20 Statistical Inference

Statistics can be more than a tool for describing data. In the social sciences we have hypotheses that move us beyond simple descriptions of populations to relationships between two or more variables. To analyze these relationships we often rely in practice on statistical inference. That is, we need to make decisions based on data collected on a small group (sample) for the larger group that we want to study (population). In Chapter 4 we introduced the concept of sampling and provided a description of the three general objectives pertaining to sampling: representativeness, size, and level of analysis. Here we move to a more focused explanation of how sampling helps researchers overcome practical limitations and what implications this has for quantitative analyses.

Sampling

As we noted earlier, when it comes to sampling, larger is generally better, the aforementioned issues aside. So why sample at all? Why not collect data on the full population? In all types of research we find the same two answers: time and money. Research occurs in the real world and collecting the ideal data is often limited by how much time and money the researcher can direct to the project. Thus every research design must take into account the practical circumstances. In doing so, researchers necessarily restrict their investigation into a sample or subset of the population that they would like to study.

In general, when we make comparisons we would like to talk about more than just the observations in the sample. We would like to talk about the larger group of interest in our research, the population. How we do this is the objective of statistical inference. More specifically, statistical inference helps researchers provide statements of confidence in our ability to generalize or infer from the sample to the population. The ability to offer an estimate of relative precision is another reason why quantitative empirical research is so useful and popular.

Consider the ANES dataset that we used for the examples in Chapter 18. How many cases or individuals do we have in our sample? Obviously, the nearly 6,000 individuals in that sample is not anywhere near the size of the full voting age population of the United States, yet we would like to use this sample of data to describe that population. In practice, we rarely study every member of the

population and instead rely on a sample and statistical inference to generalize to the population.

Statistical inference depends on knowing that the people in our sample are representative of the population. If they are representative *enough* we can generalize our conclusion from the sample to the population. This inference depends on every case in the population being randomly drawn into the sample. In statistics a **random sample** means that every case in the population has an equal chance of being drawn into the sample. Thus random is not *any* chance in statistics; it has a precise definition of equal probability.

As an example, consider blindfolded draws of slips of paper from a hat. If every slip has the same shape, size, and only one name on it, then every name has the same probability of being chosen. But we would need a huge hat to do this for several thousand, let alone the millions of people in the US population. In order to perform random sampling, we typically use a random number generator or table. In either case the output is a series of numbers having no particular pattern or order. Drawing a random sample is a simple three-stage process:

1 Attain a complete list of the population.
2 Assign unique identifying numbers to each unit or member of the population.
3 Draw the members of the sample from a random number generator or a table of random numbers.

This process is our best chance to get a sample to look like a population, provided, of course, that we have no additional information about the population.

Above we saw how we can use z-scores to assess the relationship between individual observations and the population they belong to, under a specific set of assumptions about the population parameters. We now consider a more realistic scenario that is consistent with an applied research design process where we have no information about the population μ and σ. We will take a random sample of observations from the population in order to make guesses about the population. Thus, instead of looking at one observation, we now have a sample of multiple cases. Instead of just having one value, we now have two pieces of information: a sample mean, $\bar{X}$, and a sample standard deviation, s.

So why does our sample information not perfectly match the population? Because of **sampling error**. That is, because of the process of selection our sample rarely has the same mean and standard deviation as our population. This is not an error that we can fix. Moreover, it prevents us from being exact with our estimates. Thus all inference involves uncertainty.

An example of this uncertainty around our statistical inferences that should be familiar to social science students comes from election coverage. In election polling candidates' relative chances are not reported alone but along with an estimate of error based on the polling sample. For example, a candidate might be estimated to have 65% of the vote, with a margin of error of 4 percentage points. We recognize this margin of error (discussed in more detail below) as an

indicator of uncertainty in our point estimate of anticipated vote choice. That is, the researcher is confident that this candidate has within 61% to 69% of the vote. How do we arrive at this range of values?

