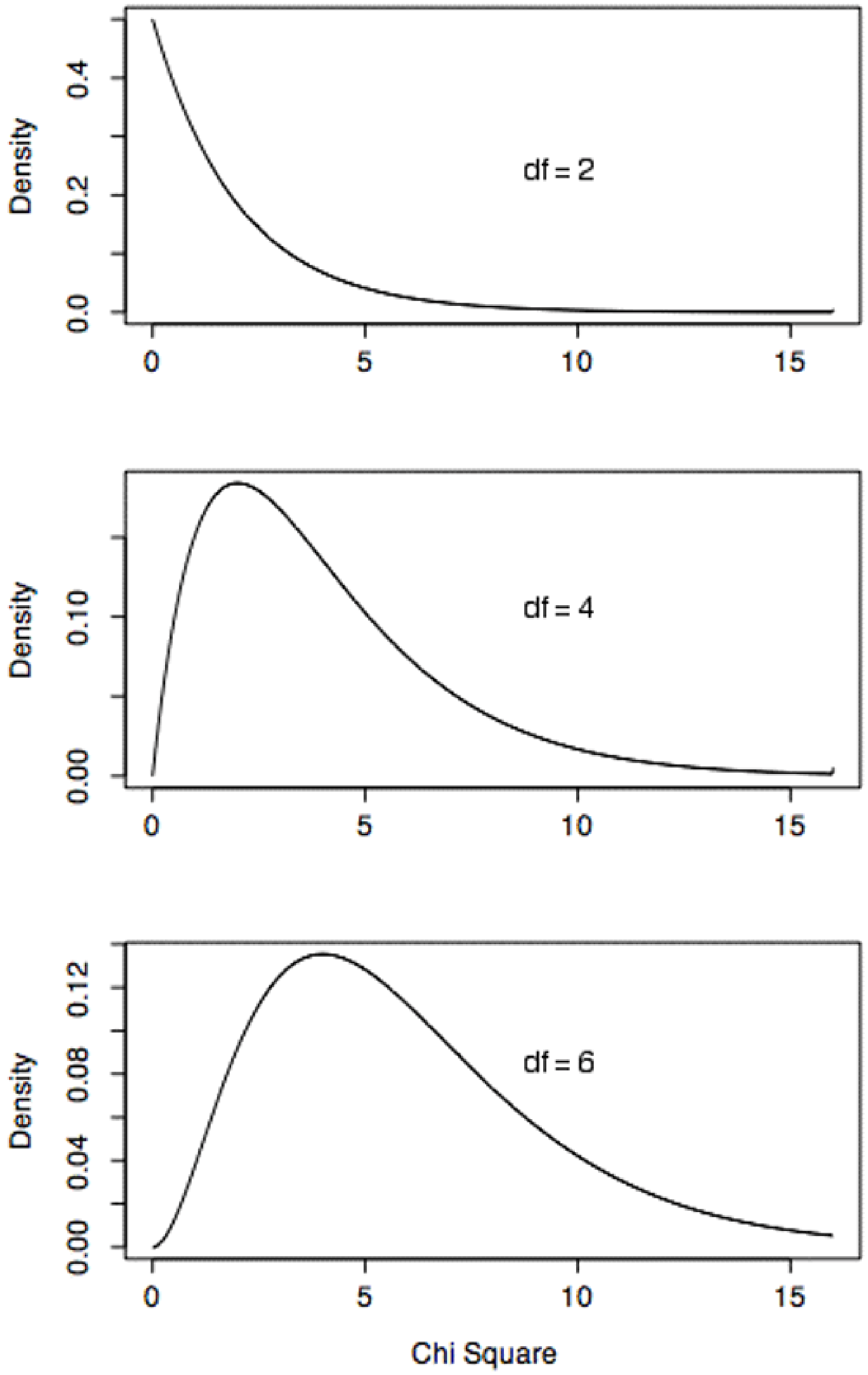


Figure 3.1: Chi Square distributions with 2, 4, and 6 degrees of freedom.

a six-sided die 36 times. However, as can be seen in Table 3.1, some outcomes occurred more frequently than others. For example, a "3" came up nine times, whereas a "4" came up only two times. Are these data consistent with the hypothesis that the die is a fair die? Naturally, we do not expect the sample frequencies of the six possible outcomes to be the same since chance differences will occur. So, the finding that the frequencies differ does not mean that the die is not fair. One way to test whether the die is fair is to conduct a significance (hypothesis) test. The null hypothesis is that the die is fair. This hypothesis is tested by computing the probability of obtaining frequencies as discrepant or more discrepant from a uniform distribution of frequencies as obtained in the sample. If this probability is sufficiently low, then the null hypothesis that the die is fair can be rejected.

Table 3.1: Outcome Frequencies from a Six-Sided Die.

Outcome	**Frequency**
1	8
2	5
3	9
4	2
5	7
6	5

The first step in conducting the significance test is to compute the expected frequency for each outcome given that the null hypothesis is true. For example, the expected frequency of a "1" is 6 since the probability of a "1" coming up is 1/6 and there were a total of 36 rolls of the die.

Expected frequency = (1/6)(36) = 6

Note that the expected frequencies are expected only in a theoretical sense. We do not really "expect" the observed frequencies to match the "expected frequencies" exactly.

The calculation continues as follows. Letting E be the expected frequency of an outcome and O be the observed frequency of that outcome, compute

$$\frac{(E-O)^2}{E}$$

for each outcome. Table 3.2 shows these calculations.

Table 3.2: Outcome Frequencies from a Six-Sided Die.

Outcome	**E**	**O**	$\frac{(E-O)^2}{E}$

1	6	8	0.667
2	6	5	0.167
3	6	9	1.500
4	6	2	2.667
5	6	7	0.167
6	6	5	0.167

Next we add up all the values in Column 4 of Table 3.2.

$$\Sigma\frac{(E-O)^2}{E} = 5.333$$

This sampling distribution of

$$\Sigma\frac{(E-O)^2}{E}$$

is approximately distributed as Chi Square with k-1 degrees of freedom, where k is the number of categories. Therefore, for this problem the test statistic is

$$X_5^2 = 5.333$$

which means the value of Chi Square with 5 degrees of freedom is 5.333.

From a Chi Square calculator[3] it can be determined that the probability of a Chi Square of 5.333 or larger is 0.377. Therefore, the null hypothesis that the die is fair cannot be rejected.

This Chi Square test can also be used to test other deviations between expected and observed frequencies. The following example shows a test of whether the variable "University GPA" in the SAT and College GPA case study is normally distributed.

The first column in Table 3.3 shows the normal distribution divided into five ranges. The second column shows the proportions of a normal distribution falling in the ranges specified in the first column. The expected frequencies (E) are calculated by multiplying the number of scores (105) by the proportion. The final column shows the observed number of scores in each range. It is clear that the observed frequencies vary greatly from the expected frequencies. Note that if the distribution were normal, then there would have been only about 35 scores between 0 and 1, whereas 60 were observed.

Table 3.3: Expected and Observed Scores for 105 University GPA Scores.

Range	**Proportion**	**E**	**O**

Above 1	0.159	16.695	9
0 to 1	0.341	35.805	60
-1 to 0	0.341	35.805	17
Below -1	0.159	16.695	19

The test of whether the observed scores deviate significantly from the expected scores is computed using the familiar calculation.

$$X_3^2 = \Sigma \frac{(E-O)^2}{E} = 30.09$$

The subscript “3” means there are three degrees of freedom. As before, the degrees of freedom is the number of outcomes minus 1, which is 4 - 1 = 3 in this example. A Chi Square distribution calculator shows that $p < 0.001$ for this Chi Square. Therefore, the null hypothesis that the scores are normally distributed can be rejected.

