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.

Table 21.1 Hypothesis tests guide

Independent Variable	Dependent Variable: Discrete	Dependent Variable: Continuous
Discrete	Chi-Square, Phi, Logit, Probit, Cramer's V	Difference of Means, ANOVA, Regression
Continuous	Logit, Probit	Regression, Correlation

Table 21.1 lists some of the common and appropriate hypothesis tests – many of which are beyond the scope of this book – by the measurement classifications of the independent and dependent variables. In the table, as is typical in the literature, we group nominal and categorical levels of measurement in the general header, **discrete**, and interval and ratio levels under **continuous** (see Table 3.4). For example, if the dependent variable is continuous and so is the independent variable, we can use regression or correlation. For discrete dependent and independent variables we might use a chi-square or phi test. And with a continuous dependent variable and a discrete indepenent variable we could use difference of means test or regression.

Cross-Tabulations

While a careful look at the frequency distribution of a single variable should be the first step in any analysis, social scientists are predominantly concerned with relationships between two or more variables, not just the distribution of a single variable. That is, the focus of research often turns to testing bivariate and multivariate hypotheses. In the case of our example above, we can easily think about relationships that might be more likely to be tested by social scientists, for example, whether particular backgrounds make individuals more likely to identify with one party or another. Even in the bivariate case, such as this, frequency distributions, appropriately structured, can provide some insight.

Returning to party identification in the 2012 ANES dataset, Table 21.2 presents two frequency distributions – one for party identification and one for gender – in a single table. The distributions for the values of party identification are noted vertically down the first column (as in a single frequency distribution) and those of gender are arranged horizontally across the first row. Thus each cell now contains information on individuals who fit in a particular category for both variables. The total values of the table (in columns and in rows) are referred to as the **marginals**.

Cross-tabulations can include raw counts as well as proportions or percentages. In the example we provide the distributions as both frequencies (counts) and percentages. When we do not have the same number of cases in each group (as above), the frequencies alone tell us little. Consider, for example, a crosstab

Table 21.2 Crosstab of partisanship by gender

	Male	Female	Row total
Democrat	1006	1355	2361
Row Percent	42.61%	57.39%	
Column Percent	37.08%	47.02%	
Total Percent	17.98%	24.22%	42.2%
Independent	999	846	1845
Row Percent	54.15%	45.85%	
Column Percent	36.82%	29.35%	
Total Percent	17.86%	15.12%	32.98%
Republican	708	681	1389
Row Percent	50.97%	49.03%	
Column Percent	26.1%	23.63%	
Total Percent	12.65%	12.17%	24.83%
Column Total	2713	2882	5595
Row Percent	48.49%	51.51%	

with equal category percentages in each cell but different frequencies. We would be tempted in this case to draw conclusions from the different frequencies. However, controlling for the sample size with percentages would show us that the different frequencies are not meaningful.

In the case of differing category sizes, we require a way to standardize frequency distributions in order to compare them. We often make use of proportions and percentages in this case. The proportion simply compares the number of cases in a given category with the total size of the distribution

$$Prop = \frac{f}{N}, \tag{21.1}$$

where f is the frequency of observations and N is the sample size. Even more frequently, as above, we make use of percentages, which are the frequency of occurrence of a category per 100 cases,

$$Pct = 100 * \frac{f}{N}. \tag{21.2}$$

From Table 21.2 what can we gather about the different groups? How does party identification look for each gender? We can see, as we have above, that Democrats were the largest share of the sample at 2,361, followed by Independents. In terms of the relationship between the two variables, 1,355 of the Democrats in the survey sample were women. While Republicans appear to split fairly equally among the sexes at 708 and 681, 999 Independents were male compared to

only 846 female Independents. The largest number of respondents are female Democrats at 1,355. This is the framework for cross-tabulations, or, more colloquially, **crosstabs**. Because they involve two variables and describe some aspects of the relationship between two variables crosstabs provide a basic bivariate analysis. Thus, typically, social scientific analysis begins with a crosstab.