Samples and Populations

So far we have considered three types of distributions: empirical distributions of actual data (e.g., Obama's feeling thermometer scores), theoretical distributions of probabilistic processes (e.g., rolling dice), and theoretical statistical distributions (e.g., the normal curve). Now we consider the distribution of sample means as a pedagogical tool to help envision what certainty means in statistical inference. While this is a hypothetical scenario that does not represent actual research, it helps explain why we are allowed to draw conclusions about populations based on a sample. The hypothetical we are considering is as follows: We have a population from which we are repeatedly drawing samples (of any size, all the same). For each sample, we calculate the mean. We then treat each of those means, as data themselves, and assess their distribution, by calculating the mean and plotting them on a histogram. We call this distribution the **sampling distribution of the sample mean**. We can also treat the standard deviations of those samples as data and assess the distributions.

The key premise of statistical inference is that we can make generalizations from samples if the sample is representative enough of the population. We can find the extent to which our sample is representative of the population based on our understanding of the characteristics of this sampling distribution of sample means. Foremost, the sampling distribution of means approximates a normal curve. Secondly, the mean of a sampling distribution of means (the mean of means) gets closer to the true population mean as N moves toward infinity. Finally, the standard deviation of a sampling distribution of means is smaller than the standard deviation of the population.

Repeating the key characteristic, when we take repeated samples from a population the mean and standard deviation of those samples are themselves normally distributed – even if the population distribution is not. In probability theory, this is what we call the **central limit theorem**, which, like the law of large numbers, is a mathematical result of probability theory. The theorem states that the means of a series of random draws from a population distribution will be approximately normally distributed provided a sufficiently large number of draws. That is, as you increase the draws the distribution of sample means looks increasingly similar to a normal distribution.

To solidify this point, Figure 20.1 shows the results of 10 and 100 draws of means from each of three familiar distributions. Moving from the distribution in the left column to 10 mean draws in the middle column and 100 mean draws in the right column we see each of the distributions begins to converge into the normal distribution. While not shown here, it is important to remember that even for less familiar or unknown distributions – indeed for any distribution with a well-defined

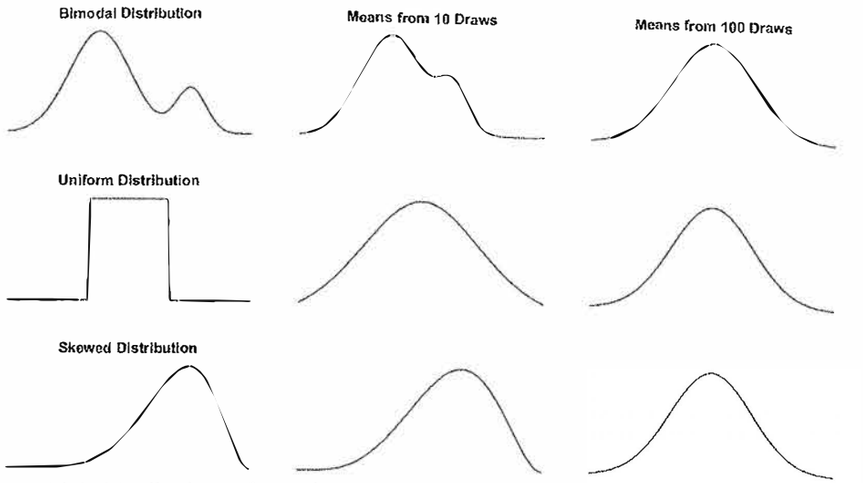

Figure 20.1 Distributions of sample means

mean and standard deviation – this theorem holds, which explains the ubiquity of the normal distribution.

With this insight we are now in a position to assess any one of our individual samples. That is, because of the characteristics of the distribution of sample means we can use what we know about the normal curve to place indicators of certainty around our estimates. Because the sampling distribution of means takes the form of the normal curve, we can say that as a score moves farther from the mean of means the probability of getting it decreases. Similarly, we know the percentage of cases falling between standard deviations and the mean.

In applied work, however, we are generally not interested in probabilities associated with a particular raw score but with samples drawn from a population. We want to make a statement about how likely our sample would be to occur, given our population mean and standard deviation. The general procedure is similar to the z-score procedure above, but, since we are assessing a sample of size N this time, we cannot simply use our population standard deviation to create z-scores. Instead we rely on the **standard error of the mean**:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}. \tag{20.1}$$

Note that we will often calculate (or more often be given by statistical software) **standard errors** associated with certain quantities. Generally speaking, these quantities represent our uncertainty about the estimate in question. If the ratio of our estimate to its standard error is high (e.g., the ratio of $\overline{X}$ to $\sigma_{\overline{X}}$), our guess is very precise. If the ratio is low, we hold less confidence in our estimate.