3.3 Contingency Tables[4]

This section shows how to use Chi Square to test the relationship between nominal variables for significance. For example, Table 3.4 shows the data from the Mediterranean Diet and Health case study.[5]

Table 3.4: Frequencies for Diet and Health Study.

		Outcome			
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

The question is whether there is a *significant relationship* between diet and outcome. Again, software can calculate a p-value for us in order to test for significance. But if we are wondering what’s going on under the hood, the first step is to compute the expected frequency for each cell based on the assumption that there is no relationship. These expected frequencies are computed from the totals as follows. We begin by computing the expected frequency for the

AHA Diet/Cancers combination. Note that 22/605 subjects developed cancer. The proportion who developed cancer is therefore 0.0364. If there were no relationship between diet and outcome, then we would expect 0.0364 of those on the AHA diet to develop cancer. Since 303 subjects were on the AHA diet, we would expect (0.0364)(303) = 11.02 cancers on the AHA diet. Similarly, we would expect (0.0364)(302) = 10.98 cancers on the Mediterranean diet. In general, the expected frequency for a cell in the ith row and the jth column is equal to

$$E_{ij} = \frac{T_i T_j}{T}$$

where E_{ij} is the expected frequency for cell i,j, Ti is the total for the ith row, Tj is the total for the jth column, and T is the total number of observations. For the AHA Diet/Cancers cell, i = 1, j = 1, Ti = 303, Tj = 22, and T = 605. Table 3.5 shows the expected frequencies (in parenthesis) for each cell in the experiment.

Table 3.5: Observed and Expected Frequencies for Diet and Health Study.

	Outcome			
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy
AHA	15	24	25	239
	(11.02)	(19.03)	(16.53)	(256.42)
Mediterranean	7	14	8	273
	(10.98)	(18.97)	(16.47)	(255.58)
Total	22	38	33	512

The significance test is conducted by computing Chi Square as follows.

$$X_3^2 = \Sigma \frac{(E-O)^2}{E} = -16.55$$

The degrees of freedom is equal to (r-1)(c-1), where r is the number of rows and c is the number of columns. For this example, the degrees of freedom is (2-1)(4-1) = 3. The Chi Square calculator[6] can be used to determine that the probability value for a Chi Square of 16.55 with three degrees of freedom is equal to 0.0009. Therefore, the null hypothesis of no relationship between diet and outcome can be rejected.

A key assumption of this Chi Square test is that each subject contributes data to only one cell. Therefore, the sum of all cell frequencies in the table must be the same as the number of subjects in the experiment. Consider an experiment in which each of 16 subjects attempted two anagram problems. The data are shown in Table 3.6.

Table 3.6: Anagram Problem Data.

	Anagram 1	**Anagram 2**
Solved	10	4
Did not Solve	6	12

It would not be valid to use the Chi Square test on these data since each subject contributed data to two cells: one cell based on their performance on Anagram 1 and one cell based on their performance on Anagram 2. The total of the cell frequencies in the table is 32, but the total number of subjects is only 16.

[1] This section is adapted from David M. Lane. "Chi Square Distribution." *Online Statistics Education: A Multimedia Course of Study*. http://onlinestatbook.com/2/chi_square/distribution.html

[2] This section is adapted from David M. Lane. "One-Way Tables (Testing Goodness of Fit)." *Online Statistics Education: A Multimedia Course of Study*. http://onlinestatbook.com/2/chi_square/one-way.html

[3] http://onlinestatbook.com/2/calculators/chi_square_prob.html

[4] This section is adapted from David M. Lane. "Contingency Tables." *Online Statistics Education: A Multimedia Course of Study*. http://onlinestatbook.com/2/chi_square/contingency.html

[5] http://onlinestatbook.com/2/case_studies/diet.html

[6] http://onlinestatbook.com/2/calculators/chi_square_prob.html

4 Causality

4.1 Causation[1]

The concept of causation is a complex one in the philosophy of science.[2] Since a full coverage of this topic is well beyond the scope of this text, we focus on two specific topics: (1) the establishment of causation in experiments and (2) the establishment of causation in non-experimental designs.

4.1.1 Establishing Causation in Experiments

Consider a simple experiment in which subjects are sampled randomly from a population and then assigned randomly to either the experimental group or the control group. Assume the condition means on the dependent variable differed. Does this mean the treatment caused the difference?

To make this discussion more concrete, assume that the experimental group received a drug for insomnia, the control group received a placebo, and the dependent variable was the number of minutes the subject slept that night. An obvious obstacle to inferring causality is that there are many unmeasured variables that affect how many hours someone sleeps. Among them are how much stress the person is under, physiological and genetic factors, how much caffeine they consumed, how much sleep they got the night before, etc. Perhaps differences between the groups on these factors are responsible for the difference in the number of minutes slept.

At first blush it might seem that the random assignment eliminates differences in unmeasured variables. However, this is not the case. Random assignment ensures that differences on unmeasured variables are chance differences. It does not ensure that there are no differences. Perhaps, by chance, many subjects in the control group were under high stress and this stress made it more difficult to fall asleep. The fact that the greater stress in the control group was due to chance does not mean it could not be responsible for the difference between the control and the experimental groups. In other words, the observed difference in “minutes slept” could have been due to a chance difference between the control group and the experimental group rather than due to the drug’s effect.

[1] This section is adapted from David M. Lane. “Causation.” *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/research_design/causation.html

[2] See http://plato.stanford.edu/search/searcher.py?query=causation

This problem seems intractable since, by definition, it is impossible to measure an "unmeasured variable" just as it is impossible to measure and control all variables that affect the dependent variable. However, although it is impossible to assess the effect of any single unmeasured variable, it is possible to assess the combined effects of all unmeasured variables. Since everyone in a given condition is treated the same in the experiment, differences in their scores on the dependent variable must be due to the unmeasured variables. Therefore, a measure of the differences among the subjects within a condition is a measure of the sum total of the effects of the unmeasured variables. The most common measure of differences is the variance. By using the within-condition variance to assess the effects of unmeasured variables, statistical methods determine the probability that these unmeasured variables could produce a difference between conditions as large or larger than the differenc[3]e obtained in the experiment. If that probability is low, then it is inferred (that's why they call it inferential statistics) that the treatment had an effect and that the differences are not entirely due to chance. Of course, there is always some nonzero probability that the difference occurred by chance so total certainty is not a possibility.

4.1.2 Causation in Non-Experimental Designs

It is almost a cliché that correlation does not mean causation. The main fallacy in inferring causation from correlation is called the third variable problem and means that a third variable is responsible for the correlation between two other variables. An excellent example used by Li (1975) to illustrate this point is the positive correlation in Taiwan in the 1970's between the use of contraception and the number of electric appliances in one's house. Of course, using contraception does not induce you to buy electrical appliances or vice versa. Instead, the third variable of education level affects both.

Does the possibility of a third-variable problem make it impossible to draw causal inferences without doing an experiment? One approach is to simply assume that you do not have a third-variable problem. This approach, although common, is not very satisfactory. However, be aware that the assumption of no third-variable problem may be hidden behind a complex causal model that contains sophisticated and elegant mathematics.

A better though, admittedly more difficult approach, is to find converging evidence. This was the approach taken to conclude that smoking causes cancer. The analysis included converging evidence from retrospective studies, prospective studies, lab studies with animals, and theoretical understandings of cancer causes.