Beyond the counts and marginals, we can get further information from our data (depending on our interests) from the row and column percents. Perhaps we want to know more about the Independent males, in which case we could look at them relative to all Independents by dividing the frequencies in each row by the number of cases in that row,

$$Pct_{Row} = 100 * \frac{f}{N_{Row}} = 54.15\%. \quad (21.3)$$

Or we could use column percents if we wanted to know the percentage of females that are Democrat relative to the entire female sample, for example. Here we divide the frequencies in each column by the number of cases in that column,

$$Pct_{Col} = 100 * \frac{f}{N_{Col}} = 47.02\%. \quad (21.4)$$

In all, the frequency distribution and its bivariate format, the crosstab, have the ability to provide a wealth of information. Thus, while somewhat limited – particularly in terms of multivariate considerations – it is good practice to begin any statistical analysis here.

Difference of Means

In social science we are often concerned with differences between groups. For example, do Republicans differ from Democrats with respect to how religious they are? The basic process involved in answering a question of this nature – where we are interested in the extent to which two samples resemble each other on some variable – is simple enough. First, we establish a hypothesis about the population. Second, we collect a sample. Next, we check to see how likely the sample results are given our hypotheses about the population. Finally, we reject or fail to reject the null hypothesis based on our confidence level.

When testing hypotheses, we typically talk about testing the null hypothesis. In this case, a null hypothesis says that the two samples are drawn from equivalent populations. That is, any difference between two samples is due to a chance occurrence or sampling error. In line with our notation, we symbolize it as

$$\mu_1 = \mu_2, \quad (21.5)$$

where μ_1 is the mean of the first population and μ_2 is the mean of the second population. Thus, in our example the null hypothesis would be that Republicans and Democrats are equally religious (or that there is no difference between them in terms of religiosity). Remember this does not mean that we are denying the difference in sample means, but that we are instead attributing that difference to sampling error when we retain the null hypothesis (i.e., we are unable to reject the null hypothesis).

If the null is retained our data suggests that there is no relationship between our variables. Of course, we as social scientists often want to establish relationships. The process we subscribe to begins with the presumption that relationships do not exist. That is, establishing differences between groups is often the rationale for research – even though failing to disprove the null is sometimes more informative, or more theoretically intriguing, than rejecting the null. If we reject the null, we cannot rule out the research hypothesis that a true population difference exists. In this case the two samples appear to have been taken from populations having different means. Or more precisely stated, the difference between sample means is too large to be accounted for by sampling error. We symbolize this difference in means as

$$\mu_1 \neq \mu_2. \quad (21.6)$$

In the previous chapter we saw how to construct a sampling distribution of mean scores. In order to understand whether we can expect a difference between sample means to be due to chance or a true population difference between the two groups, we now consider the construction of a **sampling distribution of differences between means**. This frequency distribution is just like those we have explored earlier, except that the frequency is based on a series of differences between sample means randomly drawn from a given population.

We want to make a probability statement about the occurrence of different scores in the sampling distribution of differences between means. In the past we have relied on known probability distributions, like the normal curve, to make probability statements. We do so again here. If we can assume that this sampling distribution of differences of means is distributed normally we can make statements of probability. Assuming normality we know the general characteristics of our distribution of differences between means.

But does it make sense to think of the sampling distribution of differences between means as a normal distribution? Instead of just taking a singular random sample, think again about what would happen if we took a series of random samples and made a distribution of differences between means. Consider, for example, the hypothetical data in Table 21.3 which occurs from repeatedly taking two samples, calculating the mean for each and then the difference between means. For the purposes of this example, assume also that we know the population mean. If the null hypothesis is correct then the two samples should look the same. Any difference between the population mean and any sample mean should be due purely to sampling error. Thus a distribution of differences between means would look approximately normal if we wanted to retain the null hypothesis. That is,

