7 Bivariate Hypothesis Testing

OVERVIEW

Once we have set up a hypothesis test and collected data, how do we evaluate what we have found? In this chapter we provide hands-on discussions of the basic building blocks used to make statistical inferences about the relationship between two variables. We deal with the often-misunderstood topic of "statistical significance" – focusing both on what it is and what it is not – as well as the nature of statistical uncertainty. We introduce three ways to examine relationships between two variables: tabular analysis, difference of means tests, and correlation coefficients. (We will introduce a fourth technique, bivariate regression analysis, in Chapter 8.)

7.1 BIVARIATE HYPOTHESIS TESTS AND ESTABLISHING CAUSAL RELATIONSHIPS

In the preceding chapters we introduced the core concepts of hypothesis testing. In this chapter we discuss the basic mechanics of hypothesis testing with three different examples of bivariate hypothesis testing. It is worth noting that, although this type of analysis was the main form of hypothesis testing in the professional journals up through the 1970s, it is seldom used as the *primary* means of hypothesis testing in the professional journals today.[1] This is the case because these techniques are good at helping us with only the first principle for establishing causal relationships. Namely, bivariate

[1] By definition, researchers conducting bivariate hypothesis tests are making one of two assumptions about the state of the world. They are assuming either that there are no other variables that are causally related to the dependent variable in question, or that, if there are such omitted variables, they are unrelated to the independent variable in the model. We will have much more to say about omitting independent variables from causal models in Chapter 9. For now, bear in mind that, as we have discussed in previous chapters, these assumptions rarely hold when we are describing the political world.

hypothesis tests help us to answer the question, "Are X and Y related?" By definition – "bivariate" means "two variables" – these tests cannot help us with the important question, "Have we controlled for all confounding variables Z that might make the observed association between X and Y spurious?"

Despite their limitations, the techniques covered in this chapter are important starting points for understanding the underlying logic of statistical hypothesis testing. In the sections that follow we discuss how one chooses which bivariate test to conduct and then provide detailed discussions of three such tests. Throughout this chapter, try to keep in mind the main purpose of this exercise: We are attempting to apply the lessons of the previous chapters to real-world data. We will eventually do this with more appropriate and more sophisticated tools, but the lessons that we learn in this chapter will be crucial to our understanding of these more advanced methods. Put simply, we are trying to get up and walk in the complicated world of hypothesis testing with real-world data. Once we have mastered walking, we will then begin to work on running with more advanced techniques.

7.2 CHOOSING THE RIGHT BIVARIATE HYPOTHESIS TEST

As we discussed in previous chapters, and especially in Chapters 5 and 6, researchers make a number of critical decisions before they test their hypotheses. Once they have collected their data and want to conduct a bivariate hypothesis test, they need to consider the nature of their dependent and independent variables. As we discussed in Chapter 5, we can classify variables in terms of the types of values that cases take on. Table 7.1 shows four different scenarios for testing a bivariate hypothesis; which one is most appropriate depends on the variable type of the independent variable and the dependent variable. For each case, we have listed one or more

Table 7.1. Variable types and appropriate bivariate hypothesis tests

		Independent variable type	
		Categorical	**Continuous**
Dependent variable type	Categorical	*Tabular analysis*	Probit/logit (Ch.11)
	Continuous	*Difference of means*	*Correlation coefficient*; bivariate regression model (Ch. 8)

Note: Tests in italics are discussed in this chapter.

appropriate type of bivariate hypothesis tests. In cases in which we can describe both the independent and dependent variables as categorical, we use a form of analysis referred to as **tabular analysis** to test our hypothesis. When the dependent variable is continuous and the independent variable is categorical, we use a **difference of means** test. When the independent variable is continuous and the dependent variable is categorical, analysts typically use either a probit or logit model. (These types of statistical models are discussed in Chapter 11.) Finally, when both the dependent and independent variables are continuous, we use a **correlation coefficient** in this chapter, and, in Chapter 8, we will discuss the bivariate regression model.

7.3 ALL ROADS LEAD TO *p*

One common element across a wide range of statistical hypothesis tests is the *p*-value (the *p* stands for "probability.") This value, ranging between 0 and 1, is the closest thing that we have to a bottom line in statistics. But it is often misunderstood and misused. In this section we discuss the basic logic of the *p*-value and relate it back to our discussion in Chapter 6 of using sample data to make inferences about an underlying population.

7.3.1 The Logic of *p*-Values

If we think back to the four principles for establishing causal relationships that were discussed in Chapter 3, the third hurdle is the question "Is there covariation between X and Y?" To answer this question, we need to apply standards to real-world data for determining whether there appears to be a relationship between our two variables, the independent variable X and the dependent variable Y. The tests listed in the cells in Table 7.1 are commonly accepted tests for each possible combination of data type. In each of these tests, we follow a common logic: We compare the actual relationship between X and Y in sample data with what we would expect to find if X and Y *were not* related in the underlying population. The *more different* the empirically observed relationship is from what we would expect to find if there were *not* a relationship, the more confidence we have that X and Y are related in the population. The logic of this inference from sample to population is the same as what we used in Chapter 6 to make inferences about the population mean from sample data.