A second problem is determining the direction of causality. A correlation between two variables does not indicate which variable is causing which. For example, Reinhart and Rogoff (2010)[4] found a strong correlation between public debt and GDP growth. Although some have

[3] Li, C. (1975) *Path analysis: A primer.* Boxwood Press, Pacific Grove, CA.

[4] Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a Time of Debt. Working Paper 15639, National Bureau of Economic Research, http://www.nber.org/papers/w15639

argued that public debt slows growth, most evidence supports the alternative that slow growth increases public debt.[5]

4.2 Experimental Designs[6]

There are many ways an experiment can be designed. For example, subjects can all be tested under each of the treatment conditions or a different group of subjects can be used for each treatment. An experiment might have just one independent variable or it might have several. This section describes basic experimental designs and their advantages and disadvantages.

4.2.1 Between-Subjects Designs

In a **between-subjects** design, the various experimental treatments are given to different groups of subjects. For example, in the "Teacher Ratings"[7] case study, subjects were randomly divided into two groups. Subjects were all told they were going to see a video of an instructor's lecture after which they would rate the quality of the lecture. The groups differed in that the subjects in one group were told that prior teaching evaluations indicated that the instructor was charismatic whereas subjects in the other group were told that the evaluations indicated the instructor was punitive. In this experiment, the independent variable is "Condition" and has two levels (charismatic teacher and punitive teacher). It is a between-subjects variable because different subjects were used for the two levels of the independent variable: subjects were in either the "charismatic teacher" or the "punitive teacher" condition. Thus the comparison of the charismatic-teacher condition with the punitive-teacher condition is a comparison between the subjects in one condition with the subjects in the other condition.

The two conditions were treated exactly the same except for the instructions they received. Therefore, it would appear that any difference between conditions should be attributed to the treatments themselves. However, this ignores the possibility of chance differences between the groups. That is, by chance, the raters in one condition might have, on average, been more lenient than the raters in the other condition. Randomly assigning subjects to treatments ensures that all differences between conditions are chance differences; it does not ensure there will be no differences. The key question, then, is how to distinguish real differences from chance differences. The field of inferential statistics answers just this question. Analyzing the data from this experiment reveals that the ratings in the charismatic-teacher condition were higher than those in the punitive-teacher condition. Using inferential statistics, it can be calculated that the probability of finding a difference as large or larger than the one obtained

[5] For a video on causality featuring evidence that smoking causes cancer, see http://www.learner.org/resources/series65.html

[6] This section is adapted from David M. Lane. "Experimental Designs." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/research_design/designs.html

[7] http://onlinestatbook.com/2/case_studies/ratings.html

if the treatment had no effect is only 0.018. Therefore it seems likely that the treatment had an effect and it is not the case that all differences were chance differences.

Independent variables often have several levels. For example, in the "Smiles and Leniency" case study the independent variable is "type of smile" and there are four levels of this independent variable: (1) false smile, (2) felt smile, (3) miserable smile, and (4) a neutral control. Keep in mind that although there are four levels, there is only one independent variable. Designs with more than one independent variable are considered next.

4.2.2 Multi-Factor Between-Subject Designs

In the "Bias Against Associates of the Obese"[8] experiment, the qualifications of potential job applicants were judged. Each applicant was accompanied by an associate. The experiment had two independent variables: the weight of the associate (obese or average) and the applicant's relationship to the associate (girl friend or acquaintance). This design can be described as an Associate's Weight (2) x Associate's Relationship (2) factorial design. The numbers in parentheses represent the number of levels of the independent variable. The design was a factorial design because all four combinations of associate's weight and associate's relationship were included. The dependent variable was a rating of the applicant's qualifications (on a 9-point scale).

If two separate experiments had been conducted, one to test the effect of Associate's Weight and one to test the effect of Associate's Relationship then there would be no way to assess whether the effect of Associate's Weight depended on the Associate's Relationship. One might imagine that the Associate's Weight would have a larger effect if the associate were a girl friend rather than merely an acquaintance. A factorial design allows this question to be addressed. When the effect of one variable does differ depending on the level of the other variable then it is said that there is an interaction between the variables.

Factorial designs can have three or more independent variables. In order to be a between-subjects design there must be a separate group of subjects for each combination of the levels of the independent variables.

4.2.3 Within-Subjects Designs

A **within-subjects** design differs from a between-subjects design in that the same subjects perform at all levels of the independent variable. For example consider the "ADHD Treatment"[9] case study. In this experiment, subjects diagnosed as having attention deficit disorder were each tested on a delay of gratification task after receiving methylphenidate (MPH). All

[8] http://onlinestatbook.com/2/case_studies/obesity_relation.html
[9] http://onlinestatbook.com/2/case_studies/adhd.html

subjects were tested four times, once after receiving one of the four doses. Since each subject was tested under *each* of the four levels of the independent variable "dose," the design is a within-subjects design and dose is a within-subjects variable. Within-subjects designs are sometimes called repeated-measures designs.

4.2.4 Advantage of Within-Subjects Designs

An advantage of within-subjects designs is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ greatly from one another. In an experiment on problem solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control. This typically gives within-subjects designs considerably more power (ability to find precise estimates) than between-subjects designs. That is, this makes within-subjects designs more able to detect an effect of the independent variable than are between-subjects designs.

Within-subjects designs are often called "repeated-measures" designs since repeated measurements are taken for each subject. Similarly, a within-subject variable can be called a repeated-measures factor.

4.2.5 Complex Designs

Designs can contain combinations of between-subject and within-subject variables. For example, the "Weapons and Aggression"[10] case study has one between-subject variable (gender) and two within-subject variables (the type of priming word and the type of word to be responded to).

[10] http://onlinestatbook.com/2/case_studies/guns.html

5 Models and Uncertainty

Before I leave my house each morning, I need to decide whether to take an umbrella. So I check my phone to see whether it's supposed to rain. Instead of giving me a direct yes or no answer, the weather tells me the percentage chance of rain for the day.

Why does the weather app give me a percentage? Because there's uncertainty. Science has done a lot to help us understand the weather. And as our understanding of the weather improves, our predictions get better. But we still can't predict rain perfectly.

Facing uncertainty is a common problem when we're looking at data. Whether we're trying to explain the weather, human behavior, or even plant growth, we can't make perfect predictions because there are things we can't fully explain with our current scientific knowledge.

In statistics, we have several tools that allow us to acknowledge uncertainty. This enables us to build models like the ones powering my weather app—models that give us a prediction that includes a description of how uncertain we are. Some days we are 100% sure it will rain, other days only 60%.

In order to build these models that acknowledge uncertainty, we need a way to talk about what we do know and what we don't know. Let me give a very simple example of a model that accounts for uncertainty:

$$happiness = 3.0 + 2.3 \times income + \varepsilon \tag{5.1}$$

This model attempts to explain one's level of happiness based on their income. You might notice that it looks very similar to the regression equations we saw in Chapter 3. That's because regression is one of the main tools used to estimate a model that includes uncertainty.

What does this model mean in practical terms? Well, there are no obvious units we can use to quantify the amount of happiness someone experiences, so the exact values of the numbers we see are not particularly meaningful. But the fact that there's a positive number (2.3) that is being multiplied by income implies that as income gets bigger, happiness gets larger.

The key part of this equation that I want to focus on is the little Greek letter at the end of the equation: ε. This letter is called "epsilon," and it is often used to represent what we call an **error term** (also sometimes called a **disturbance term**). The error term (ε) represents everything else besides income that affects happiness. By including an error term, we are acknowledging that we can't perfectly predict one's level of happiness based on their income.