Table 21.3 Distribution of differences between means

Differences	Frequency
5	1
4	2
3	5
2	7
1	10
0	18
–1	10
–2	8
–3	5
–4	3
–5	1

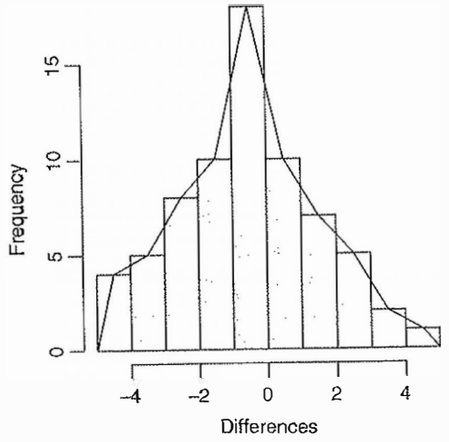

Figure 21.1 Histogram with polygon of differences between means

if there was truly no difference between means due to actual differences in the population, then the differences that do show up in the samples should look like random fluctuations about the mean with most scores close to the mean and few scores in the tail. Indeed the difference in means should overestimate and underestimate the mean in roughly equal numbers. In addition, the mean of the difference in means should be close to zero as the true central tendency of cases is to have no difference between the sample means. Figure 21.1 plots the histogram of differences overlayed with a polygon. The polygon shows that after only 70 draws from each sample we can already see a shape that somewhat resembles the normal curve, which we should expect given the central limit theorem. Moreover, the mean is zero, which suggests the samples are very similar to each other.

Again, note that we do not take many samples from a population in practice. Given what we know about the population and the normal curve, our reasoning for rejecting or retaining the null hypothesis can be constructed in terms of the score's distance to the mean – in this case the difference between means. If the difference of means that we found lies so far from the mean of differences between means for the null hypothesis (i.e., 0) that it only has a small probability of occurrence in the sampling distribution of differences between means, we reject the null. Contrarily, if the difference of means falls close to the mean of differences between means such that the probability of its occurrence in the sampling distribution of differences between means is high, we find ourselves unable to reject the null.

As we have done in the past, we need to transform our parameter of interest into a standardized unit to determine where it falls on the distribution. In this case we are dealing with sample mean differences that we need to translate into standardized units, so we calculate it accordingly:

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sigma_{\overline{X}_1 - \overline{X}_2}}, \tag{21.7}$$

where $\overline{X}_1 - \overline{X}_2$ is the difference between the mean of the first sample and the mean of the second sample. We assume 0 for the mean of the sampling distribution of differences between means based on our null hypothesis:

$$\mu_1 - \mu_2 = 0. \tag{21.8}$$

We rarely have knowledge of the standard deviation of the distribution of mean differences; again, it is too costly to draw enough pairs of sample means from the population to calculate it. Moreover, in practical research we can rarely assume that our sample sizes or variances are equal. Not unlike our problem with the standard deviation in the sampling distribution of means then, we need an estimate for the standard deviation that combines information from both samples. That is, the variance and sample size need to be accounted for to give us an idea of how different $\overline{X}_1$ is from $\overline{X}_2$ due to sampling error alone. We therefore calculate an approximation of it from the two samples that we actually draw. We will call this our **standard error of the difference between means**:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}. \tag{21.9}$$

With the standard error we can rewrite the test statistic for the difference of means test in terms of how it is used in actual research:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}. \tag{21.10}$$

There are a few variants on the difference of means test. One common version uses the independent variable to break observations into groups over time.

A second variant involves the differences between proportions. The procedures for comparing scores between the same group tested twice (i.e., panel data) and for comparing proportions involve different assumptions and slightly different formulae than that above, where we are testing mean differences in two different populations. While we will not cover this material here, you should be aware of the difference. Also note that the formulae above can be simplifed when we can assume equal sample sizes or equal variances. We have not made those assumptions with this data, which is more common in observational data.

Parametric Models

The statistical analysis above (difference of means) as well as some of the others we introduce below (correlation and regression) assume that the distributions of the variables being assessed belong to a large collection of known parameterized families of probability distributions. A parameter is just a characteristic of a population that we can use to describe the distribution. For example, in the case of the difference of means, we rely on the normal distribution, with its familiar parameters of μ and σ. Thus, we call all tests of this nature **parametric**.

While largely beyond the scope of this book, it is important to note that there also exist **nonparametric** models. These models similarly employ a mathematical procedure for hypothesis testing, but, unlike parametric statistics, they make no assumptions about the underlying distributions of the variables. Here the model structure is not specified at the onset by assuming a known probability distribution, but, instead, determined by the data. For example, a histogram is a simple nonparametric estimate of a probability distribution. As such, the term nonparametric is not meant to imply that such models completely lack parameters but that the number and nature of the parameters are flexible and not fixed in advance. There are also **semiparametric** models that have both parametric and nonparametric components.