Returning to our sampling distribution of sample means now with the standard error of the mean and some population parameters (which we do not typically

have in practice but are given here for pedagogical purposes), we can arrive at a z-score for the sample means distribution in the same way we arrived at a z-score for any X value above:

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}. \tag{20.2}$$

This z-score provides us the probability of finding the mean score from the sample in the population. That is, if we then consult Table A.1 for our calculated z-score we can find the probability of randomly choosing this sample (with its particular mean and standard deviation) from a population.

Relatedly, we can also find the range of mean values within which our true population mean is likely to fall, which brings us to the concept of the **confidence interval** (CI). The confidence interval is just our estimate or statistic wrapped in some range of uncertainty. Here the estimate is the mean, but we will want to create confidence intervals for the statistics we introduce later in the book as well. The range of uncertainty around a statistic is conveyed by the **margin of error** (MOE), which expresses the amount of sampling error in our results. The larger the margin of error the less confidence we hold that our observed estimate is close to that of the population. Thus, the confidence interval is simply an estimate plus and minus the margin of error:

$$CI = \text{Statistic} \pm MOE, \tag{20.3}$$

where the margin of error is a particular chosen value of a standardized score (e.g., z-score) multiplied by the standard error of the statistic:

$$MOE = \text{standardized score} \times \sigma_{Statistic}. \tag{20.4}$$

Equivalently, a confidence interval is just an estimate plus and minus a standardized score times the standard error of estimate.

In the case of the mean for the normal distribution our confidence interval is calculated accordingly:

$$CI = \overline{X} \pm 1.96 \times \sigma_{\overline{X}}. \tag{20.5}$$

The chosen value of the standardized score, 1.96, corresponds to the level of confidence we choose to hold in our estimate. That is, here we arrive at the uncertainty by adding and substracting from the estimate our standard error multiplied by a particular z-score, 1.96. But where does 1.96 come from?

Hypothesis Testing

Throughout the statistics section of the book we have asked you to consider a number of research questions and how we might answer them. In practice, however, researchers more formally put forward statements of expectation to test with their data, which we call **hypotheses** (introduced in Chapter 2). Hypothesis

testing in statistics follows a specific process. We begin by offering a research hypothesis, or statement of expectation. For any hypothesis we also propose a **null hypothesis**, or opposing expectation that we will try to reject with our hypothesis test. The null hypothesis typically holds that the observed results occured by chance; i.e., sampling error. We next obtain a sample and calculate the relevant statistic. Finally, we calculate the probability of observing the statistic by chance, under the assumption that the null hypothesis is true. Based on this probability we decide whether or not to reject the null hypothesis.

The decision of whether or not to reject the null hypothesis is made easier and more consistent by accepting a conventional threshold. We refer to the **significance level** in terms of α, which equates to one less than our chosen **confidence level**:

$$\alpha = 1 - \text{confidence level}. \tag{20.6}$$

α corresponds to the area of the distribution in the tails. It is simply the probability of rejecting the null hypothesis if the null hypothesis is true. We might think of it as an expression of our chosen probability of being wrong. Thus we decide to reject the null hypothesis only when we are really confident, which means that we should choose as our threshold an α that is very small. But exactly how small is very small? The standard confidence level in the natural and social sciences is 0.95 (or 95%), which corresponds to α of 0.05 (or 5%) and a z-score of ± 1.96. The left graph in Figure 20.2 illustrates that this α means there is 2.5% in each of the tails. We call the z-scores demarcating the confidence level **critical values**, because we know that attaining a larger z-score than ± 1.96 ($z < -1.96$ or $z > 1.96$) conveys that the result is statistically significant at this confidence level. That is, statistical significance conveys that we reject the null hypothesis that the observed results occured by chance, since they are far, by conventional standards, from the null hypothesis under the assumed distribution. Thus the α value serves as a practical cutoff point at which the null hypothesis can be rejected in the context of sampling error. When rejecting the null hypothesis (at the conventional 95% level of confidence) we are conveying that there is less than a 5 in 100 chance that we have done so when we should have failed to reject it.