We think that knowing one's income will help us predict their happiness, but we know there are other factors we won't be able to measure or identify that will also affect happiness. Thus, if all we know about someone is their income, we will have uncertainty about their exact level of happiness. By including an error term (ε) in the model, we make clear that we only claim to have a partial understanding of happiness, not a complete one.

Think for a moment about how few topics we could study if we didn't have the freedom to build models that include uncertainty. We'd only be able to build a model of a dependent variable after we had identified (and measured) *all* of the factors that affect that variable! We wouldn't be able to build a model of rain since we don't know all of the factors that affect the rain. We couldn't build a model of voting behavior since we don't know everything that affects how someone will vote. By including an error term in our model, we can build models even when our understanding of something is incomplete.

The first part of our model that appears on the right side of the equation ($3.0 + 2.3 \times income$) is sometimes described as the *systematic* part of our model. It's what we would use to build a prediction of happiness if all we know about some is their income level. Suppose, for example, that someone has an income of 4 units (perhaps income is measured in tens of thousands of dollars of annual income, so a salary of $40,000 is coded as a 4). According to our model, that person's happiness would be:

$$happiness = 3.0 + 2.3 \times (4) + \varepsilon$$

$$happiness = 12.2 + \varepsilon$$

We, therefore, predict that someone with an income of 4 will have a happiness of 12.2, but we also acknowledge that their actual happiness will likely be a bit different from our prediction since our model indicates that their actual happiness will equal 12.2 plus the value of the error term (ε).

The error term describes something unknown, so we can't measure it or directly observe it. But what we can do is talk about its characteristics using concepts from probability theory. Specifically, we're going to describe the value of the error term as being randomly selected. You may have dealt with randomness in math classes before using examples such as coin flips, die rolls, or drawing cards from a 52-card deck. Just as the likelihood of different outcomes from parlor games can be described using probability, we're going to use probability to describe different possible values for the error term of a statistical model.

5.1 Assumptions about error terms

It's easy to write out an equation that includes an error term, but we are not going to be able to do much with our model unless we make some assumptions about the error term. One of the most important (and challenging) parts of doing statistical analysis is making assumptions about the possible values of the error term. Different assumptions about the error term can result in very different conclusions.

Let's again consider the simple model of happiness that was introduced above:

$$happiness = 3.0 + 2.3 \times income + \varepsilon$$

We might assume the following things about the error term (ε):

1. The values of the error term (ε) can be described by a normal distribution with a mean of 0
2. Knowing someone's income doesn't help us predict the values of the error term (ε)

What do these two assumptions mean?

First, if the error term (ε) follows a normal distribution with a mean of zero, that means that (according to our model), people are just as likely to have a positive value of the error term as they are to have a negative value of the error term. In other words, all those factors we haven't accounted for in our model are equally likely to push people in the direction of being happier or in the direction of being less happy. Our model and assumptions tell us that if we predict happiness purely based on income, we'll *overestimate* some people's happiness, and we'll *underestimate* an equal number of people's happiness.[1]

Second, these assumptions allow us to describe how much individual observations will tend to deviate from our income-based predictions. We haven't specified in our assumptions what the standard deviation is for the normal distribution for the error term (ε), but statistical analysis will let us estimate the standard deviation of an error term. And we know that there is a 95% chance of drawing a value within two standard deviations of the mean for any normal distribution. So whatever the standard deviation of the error term (ε) is, we would expect that 95% of the time, the error term will take on a value that is within two standard deviations of zero. Conversely, 5% of the time, the error term will take on a value that is more than two standard deviations from zero. Suppose that the standard deviation of the error term (ε) happens to be three. If we have a dataset containing the income and happiness of 1,000 randomly selected people, we would expect that about 950 of these people will have a level of happiness that falls within six units of our income-based prediction. But for about 50 of these people, our prediction of their happiness will be off by more than six units.

Third, our assumptions imply that income is not tied in any consistent way to (the total sum of) factors other than income that also affect peoples' happiness. Remember, the error term (ε) represents all factors other than income that affect satisfaction. If income is related to these other factors, then the value of income should help us predict the value of the error term. For example, if having a stable environment in childhood tends to cause both higher incomes

[1] Note that these deviations from our prediction don't imply that our model is wrong; our model explicitly acknowledges that we'll get only imperfect estimates if we predict happiness based on income, since the unobserved error term (ε) also contributes to happiness.

and greater happiness in adulthood, the error term will partially reflect the effect of childhood stability on happiness, so high incomes (which are partially caused by childhood stability) will be probably be predictive of a more positive error term. This would constitute a violation of our assumptions since we specifically indicated that income wasn't predictive of the error term. As this example illustrates, our assumptions about error terms are often quite strict, making it rather difficult in practice to build good models that account for uncertainty.

5.2 Models and probabilistic thinking

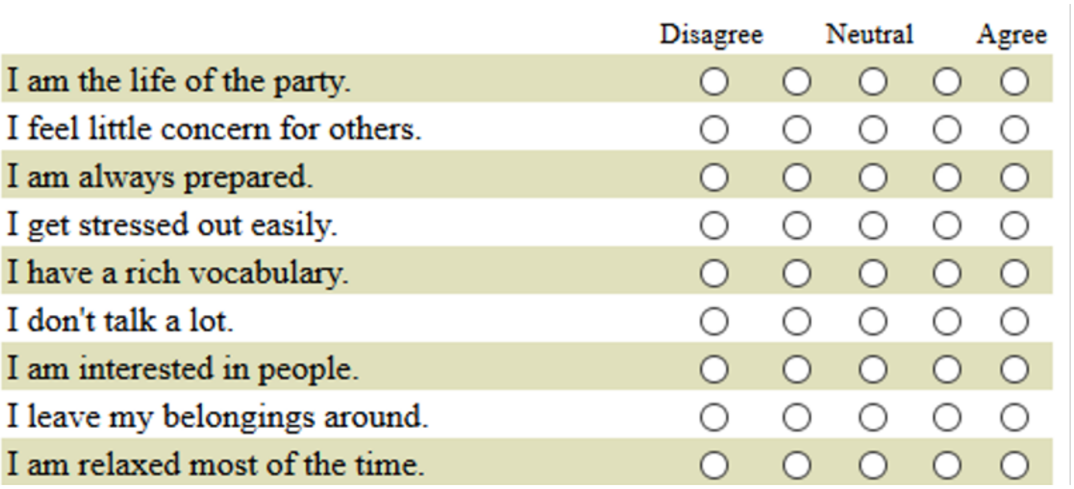
Despite the difficulty inherent in building models that accommodate uncertainty, we have little alternative unless we wish to only build models of things we think we can predict with 100% accuracy. And fortunately, our models do not always have to be perfectly correct in order to generate useful predictions or explanations. As the statistician George Box famously said, "all models are wrong, but some are useful."

An important part of learning to do good statistical analysis is learning to think clearly about models so that you can pick out a model that is useful for whatever it is you want to accomplish. And the first step toward understanding many statistical models is learning to think about the world in probabilistic terms, as we've done here in this reading. Probabilistic thinking asks questions like:

- Based on what I do know and what I don't know, what can I predict?
- How does adding or removing different pieces of information change my prediction?
- How much uncertainty is there in my prediction?
- How often will my prediction differ greatly from what actually happens (even if my model is correct)?