Generally speaking, nonparametric tests have less **statistical power** than the appropriate parametric tests (though this depends on the kind of nonparametric test), but are more robust when the assumptions underlying the parametric test are not satisfied. Power refers to the probability of rejecting the null hypothesis when it is truly false. The results of a parametric test for a sample that does not appropriately match the assumed distribution are not meaningful. In these cases we should rely on nonparametric tests.[187]

Example: Group Membership and Ideology

Let us take an example related to questions of social capital. We have a hypothesis that liberals will belong to more social groups than conservatives. That is, liberals are more likely than conservatives to be members in formal groups, like civic and social groups, professional associations, and political organizations. The null

Table 21.4 Summary statistics of group membership by ideology

Liberals	Conservatives
$N_1 = 25$	$N_2 = 37$
$\bar{X}_1 = 60$	$\bar{X}_2 = 49$
$s_1 = 8$	$s_2 = 7$

hypothesis is thus that the $\mu_{liberals} = \mu_{conservatives}$. Our research hypothesis is that $\mu_{liberals} > \mu_{conservatives}$. However, for ease of presentation, let us agree that the opposite result (conservatives are more involved) is also a matter of interest and good possibility. Thus we are simply testing whether a relationship exists between ideology and group membership. In this case, our research hypothesis is that $\mu_{liberals} \neq \mu_{conservatives}$.

Table 21.4 provides the summary statistics for the data on liberal and conservative opinions on group membership. Here, the dependent variable is a multi-item index of engagement in formal group activities ranging from 0 to 100. The independent variable is a simple liberal (1) or conservative (0) dichotomy. We are dealing with an independent variable that is nominal and a dependent variable that is continuous, which, along with the hypothesis, make it appropriate for a difference of means test.

We begin by calculating the standard error of the difference between means:

$$\begin{aligned} s_{\bar{X}_1 - \bar{X}_2} &= \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \\ &= \sqrt{\frac{64}{25} + \frac{49}{37}} \\ &= \sqrt{3.88} \\ &= 1.97. \end{aligned} \tag{21.11}$$

If we test the hypothesis that there is a difference between liberals and conservatives on group membership we need to use a t-test to see if the samples are truly different in this case, not a z-score. Recall that z-scores are limited to situations in which we know the true population standard deviation or we have very large distributions, which is not the case in this example. Here we are estimating each σ from our s_1 and s_2, so we use t-tests.

Recall the t-ratio (this time with our new standard error):

$$\begin{aligned} t &= \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \\ &= \frac{60 - 51}{1.97} \\ &= 4.57. \end{aligned} \tag{21.12}$$

Next, we check our t-ratio with the values in Appendix A.2. In this case our degrees of freedom involve two samples, so we have:

$$\begin{aligned} df &= (N_1 - 1) + (N_2 - 1) \\ &= N_1 + N_2 - 2 \\ &= 25 + 37 - 2 \\ &= 60. \end{aligned} \tag{21.13}$$

For a df of 60, compare 4.57 to the critical values corresponding to the chosen α; 4.57 is larger than the critical value of 2 for the conventional $\alpha = 0.05$ as well as those for the stricter cutoffs, including $\alpha = 0.001$. We believe 0.05 is strict enough for a hypothesis of this nature (and of any nature in this book), and so we reject the null hypothesis. According to these results it is very unlikely that liberals and conservatives come from the same population with respect to their group membership activity.

In addition, we can analyze our data by appealing to confidence intervals. We have an observed difference of 9 (60 – 51), a standard error of 1.97, and degrees of freedom of 60, so we can state that the population difference of means should fall between our confidence interval. Recall, that this is simply the Observed Difference ± Critical Value × Standard Error:

$$\begin{aligned} CI &= 9 \pm 2.00 \times 1.97 \\ &= 9 \pm 3.94 \\ &= [5.06, 12.94]. \end{aligned} \tag{21.14}$$

Notice that this tells us that the population difference must be positive, not zero, since the range of values does not include zero. Thus, we reject the null that the population mean difference is zero. Again, we will only reject the null if zero is not included within our confidence interval.