The interpretation of the confidence level is made clearer by reconsidering the hypothetical example of repeatedly drawing random samples and calculating the

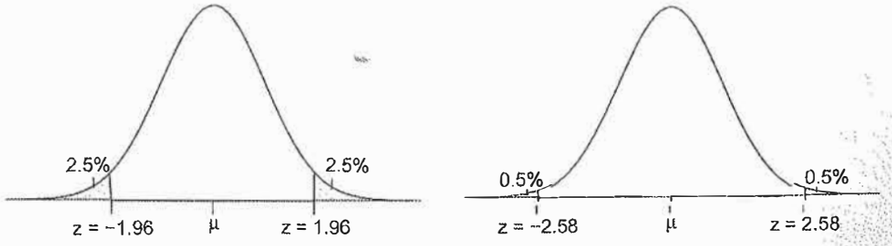

Figure 20.2 Varying alpha

mean for each one. If we were to draw 100 samples from the population, 95 of the times the confidence interval would cover the true mean. Looking under the normal curve 95% of the area falls around the mean between the z-scores of −1.96 and 1.96; the margin of error for the normal distribution follows accordingly:

$$MOE = 1.96 \times \sigma_{\bar{X}}. \tag{20.7}$$

Thus, the interval has a 95% chance of including the true value. The reason for this threshold, however, is purely conventional. We are merely accepting this significance level to make scientific work more consistent. For example, 94% also seems to be a good level of confidence, and 96% even more so. The right graph in Figure 20.2 shows a tougher test for rejecting the null hypothesis. With a 99% significance level, corresponding to z-scores of −2.58 and 2.58, there is only 0.5% in each of the tails. Why then do we choose 95%?

Importantly, 95% is just a rule of thumb. There is no statistical reason why it is most commonly used as the cutpoint. It is merely the norm in most scientific disciplines to hold estimates at this level of certainty. While it is a largely accepted matter of convention, researchers should therefore be careful in relying strictly on 95% as a cutpoint. In addition, since we have a standard choice of 95% confidence intervals, scientists have 5% chance of being wrong or $\alpha = 0.05$. That is, even though we have a random sample, samples can sometimes be very unrepresentative of the population by chance. We accept that there is always some sampling error by providing a range of confidence or certainty around each of our estimates, as with a confidence interval.

We attain statistical significance when the **p-value** for the statistic is less than α. In terms of the distribution, the p-value is the area from the z-score (or another test statistic discussed below) to the tails. It is therefore the exact probability of observing a sample statistic as or more extreme than the observed one if the null is true. Thus, if the z-score is within the tails denoted by the critical values, the p-value will be less than or equal to α, and we reject the null hypothesis. The p-value associated with the standard 95% confidence level is 0.05, so, in practice, anytime we arrive at a p-value less than or equal to 0.05 we reject the null.

There are two types of errors in statistical hypothesis testing: **Type 1** and **Type 2**. Figure 20.3 illustrates these errors in a 2 × 2 table. The columns refer to the two possible states of the null hypothesis in reality. The rows refer to the two possible states as perceived or measured through the statistical tests. Type 1 errors mean we rejected the null when we should have retained it. In other words the error is the rejection of the null hypothesis when in reality the null hypothesis is true, or a **false positive**. Type 2 errors mean we retained the null when we should have rejected it. This is the failure to reject the null hypothesis when in reality it is false, or a **false negative**.

It is important to remember that these concepts are intertwined in inferential statistics. In quantitative analysis – specifically, when relying on levels of significance to reject the null hypothesis – we can provide additional consideration depending on our concerns over particular errors. If we are worried about Type 1 errors, we can increase the stringency of α; e.g., move from $\alpha = 0.05$ to $\alpha = 0.01$.

	True	False
Fail to Reject	Correct	Type II Error
Reject	Type I Error	Correct

Figure 20.3 Hypothesis test results against real state of null hypothesis

The right graph in Figure 20.2 demonstrates that we have shrunk the amount of area in our tails by decreasing α. Thus the probability of getting a Type 1 error is just α. However, if we are more worried about Type 2 errors, we can increase the size of the sample so that we are more likely to reject the null hypothesis when it should be rejected. That is, failing to reject a null hypothesis is less likely by random chance if the sample is larger. The probability of getting a Type 2 error is thus directly related to our sample size.