6 Regression with Qualitative Independent Variables

Let's say I'm interested in studying how personality relates to gender. The most common personality measure in psychology is called the "Big Five" personality inventory. There is a standard set of 50 survey items that researchers can use to measure five aspects of personality. Figure 6.1 is an example of some of these questions and how they are formatted:



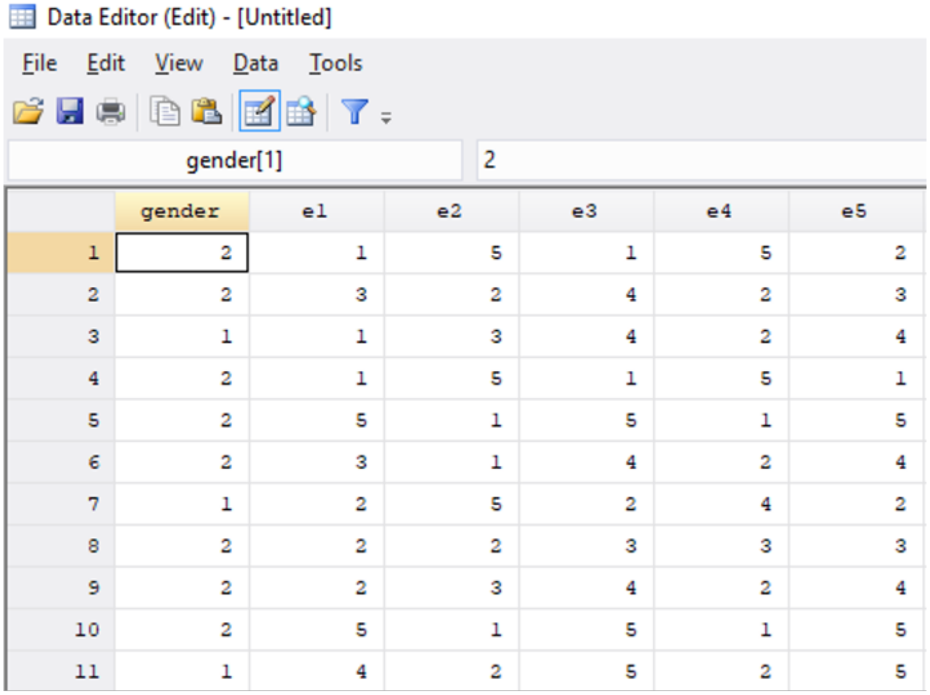
	Disagree		Neutral		Agree
I am the life of the party.	○	○	○	○	○
I feel little concern for others.	○	○	○	○	○
I am always prepared.	○	○	○	○	○
I get stressed out easily.	○	○	○	○	○
I have a rich vocabulary.	○	○	○	○	○
I don't talk a lot.	○	○	○	○	○
I am interested in people.	○	○	○	○	○
I leave my belongings around.	○	○	○	○	○
I am relaxed most of the time.	○	○	○	○	○

Figure 6.1

For now, I decide to focus on whether people are introverted or extroverted. Extroverts are outgoing and tend to enjoy interacting with others. Extroverts will tend to agree with the statement "I am the life of the party" while introverts will tend to agree with the item "I don't talk a lot."

I find a dataset that contains lots of responses to the Big Five personality questions as well as information on the gender of each respondent.[1] There are 10 different questions related to extroversion, and the dataset has one variable (column of data) for each of these 10 questions. The column labeled e1 shows responses to the item "I am the life of the party." A value of 1 means the respondent disagrees with this statement, while a 3 indicates neutral, and a 5 means they disagree.

[1] https://openpsychometrics.org/_rawdata/ (the file I used is called "BIG5.zip")



gender[1] 2

	gender	e1	e2	e3	e4	e5
1	2	1	5	1	5	2
2	2	3	2	4	2	3
3	1	1	3	4	2	4
4	2	1	5	1	5	1
5	2	5	1	5	1	5
6	2	3	1	4	2	4
7	1	2	5	2	4	2
8	2	2	2	3	3	3
9	2	2	3	4	2	4
10	2	5	1	5	1	5
11	1	4	2	5	2	5

Figure 6.2

For all of the odd-numbered extroversion questions (e1, e3, e5, etc.), agreement indicates extroversion. For the even-numbered items (e2, e4, e6, etc.), agreement indicates introversion. To create a single extroversion variable that combines responses from all 10 survey items, I create a tally, adding up all the values for odd-numbered questions and then subtracting the responses to the even-numbered questions. An extreme extrovert will have a 5 for all the odd-numbered questions and a 1 for all of the even-numbered ones, giving them a score of 20 (5x5-5x1=20). An extreme introvert will have a -20 since they will answer 1 to all the odd-numbered questions and 5 to all the even-numbered ones (5x1-5x5=-20).

Most people lie somewhere in the middle between introversion and extroversion:

Our gender variable was measured by asking respondents “What is your gender?” and they could choose from male, female, or other. In a moment, we’ll consider those who responded “other,” but for now, let’s just look at those who chose either male or female.

6.1 Predicting extraversion using gender

If I want to describe differences in extraversion by gender in this dataset, I can compute the mean value of extraversion for males and for females. It turns out that males have an average extraversion of -0.46 while females’ average level of extraversion is 0.53. Thus, the average

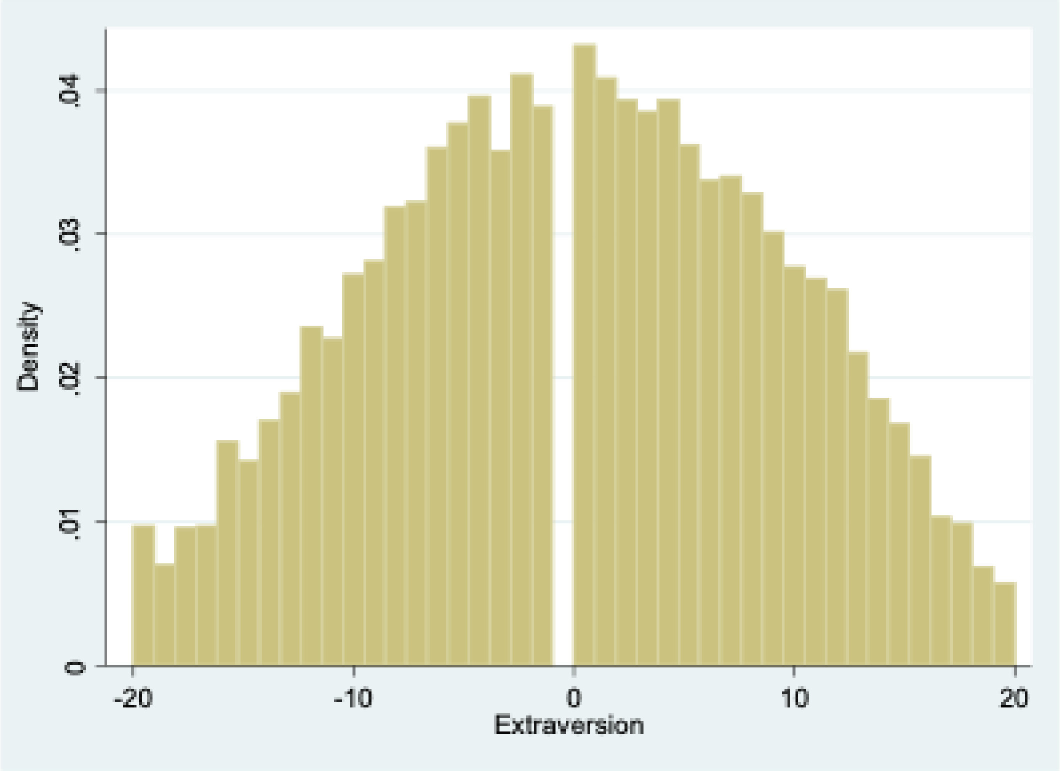


Figure 6.3

female is about 1-point more extraverted than the average male. But of course, there is lots of variation in extraversion among both groups:

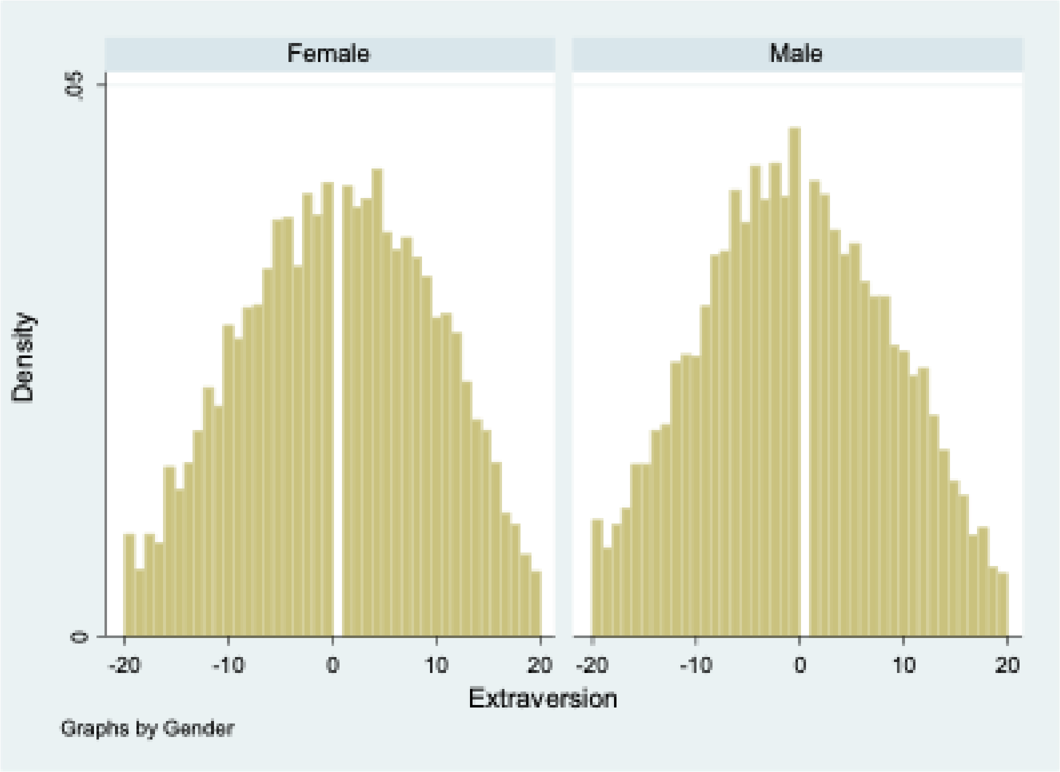


Figure 6.4

There are plenty of females who are introverts and plenty of males who are extroverts.

If you asked me to guess the extroversion level of someone and the only thing you told me about them was their gender, my best bet would probably be to guess the average extroversion level for someone of that gender. So for a female I knew nothing else about, I would guess their extroversion to be 0.53, while for a male I'd guess -0.46.

Social scientists use the **dependent variable** to describe the variable they're making a prediction about and **independent variable** to describe the variables that help them make that prediction. So in this example, extraversion is my dependent variable and gender is my independent variable.

When we're working with data, sometimes it's helpful to express how I would make a guess about a dependent variable (extraversion) based on other factors (gender) using a mathematical formula. In fact, this is exactly what we do when we run a regression. There are many ways I could write this formula, but I'll show just two for now. First, I could write:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male \tag{6.1}$$

Notice I've added a "hat" above the name of the variable *Extraversion*; this hat means that I'm making a guess about the value of that variable (I'm guessing the level of extraversion based on gender). The equation has two other variables *Female* and *Male*, and these two variables will take on a value of 1 if the person's gender is equal to the name of the variable and will otherwise take on a value of 0. For a female, *Female* will equal 1 and *Male* will equal 0, giving us:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) = 0.53$$

So our guess for the level of extroversion ($\widehat{Extraversion}$) of a female we know nothing about is 0.53.

For a male, our guess is:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (1) = -0.46$$

There's a second way I can write my formula, which will turn out to be more useful in the future when we come to consider multiple factors at the same time that might help us predict the value of a dependent variable. Rather than having two variables to represent gender in my equation, I can just use one:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male \tag{6.2}$$

In Equation 6.2, we start from female as our baseline. Notice that the first number we see (0.53) is our guess for the value of extraversion for a female. When we're considering a female, Male=0, so:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 0 = 0.53$$

Thus, we get the right prediction for females from this equation, even though we didn't include a variable specifically for females. If we have a male, Male=1, so we get:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 1 = -0.46$$

This is the same prediction we got before. Remember, I decided to initially just analyze respondents who selected either male or female. Since we are only considering two categories (male or female), and each respondent is either a male or a female, saying *Male* = 1 lets me know that *Female* = 0. It's actually repetitive in this context to both say that *Male* = 1 and *Female* = 0. Similarly, saying *Male* = 0 implies that *Female* = 1. So I can simplify my equation by just including one variable to indicate binary gender.

Notice that in Equation 6.2, the number next to *Male* is equal to the difference between the average level of extraversion for females and the average level for males (0.53-(-0.46)=0.99). This is because Equation 6.2 starts with females as the baseline, so to get our prediction for males, we have to adjust our baseline prediction by the average difference for males.

Equation 6.2 is also typically how we will arrange our equation when we're running a regression.

6.2 Prediction with more than two categories for gender

I now move beyond the gender binary and consider the "other" category in survey responses. I'll refer to this other category as "non-binary" gender. The average level of extraversion among those with non-binary gender is -5.66. So non-binary people tend to be quite a bit more introverted than those who identify as male or female. As with males and females, there is considerable variation among non-binary people:

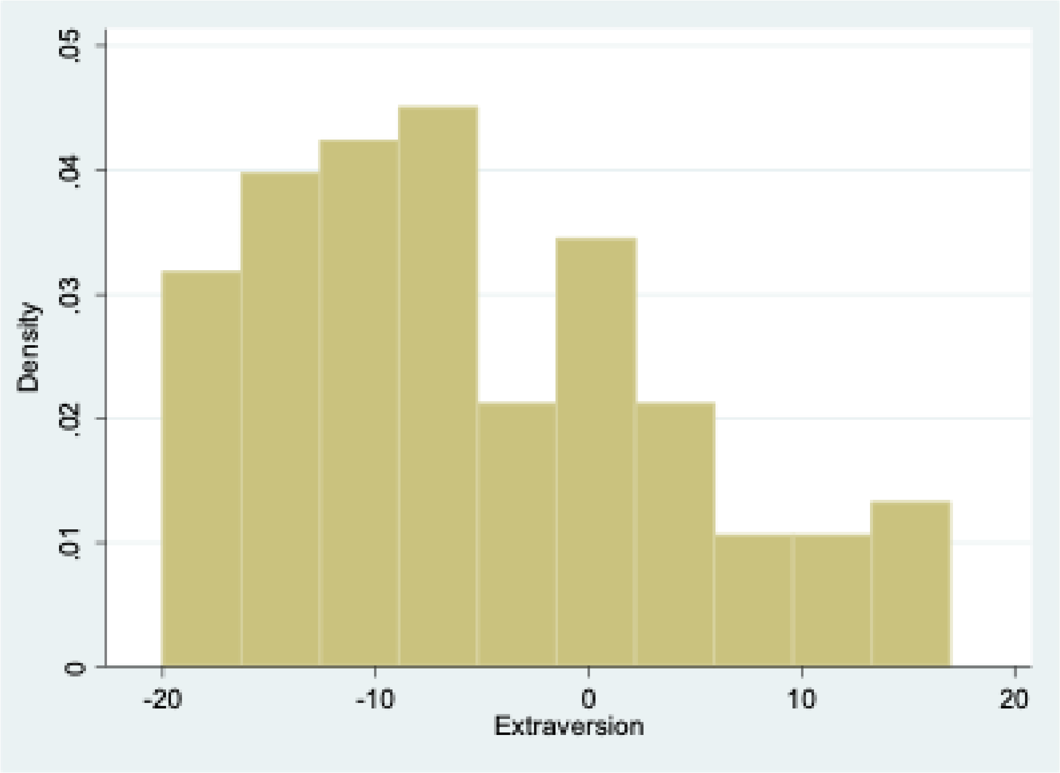


Figure 6.5

The number of non-binary respondents is relatively small (102), so it's not terrible surprising that this histogram looks a bit choppier than the ones we saw before.

Again, if we had to make a guess about the level of extraversion of someone, and all we knew about that person was that their gender was non-binary, we would probably want to guess the mean value among non-binary respondents (-5.66). Modifying Equation 6.1 to incorporate a third category is relatively straightforward:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male - 5.69 \times Other \tag{6.3}$$

For someone who identifies as female, we would plug in $Female = 1$, $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) - 5.66 \times (0) = 0.53$$

If someone identifies as non-binary, we would use $Female = 0$, $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (0) - 5.66 \times (1) = -5.66$$

We can also return to the format of Equation 6.2 but modify it to include the other category. This is how we will typically write our equation if we are doing a regression:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male - 6.19 \times Other \tag{6.4}$$

Now that there are three possible values for gender (female, male, and other), knowing the value of *Male* doesn't necessarily allow us to conclude what the vale of female is. If , the individual could identify as either female or non-binary. So we have to include a second variable. In this case, we chose to include the variable *Other*. If we know the values of *Male* and *Other*, we can always figure out the value of *Female* by process of elimination.

For a non-binary person, we plug in $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (1) = -5.66$$

When considering a female, we use $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (0) = 0.53$$

Equation 6.3 and Equation 6.4 communicate an equivalent method of making a prediction about extraversion based on gender; they just offer this information in two different formats.

Equation 6.4 might be a bit trickier to understand for now, but it will become very useful in the future.

Notice that we can talk about gender either as one qualitative variable with three possible values (female, male, or other), or we can talk about it as a series of three dummy variables (*Female*, *Male*, and *Other*) that can take each on a value of either 0 or 1. This can make things a bit confusing, but the important thing to remember is that when we have a qualitative variable with more than two categories, we'll need to break out the categories into a set of dummy variables for purposes of representing the qualitative variable in an equation.

However, as Equation 6.2 and Equation 6.4 illustrate, we don't necessarily need a dummy variable for every single category. Specifically, whenever we want to create an equation with a qualitative independent variable in a format like Equation 6.2 or Equation 6.4, the number of dummy variables should be equal to the number of categories minus one. Since our gender variable can take on three possible values in this example, we included two independent variables in Equation 6.4. No dummy variable is included for female, so we call female the **omitted category** or the **baseline category**. Remember, the first number in Equation 6.4 is 0.53, which represents our guess for females—the baseline category. If we instead had a qualitative variable with five categories, we would include four dummy variables in our equation.

7 Regression with Qualitative Dependent Variables

Suppose I want to build a model of voting. I decide to use the 2016 American National Election Studies[1] survey results to try to understand how race is associated with voting. Respondents in the 2016 survey were asked about who they voted for in 2012, and I'm going to focus on their 2012 voting patterns for now. Here are the distributions for my two main variables of interest:

~~~
```{stata}

. tab vote
PRE: RECALL OF LAST (2012) PRESIDENTAL  |
                            VOTE CHOICE |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                        1. Barack Obama |      1,728       56.58       56.58
                         2. Mitt Romney |      1,268       41.52       98.10
                       5. Other SPECIFY |         58        1.90      100.00
----------------------------------------+-----------------------------------
                                  Total |      3,054      100.00

. tab race
  PRE: SUMMARY - R SELF-IDENTIFIED RACE |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                 1. White, non-Hispanic |      3,038       71.68       71.68
                 2. Black, non-Hispanic |        398        9.39       81.08
3. Asian, native Hawaiian or other Paci |        148        3.49       84.57
4. Native American or Alaska Native, no |         27        0.64       85.21
                            5. Hispanic |        450       10.62       95.82
6. Other non-Hispanic incl multiple rac |        177        4.18      100.00
----------------------------------------+-----------------------------------
                                  Total |      4,238      100.00
```
~~~

[1] https://electionstudies.org/data-center/2016-time-series-study/

Notice that my dependent variable (vote) is qualitative. It can take on three possible values: voted for Obama, voted for Romney, or voted for other. I can build a simple set of regression models to see how race predicts vote choice. The key is to first convert each of the three categories for my dependent variable into its own dummy variable. I can accomplish this with the following code:

~~~
```{stata}
tab vote, gen(vote_)
```
~~~

I now have several new variables in my dataset that have names starting with “race_”:

~~~
```{stata}
. tab vote_1

   vote==1. |
     Barack |
      Obama |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,326       43.42       43.42
          1 |      1,728       56.58      100.00
------------+-----------------------------------
      Total |      3,054      100.00

. tab vote_2

   vote==2. |
Mitt Romney |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,786       58.48       58.48
          1 |      1,268       41.52      100.00
------------+-----------------------------------
      Total |      3,054      100.00

. tab vote_3

   vote==5. |
      Other |
    SPECIFY |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,996       98.10       98.10
          1 |         58        1.90      100.00
~~~

~~~
------------+-----------------------------------
      Total |      3,054      100.00
```
~~~

I also convert my race variable into a set of dummy variables by running:

~~~
```{stata}
tab race, gen(race_)
```
~~~

I can then run three regressions, one for each value of my dependent variable. Let's start with voting for Obama (vote_1):

~~~
```{stata}
. reg vote_1 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =     76.29
       Model |  83.3981974         5  16.6796395   Prob > F        =    0.0000
    Residual |  662.426572     3,030  .218622631   R-squared       =    0.1118
-------------+----------------------------------   Adj R-squared   =    0.1104
       Total |  745.824769     3,035  .245741275   Root MSE        =    .46757

------------------------------------------------------------------------------
      vote_1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |   .4972868   .0281049    17.69   0.000     .4421802    .5523934
      race_3 |   .2078207   .0541766     3.84   0.000     .1015941    .3140472
      race_4 |   .1028423   .1353307     0.76   0.447     -.162507    .3681916
      race_5 |   .3135004    .032158     9.75   0.000     .2504466    .3765542
      race_6 |   .1042547   .0441427     2.36   0.018      .017702    .1908075
       _cons |    .480491   .0097901    49.08   0.000     .4612952    .4996868
```
~~~

Since our independent variable is qualitative, we have an omitted category. In this case, we've left category 1 (race_1) out of our regression, which indicates non-Hispanic white respondents. Our constant indicates that predicted value of the dependent variable when all independent variables are equal to zero. We can see this by writing out the regression equation:

$$\widehat{vote\_}1 = .48 + .50race\_2 + .21race\_3 + .10race\_4 + .31race\_5 + .10race\_6 \tag{7.1}$$

For non-Hispanic white respondents, race_1 equals one and all other race dummy variables equal zero, so we get:

$$\widehat{vote\_}1 = .48 + .50(0) + .21(0) + .10(0) + .31(0) + .10(0) = .48$$

Remember, vote_1 is equal to zero if the respondent didn't vote for Obama, and it is equal to one if the respondent did vote for Obama. Our predicted value is neither zero nor one; instead, we get .48. This can be interpreted as indicating the probability of a one. In other words, a non-Hispanic white has a .48 probability of voting for Obama. We can also convert this probability to a percentage by moving the decimal place two spots to the right: a non-Hispanic white is estimated to have a 48% chance of voting for Obama, according to this model.