Correlation

We began our exploration of statistical inference by generalizing differences we find in samples to differences in populations (see Chapter 20). Using data on sample differences, we described differences in populations with a particular level of certainty. In this chapter we took a similar approach to consider the difference of means on a continuous variable across groups as indicated by a nominal level variable. Finding a statistically significant relationship (rejecting the null hypothesis at a particular level of confidence) indicates that the extent of the difference in means is unlikely to be the result of sampling error.

With correlation we move from considering the relationship between an interval and nominal level variable, as with difference of means, to the relationship between two interval variables. In this context, correlation provides a measure of the strength of the relationship. Thus, **correlation** is a measure of association between two or more variables. Fortunately, correlation is conceptually straightfoward as it can be thought of as an extension of our data visualization tool, the scatter plot.

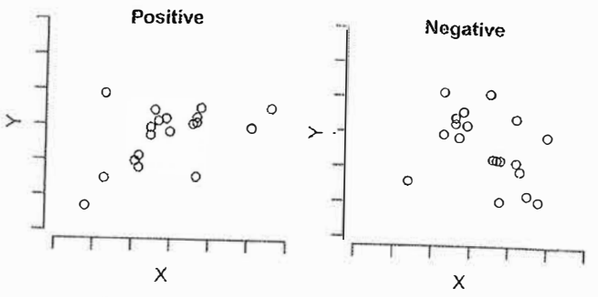

Figure 21.2 Scatter plots and direction of correlation

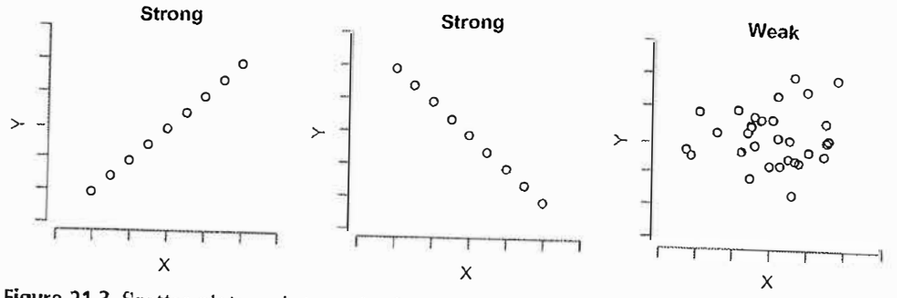

Figure 21.3 Scatter plots and strength of correlation

Visually, a scatter plot should give us a good idea whether our null hypothesis of no correlation can be rejected. Recall that in scatter plots we typically locate the X variable values along the x-axis (the horizontal base) and the Y variable values along the y-axis (the vertical base). Correlation consists of two components: strength and direction. The direction of correlation can be either negative or positive. If high scores on X correspond to low scores on Y, we have negative correlation. If high scores on X correspond to high scores on Y, we have positive correlation. In other words, if most points are in the bottom left and top right, a positive correlation is plausible, as in the left panel in Figure 21.2. If most points are in the top left and bottom right, as in the right panel, a negative correlation is plausible.

The strength of correlation can be strong or weak. If the scatter plot looks like a straight line, as in the left two panels of Figure 21.3, we are likely to have a strong correlation. If we get something that looks like a cloud of points, as in the far right panel, we have a weak correlation. Thus a correlation is strong if for a unit change in variable X, we can expect a specific change in variable Y in a particular direction (positive or negative). A correlation is weak if for a unit change in variable X, we are not sure what the change would be in variable Y. In other words, for a strong correlation we can look at X and predict the changes in Y; for a weak correlation we have a harder time doing that.

Scatter plots, however, do not allow us to make statistical inferences. Recall that statistical inference is the process of inferring from a sample to a population. In order to make statistical inferences, first we need to create a test statistic of

association. The most common statistic of association between two interval level variables and the one we discuss below is **Pearson's** r. Then we need a statistical procedure for evaluating the significance of this test statistic. To reiterate, the size of Pearson's r does not by itself allow us to draw a conclusion about statistical significance. There are two separate though related analyses to conduct to arrive at statistical inference. Thus, although the formulae differ, the logic of statistical inference here is much the same as in the difference of means test above.