Recall that α specifically refers to the size of the tail regions under the curve. It is the threshold value below which it is considered so small that the null hypothesis can be rejected, and is determined ahead of time by the researcher who is balancing the costs of Type 1 and Type 2 errors. In practice researchers often merely check to see if the z-score or test statistic exceeds the critical value (e.g., 1.96) associated with our chosen α (e.g., 0.95). If so we can say that the results are statistically significant at the α level. However, we need to be vigilant about the interpretation of these concepts in light of their derivation and somewhat arbitrary nature.

Estimating Population Parameters

Returning to our estimation of the population mean, we are still not in a realistic position from the perspective of practical research. In practice, we only have a sample mean and sample standard deviation. So how do we use that information to arrive at an estimate of the population mean? We begin by finding an estimate of the standard error of the mean. To do so, we simply divide our sample standard deviation, s, by the square root of N; thus the sample standard error of the mean:

$$s_{\bar{X}} = \frac{s}{\sqrt{(N)}}. \tag{20.8}$$

Thus in small samples this correction gives a fair estimate of the variability in the entire population. Of course, in large samples this correction is trivial and the sample means tend to be reliable estimates of the population means.

Though it brings us closer to how we use sample estimates in actual research, estimating the standard error of the mean creates a new problem: the sampling distribution of means is no longer normal due to using a random variable $s_{\overline{X}}$ in place of the population parameter $\sigma_{\overline{X}}$. The extra uncertainty in the estimated standard error makes the sampling distribution of means wider. Our distribution now has greater dispersion than a normal distribution, so we cannot use z-scores, which refer only to normal distributions. Instead, the ratio follows a **t-distribution**, where our standardized score is now:

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}. \tag{20.9}$$

The t-distribution is, however, similar in two important ways to the normal distribution: it is symmetric and the area under the curve can be characterized by knowing the mean and standard deviation. We are still interested in stating the range in which we can be confident that the population mean falls. That is, we will need to calculate a margin of error for our estimate. However, since we are using t-ratios instead of z-scores, the appropriate cutpoints are not always 1.96. The t-ratio, unlike the z-score, depends on **degrees of freedom** (df), where

$$df = N - 1. \tag{20.10}$$

In the calculation of a statistic, the degrees of freedom is the number of values that remain variable. In other words, it is the number of observations less the number of parameters used to estimate the statistic. The greater the df, the larger the sample size, and thus the closer the t-distribution is to a normal distribution, as shown in Figure 20.4. So when the sample is large there is no difference between a z-score and t-ratio, and thus we can rely on the familiar z-score instead of the t-ratio. When the sample is small we rely on the t-ratios.

Take, for example, the sample variance for X. Because it requires the calculation of a single parameter, the mean, it has $N - 1$ degrees of freedom. The rationale is that because the sample standard deviation is smaller than the standard deviation would be when calculated from the population, we inflate the sample variance slightly with $N - 1$ in the denominator instead of N. In other words the sample mean fits the sample better than the population mean might, and so the sample standard deviation has a slight bias in that it is a smaller representative of

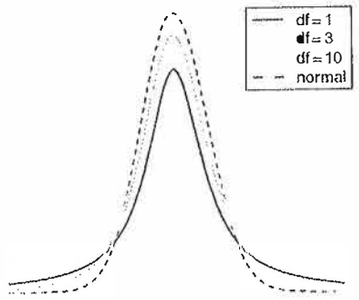

Figure 20.4 T-distribution at different degrees of freedom

the population standard deviation. So we make a quick correction in this by taking out a bit from the denominator. Accordingly, we arrive at less biased or, as they are frequently called, **unbiased estimates** of the population parameter:

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{N - 1}. \tag{20.11}$$

Thus in small samples this correction gives a fair estimate of the variability in the entire population. Of course, in large samples the sample means tend to be reliable estimates of the population means. Like before with the normal curve, we can use a table to inform us of the area under the t-distribution. When we rely on the t-ratio table in Appendix Table A.2, we need two pieces of information, in addition to the standardized score, in order to find where our estimate sits on the distribution. First, we need the degrees of freedom and, second, we need a confidence level. As we have noted above, the margin of error depends on the sample size, but we choose the confidence level. For the mean of a small sample, then, we can get our confidence intervals from:

$$\overline{X} \pm t \times \sigma_{\overline{X}}. \tag{20.12}$$

Again, if our sample size is large enough, the t-ratio is the same as the z-score, 1.96. If the sample size is smaller, that number increases to above 2. For example, that multiplier is equal to 2.021 for a sample of 40, and 2.228 for a sample of 10. Thus the smaller the sample the more uncertainty around our estimate.