Now, let's look at the slope coefficients. The coefficient for black (race_2) equals .50. Thus, a one-unit increase in race_2 is associated with a .50-unit increase in vote_1. Let's break that down a bit to see if we can create a clearer interpretation. Since race_2 is a dummy variable and non-Hispanic white is the omitted category, a one-unit increase in race_2 correspondents to having a black respondent instead of a white respondent. And since our dependent variable is binary, we should think in terms of probabilities, which can be converted to percentages: a .50-unit increase in vote_1 means a 50 percentage-point increase in the probability of voting for Obama. So putting this altogether, we'd say: (non-Hispanic) black voters are 50 percentage points more likely to vote for Obama than (non-Hispanic) white voters, according to this model.

Similarly, Asian voters are 21 percentage points more likely to vote for Obama than (non-Hispanic) white voters. Native Americans are 10 percentage points more likely to vote for Obama than (non-Hispanic) white voters. Hispanics are 31 percentage points more likely to vote for Obama than non-Hispanic white voters. And voters identifying as multiracial or other race are 10 percentage points more likely to vote for Obama than (non-Hispanic) white voters. All of these differences are statistically significant, except for Native American versus white voters (probably because there are only 27 Native Americans in the sample, making the estimate of this difference very imprecise).

Let's move onto running a regression for the second category of our dependent variable:

~~~
```{stata}
. reg vote_2 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =     72.35
       Model |  78.6117037         5  15.7223407   Prob > F        =    0.0000
    Residual |  658.463395     3,030  .217314652   R-squared       =    0.1067
-------------+----------------------------------   Adj R-squared   =    0.1052
~~~

~~~
       Total |  737.075099     3,035  .242858352   Root MSE        =    .46617

------------------------------------------------------------------------------
      vote_2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |   -.483031   .0280207   -17.24   0.000    -.5379725   -.4280895
      race_3 |  -.2002027   .0540143    -3.71   0.000     -.306111   -.0942944
      race_4 |  -.0822373   .1349253    -0.61   0.542    -.3467917     .182317
      race_5 |  -.3014791   .0320617    -9.40   0.000     -.364344   -.2386142
      race_6 |  -.1344972   .0440105    -3.06   0.002    -.2207906   -.0482038
       _cons |    .498904   .0097607    51.11   0.000     .4797657    .5180423
```
~~~

Now we're looking at predictions of voting for Mitt Romney. Our constant is .50, indicating that a non-Hispanic white voter has a 50% chance of voting for Mitt Romney. The coefficient of -.48 for race_2 indicates that (non-Hispanic) black voters are 48 percentage points less likely to vote for Mitt Romney than (non-Hispanic) white voters. I won't go on to interpret the rest of the coefficients, but they follow the same pattern.

Finally, let's look at a regression with vote_3 as the dependent variable:

~~~
```{stata}
. reg vote_3 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =      2.23
       Model |   .20833556         5  .041667112   Prob > F        =    0.0490
    Residual |  56.6836275     3,030  .018707468   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  56.8919631     3,035  .018745293   Root MSE        =    .13678

------------------------------------------------------------------------------
      vote_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |  -.0142558   .0082213    -1.73   0.083    -.0303757    .0018642
      race_3 |   -.007618   .0158479    -0.48   0.631    -.0386917    .0234557
      race_4 |   -.020605   .0395873    -0.52   0.603    -.0982258    .0570158
      race_5 |  -.0120213    .009407    -1.28   0.201     -.030466    .0064234
      race_6 |   .0302425   .0129128     2.34   0.019     .0049238    .0555611
       _cons |    .020605   .0028638     7.19   0.000     .0149898    .0262202
```
~~~

This regression provides some insights into who supported third-party candidates in the 2012 election. First, our constant indicates that a non-Hispanic white voter has a 2% chance of voting third-party. (Non-Hispanic) black voters are one percentage point less likely to vote third-party than white voters, although this difference is only significant at the .10 level. The only other significant slope coefficient is for race_6, where we see that people who identify as multiracial or other race are estimated to be three percentage points more likely to vote third-party than (non-Hispanic) white respondents.

Now that we've run one regression for each category of our dependent variable, we've completed an analysis. Note that using regular linear regression (the reg function in Stata) is not the only way (or even necessarily the preferred way) to analyze a qualitative dependent variable. There are other models (e.g., multinomial logistic regression) that are specifically designed to be used with a qualitative dependent variable. However, using simple linear regression is a good way to get started looking at qualitative variables if you haven't learned these fancier models and how to properly interpret them.

One final thing I want to show you is that our results will be in a slightly different format but will be in one sense equivalent if we decide to use a different category as our omitted category when using a qualitative independent variable. Let's say we want to make black (race_2) our reference category. Compare the following results to what we saw near the top of this page:

~~~
```{stata}
. reg vote_3 race_1 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =      2.23
       Model |  .20833556         5  .041667112   Prob > F        =    0.0490
    Residual |  56.6836275     3,030  .018707468   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  56.8919631     3,035  .018745293   Root MSE        =    .13678


------------------------------------------------------------------------------
      vote_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_1 |   .0142558   .0082213     1.73   0.083    -.0018642    .0303757
      race_3 |   .0066378    .017388     0.38   0.703    -.0274557    .0407313
      race_4 |  -.0063492   .0402287    -0.16   0.875    -.0852274     .072529
      race_5 |   .0022345   .0118186     0.19   0.850    -.0209387    .0254077
      race_6 |   .0444983   .0147623     3.01   0.003      .015553    .0734435
       _cons |   .0063492   .0077064     0.82   0.410    -.0087611    .0214595
```
~~~

Now, our constant tells us that a black voter has a .6% chance of voting third-party. This is the same prediction we would get from our prior model where race_1 was the omitted category: to find our prediction for black voters from the prior results we would have added the coefficient for race_2 (-.014) to the constant (.021), yielding .6% or .006 (or .007 if we use the rounded numbers shown in parentheses).

The coefficient for race_1 tells us about how white voters differ from black voters. Notice that the p-value is exactly the same as what we saw in the prior table for race_2, and the coefficient for race_1 in this table is the same as the coefficient for race_2 in the prior table, except the sign has changed. That's because comparing black to white is the same as comparing white to black, except that we're going in the opposite direction.

You can go on to play around with these two sets of results more on your own if you'd like. Both regression equations will yield the same prediction for a voter of any given race. The difference lies only in the starting point, as represented fby the constant. However, the p-values will usually differ because they are describing a different comparison (e.g., comparing Asian to black in this table versus comparing Asian to white in the prior table). Thus, it doesn't really matter which category you pick as your omitted category, except that you may care more about some comparisons than others. You can also run the same regression multiple times but with different omitted categories so that you can get the p-values for a full set of comparisons across groups.