To understand Pearson's r, we first need to understand how we compute the **covariance** between two variables. Covariance is a measure of how much two variables change together or "covary":

$$\text{Cov} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{N - 1}. \tag{21.15}$$

As is evident from the formula, we use the deviations to measure change in each variable. By multiplying corresponding deviations we see that the larger the corresponding deviations the larger the covariance. Notice that covariance could be negative as well. Because we have to add the combined deviations together to get the sense of how the two variables vary together, we need to account for the fact that the sum of the combined deviations could be large due to the simple fact that the sample size is large. Thus we control for sample size and divide by $N - 1$, instead of just N, since the population mean is unknown and we are relying on the sample mean to estimate it.

While it appears that we can use the covariance as our measure of correlation, using the covariance formula may be misleading at times. What if we have a large amount of deviation in the X variable and not much deviation in the Y variable? Such will give us a large covariance. Alternatively, if we have a moderate amount of deviation in the X variable and a moderate amount of deviation in Y variable, this will give us covariance of a reasonable size. However, this does not necessarily mean that in the first case the variables correlate stronger than in the second case. All this suggests that our measure of correlation needs to adjust our covariance by the amount of deviation present in each variable. A good measure of deviation in a variable is the standard deviation. Thus we can gauge the degree of association between the two variables by dividing covariance by the product of standard deviations.

Given the logic above, we can get a particular measure of correlation simply by dividing the covariance over the combined standard deviations and simplifying. Thus, Pearson's r is:

$$\begin{aligned} r &= \frac{\dfrac{\Sigma(X - \overline{X})(Y - \overline{Y})}{N - 1}}{\sqrt{\dfrac{\Sigma(X - \overline{X})^2}{N}} \times \sqrt{\dfrac{\Sigma(Y - \overline{Y})^2}{N}}} \\ &= \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2 \Sigma(Y - \overline{Y})^2}}. \end{aligned} \tag{21.16}$$

The formula for Pearson's r limits the range of values the statistic can take to falling between −1.00 and 1.00, with larger absolute values representing stronger correlations. Values of −1.00 or 1.00 represent a perfect linear relationship (i.e., you could draw a straight line and every point would fall on it). If the correlation coefficient is negative, then we have a negative correlation. If it is positive, we have a positive correlation. If the correlation is 0, we have the weakest possible correlation (no correlation, think of a cloud of points). Pearson's r therefore allows us to make statements that smaller values indicate weaker relationships. Although there is no universally accepted rule of thumb for distinguishing between strong and weak relationships, the closer to 1 or −1 the stronger, and the closer to 0 the weaker.

Note that the value of r does not tell us the slope of the best fitting line through our scatter plot, which we will discuss below with regression. An r of $|1|$ – called the **absolute value** of 1; i.e., the non-negative value of 1 – for example, does not mean that for a unit increase in one variable there is a full unit increase in another. Instead it conveys that there is no variation between the data points and the line of best fit. That is, we can have different slopes in the scatter plot with equivalent values of r.

Of course, we are typically working with samples, not populations. That is, using our sample data we would like to say whether or not the variables in the populations correlate. Here the null hypothesis refers to the situation in which the population characteristics are not correlated. Thus the null hypothesis is that $r = 0$, while our alternative hypothesis is that $r \neq 0$. It is worth explicitly restating our null hypothesis: there is no linear relationship between X and Y. Statistical theory tells us that the critical value associated with the test statistic represents the probability of finding this value of r as or more extreme than what you would get if no linear relationship actually exists. As before, we are willing to reject the null if this probability (p-value) is less than 0.05; or equivalently, if our test statistic exceeds the critical value for t.