Example: Average Number of Parties in Democracies

Let's return to the concept of democracy and consider a comparative politics example. Here we would like to know about how many political parties we should expect to find in a democratic country. That is, what is the average number of parties in a democracy? We do not have the time to collect the number of parties for all democratic countries in the world and instead only collected this data for a random sample of 60 of them.

Akin to practical research, we do not know the population mean – which would answer our question – but we can use our sample to make an educated guess. We begin by calculating the sample mean, $\overline{X} = 3.75$, and the standard deviation, $s = 1.05$. Because we have a sample, we need to account for sampling error, therefore we calculate the standard error of the mean:

$$\begin{aligned} s_{\overline{X}} &= \frac{s}{\sqrt{(N)}} \\ &= \frac{1.05}{\sqrt{60}} \\ &= 0.14. \end{aligned} \tag{20.13}$$

With our degrees of freedom for the standard error of the mean, $N = 60$, we consult Appendix Table A.2 to find that for *alpha* = 0.05 our appropriate t-value

is 2.0. This tells us that for a t-distribution with df = 100, 95% of the area under the curve falls between $t = -2.00$ and $t = 2.00$. Finally, we plug this into our formula for the confidence interval:

$$\begin{aligned} CI &= \overline{X} \pm t \times s_{\overline{X}} \\ &= 3.75 \pm (2.00 \times 0.14) \\ &= 3.75 \pm .50 \\ &= [3.47, 4.03]. \end{aligned} \tag{20.14}$$

We can thus say that the mean number of parties in our population is between about 3.5 and 4.

CONCLUSIONS

In statistical inference we generalize from a sample to a population by making assumptions about the true distribution of a variable. Reference to a probability distribution, like the normal curve, helps us understand how likely the results we have found in our sample are due to chance. In order to do so, we also need to ensure that we have a representative sample, which can be accomplished through random sampling, and can agree on a threshold for statistical confidence.

KEY TERMS

- Random sample
- Sampling error
- Sampling distribution of the sample mean
- Central limit theorem
- Standard error of the mean
- Standard errors
- Confidence intervals
- Margin of error
- Hypotheses
- Null hypothesis
- Significance level
- Confidence level
- Critical values
- p-value
- Type 1 error
- Type 2 error
- False positive
- False negative
- t-distribution
- Degrees of freedom
- Unbiased estimates

21 Bivariate Statistics

Bivariate statistics allow us to test the relationship between two variables. While simple, they provide great empirical leverage for hypotheses of association and, with the appropriate research design, causality. We next hone our ability to make controlled comparisons and introduce inference making about sample means with the difference of means test. We then proceed to correlation, which moves us beyond making a simple claim of a relationship or no relationship between two variables to a measure of both the strength and direction of the relationship.

Revisiting Levels of Measurement

In the previous chapter, we found out some important substantive information about our population means. Indeed we learned how to estimate a population mean from the information that we gain in a single sample. However, we are still not in a position to assess a hypothesis. We do not generally have expectations as simple as "the mean number of parties in a democracy is *x*." Our hypotheses typically suggest a relationship between an independent and dependent variable, and we have not made reference to bivariate analyses yet. We will explore a number of statistical tests for assessing hypotheses. We can divide hypothesis tests into three types based on the information they provide about the relationship between the independent and dependent variable:

1. Those that simply analyze whether or not there is a relationship between variables (e.g., Difference of Means).
2. Measures of association, which tell us the direction and strength of the relationship between two variables. (e.g., Correlation).
3. Measures of average effect, which tell us the amount of change in the dependent variable given a unit change in the independent variable (e.g., Regression).

It is important to note that knowing which test to use requires us to think not just about the question we would like to answer but also our variables' levels of measurement, their population distributions, and the sample size. In addition, all the tests discussed here demand that the data come from a random sample. In Chapter 3 we learned about levels of measurement. In quantitative analyses understanding the level of measurement is essential because it helps us decide which hypothesis test to use.