Note that in practice one might provide a correlation coefficient as a summary statistic of the data without testing whether it is statistically different from zero. However, hypothesis testing with Pearson's r has some basic assumptions. There should be a straight line (not curvilinear) relationship between two variables. One may also detect nonlinear relationships between variables with this approach, but Pearson's r cannot be used to test these relationships, which would require a different test statistic. Second, the variables should be measured at the interval level and normally distributed, though the latter is of less importance in reasonably large samples since we can invoke the central limit theorem. Finally, random sampling is needed to allow us to generalize from the sample to the population.

In our example below we will present the relationship between two variables, issue dimensions and political parties, in terms of an independent and dependent variable. Correlation, however, does not require the specification of an independent and dependent variable. Correlation is simply a statement about association, not about causation. For instance, we might examine the correlation between two

independent variables. If we control for two highly correlated independent variables at the same time, this causes statistical problems for multivariate regression. This very high correlation is called **collinearity**, and complicates estimation, which we will talk about more in the context of multiple regression.

Correlation can also be a useful tool in the operationalization process. If we want to use multiple measures to get at the same concept, there should be some correlation between the measures. This **construct validity** suggests that valid measures should be correlated with related features of the concept. On the other hand, if correlation is very high, using both measures may be redundant.

Outside of these assumptions, the general process of calculating the correlation coefficient and testing statistical significance is similar to what we did for the difference of means. We begin by calculating our means and standard deviations and plugging them into the (intuitive) formula for r. We next calculate our test statistic, t-ratio, to generalize from our sample to the population. In the case of correlation we use

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}. \tag{21.17}$$

To test the null we also need to state our level of confidence and calculate the degrees of freedom. In correlation with two samples the degrees of freedom is $N - 2$. As per usual, we then go to the back of the book – i.e., find the t-ratio at the appropriate confidence level and degrees of freedom in the Appendix Table A.2 – to check whether the t-ratio is larger than the critical value for t. If it is, we reject the null of no correlation. If it is not, we fail to reject the null.

In closing, it is worth repeating a familiar statistics mantra: correlation is not causation. As we noted, there are a number of reasons why we might find a correlation without ever expecting there to be a causal relationship. Even if we do expect a causal relationship, we still must have the proper temporal sequence, and rule out concerns such as antecedent variable and spurious correlation to be confident in a causal relationship.

Example: Issue Dimensions and Parties

In democratic countries are multiple issue dimensions associated with a greater number of parties? That is, should we expect more parties when there are more issue dimensions on which parties can contend for power? The concept of issue dimensions suggests that there is more than a single left/right ideological dimension to politics and instead politics can involve multiple dimensions, including socioeconomic and sociocultural issues. A straightforward hypothesis stemming from these classic questions is the expectation of a positive correlation between issue dimensions and parties.

For this exercise we have some hypothetical sample data from a population of democratic countries that provides us with the two variables necessary to test the

Table 21.5 Issue dimensions and parties data

	Issue dimensions	Parties
Switzerland	3.73	4.84
Italy	2.99	4.90
Netherlands	3.25	4.65
France	2.76	4.03
Portugal	2.32	3.43
Germany	3.24	3.43
Spain	2.11	2.76
United Kingdom	2.75	2.99
Mean	2.89	3.88

Table 21.6 Calculating r for issue dimensions and parties

	X Deviation	Y Deviation	Product	X Dev^2	Y Dev^2
Switzerland	0.84	0.96	0.81	0.71	0.92
Italy	0.1	1.02	0.10	0.01	1.04
Netherlands	0.36	0.77	0.28	0.13	0.59
France	–0.13	0.15	–0.02	0.02	0.02
Portugal	–0.57	–0.45	0.26	0.32	0.20
Germany	0.35	–0.45	–0.16	0.12	0.20
Spain	–0.78	–1.12	0.87	0.61	1.25
United Kingdom	–0.14	–0.89	0.12	0.02	0.79
Sum of products			$SP = 2.26$		
Sum of squares				$SS_X = 1.94$	$SS_Y = 5.03$

correlation, one that notes the number of issue dimensions and another the number of parties. In order to test the correlation between these two variables, we first calculate the Pearson's correlation coefficient (r) then the t-ratio to arrive at a test of statistical significance.

We begin by calculating the distances between the raw values and its mean value for each variable, which are called the X deviation and Y deviation, respectively. As shown in Table 21.6, to obtain the X deviation, we simply subtract $\overline{X}$ from the X value for that observation. We do the same for Y. Deviations can be illustrated within a scatter plot that includes a vertical line at the mean value of X and a horizontal line at the mean value of Y. Comparing Figure 21.4 to Table 21.6 we see that points or countries in the top right corner will have two positive deviations, and those in the bottom left corner will have two negative deviations. Points in the top left corner will have a positive Y deviation and negative X deviation, and those in the bottom right corner will have a negative Y deviation and positive X deviation.

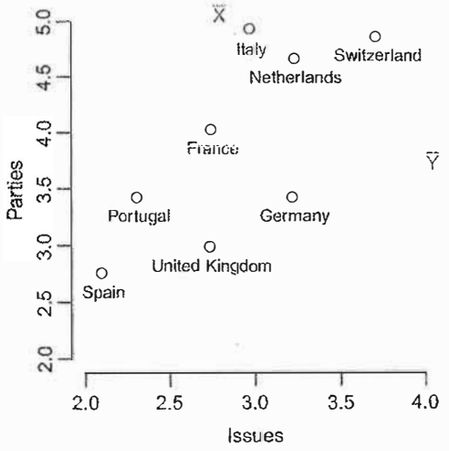

Figure 21.4 Scatter plot of parties and issues

Next, for each observation, we multiply the X deviation by the Y deviation. If this product is positive, the observation is consistent with a positive relationship; if the product is negative, this observation is consistent with a negative relationship. Then we sum these values. This is the **sum of products** (SP), the numerator of our formula for r. So far, we can note two things. The sum of products is positive, which means the relationship is positive. All but one of the individual products is positive, which means that the relationship appears to be fairly consistent. However, the scale of this output is unrelated to the values of the variables, so we cannot make a statement about the strength of association yet.

By calculating the denominator, we succeed in constraining our test statistic to the range $[-1, 1]$. We can now interpret the strength of the relationship. The denominator calculates the deviations of each variable in relation to its own mean, but without respect to the other variable. In other words, we are calculating something akin to the variance of each variable (except that we do not divide by N), which we call the **sum of squares**: SS_X for X and SS_Y for Y.

We now have all the elements we need for our formula:

$$\begin{aligned} r &= \frac{SP}{\sqrt{SS_X \times SS_Y}} \\ &= \frac{2.26}{\sqrt{1.94 \times 5.03}} \\ &= 0.72. \end{aligned} \tag{21.18}$$

What, then, do we make of this value of 0.72? We know there is a positive relationship between issue dimensions and number of parties, and that the relationship is quite strong. In giving a substantive interpretation of correlation, these are the two necessary elements: direction and strength.

But is it statistically significant? The same value of r may or may not be statistically significant depending on the sample size. So, answering this question is a two-step process. First, we must translate r into a t-statistic, using a formula that involves only r and N. To calculate t, we simply need our value of r and our number of observations

$$\begin{aligned} t &= \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \\ &= \frac{0.72\sqrt{8-2}}{\sqrt{1-0.72^2}} \\ &= \frac{1.76}{0.69} \\ &= 2.54. \end{aligned} \tag{21.19}$$

Then we just need to calculate our degrees of freedom to determine the critical value. The degrees of freedom are simply:

$$\begin{aligned} df &= N-2 \\ &= 6. \end{aligned} \tag{21.20}$$

With six degrees of freedom, our critical value for t is 2.45. Thus, with our t-ratio of 2.54, we can reject the null.

CONCLUSIONS

Social scientists are typically interested in the relationships between two or more variables. Above we have introduced two bivariate hypothesis tests for continuous dependent variables. In quantitative analyses understanding the level of measurement is essential because it helps us decide which hypothesis test to use. The first test, difference of means, allows researchers to make comparisons of sample means. The second, correlation, moves us beyond making a simple claim of a relationship or no relationship between two variables to a measure that conveys both the strength and direction of the relationship.

KEY TERMS

- Discrete variable
- Continuous variable
- Marginals
- Cross-tabulation (crosstab)
- Sampling distribution of differences between means
- Standard error of the difference between means
- Parametric