The statistic that is most commonly associated with this type of logical exercise is the ***p*-value**. The *p*-value, which ranges between 0 and 1, is the probability that we would see the relationship that we are finding because of random chance. Put another way, the *p*-value tells us the probability that we would see the observed relationship between the two variables in our sample data if there were truly no relationship between them in the unobserved population. Thus, the lower the p-value, the greater confidence we have that there *is* a systematic relationship between the two variables for which we estimated the particular p-value.

One common characteristic across most statistical techniques is that, for a particular measured relationship, the more data on which the measurement is based, the lower our p-value will be. This is consistent with one of the lessons of Chapter 6 about sample size: The larger the sample size, the more confident we can be that our sample will more accurately represent the population.[2] (See Subsection 6.4.2 for a reminder.)

7.3.2 The Limitations of p-Values

Although p-values are powerful indicators of whether or not two variables are related, they are limited. In this subsection we review those limitations. It is important that we also understand what a p-value is not: The logic of a p-value is not reversible. In other words, $p = .001$ does not mean that there is a .999 chance that something systematic is going on. Also, it is important to realize that, although a p-value tells us something about our confidence that there is a relationship between two variables, it does not tell us whether that relationship is causal.

In addition, it might be tempting to assume that, when a p-value is very close to zero, this indicates that the relationship between X and Y is very *strong*. This is not necessarily true (though it might be true). As we previously noted, p-values represent our degree of confidence that there is a relationship in the underlying population. So we should naturally expect smaller p-values as our sample sizes increase. But a larger sample size does not magically make a relationship stronger; it *does* increase our confidence that the observed relationship in our sample accurately represents the underlying population. We saw a similar type of relationship in Chapter 6 when we calculated standard errors. Because the number of cases is in the denominator of the standard error formula, an increased number of cases leads to a smaller standard error and a more narrow confidence interval for our inferences about the population.

Another limitation of p-values is that they do not directly reflect the quality of the measurement procedure for our variables. Thus, if we are more confident in our measurement, we should be more confident in a particular p-value. The flip side of this is that, if we are not very confident

2 Also, the smaller the sample size, the more likely it is that we will get a result that is not very representative of the population.

in our measurement of one or both of our variables, we should be less confident in a particular p-value.

Finally, we should keep in mind that p-values are always based on the assumption that you are drawing a perfectly random sample from the underlying population. Mathematically, this is expressed as

$$p_i = P \; \forall i.$$

This translates into "the probability of an individual case from our population ending up in our sample, p_i, is assumed to equal P for all of the individual cases i." If this assumption were valid, we would have a truly random sample. Because this is a standard that is almost never met, we should use this in our assessment of a particular p-value. The further we are from a truly random sample, the less confidence we should have in our p-value.

7.3.3 From p-Values to Statistical Significance

As we outlined in the preceding subsection, lower p-values increase our confidence that there is indeed a relationship between the two variables in question. A common way of referring to such a situation is to state that the relationship between the two variables is **statistically significant**. Although this type of statement has a ring of authoritative finality, it is always a qualified statement. In other words, an assertion of statistical significance depends on a number of other factors. One of these factors is the set of assumptions from the previous section. "Statistical significance" is achieved only to the extent that the assumptions underlying the calculation of the p-value hold. In addition, there are a variety of different standards for what is a statistically significant p-value. Most social scientists use the standard of a p-value of .05. If p is less than .05, they consider a relationship to be statistically significant. Others use a more stringent standard of .01, or a more loose standard of .1.[3]

We cannot emphasize strongly enough that finding that X and Y have a statistically significant relationship does *not* necessarily mean that the relationship between X and Y is strong or especially that the relationship is causal. To evaluate whether or not a relationship is strong, we need to use our substantive knowledge about what it means for the value of Y to change by a particular amount. We will discuss assessments of the strength of relationships in greater detail in Chapter 9. To evaluate the case for a

[3] More recently, there has been a trend toward reporting the estimated p-value and letting readers make their own assessments of statistical significance.

causal relationship, we need to evaluate how well our theory has performed in terms of all four of the causal hurdles from Chapter 3.

7.3.4 The Null Hypothesis and *p*-Values

In Chapter 1 we introduced the concept of the null hypothesis. Our definition was "A null hypothesis is also a theory-based statement but it is about what we would expect to observe if our theory were incorrect." Thus, following the logic that we previously outlined, if our theory-driven hypothesis is that there is covariation between X and Y, then the corresponding null hypothesis is that there is no covariation between X and Y. In this context, another interpretation of the p-value is that it conveys the level of confidence with which we can reject the null hypothesis.

7.4 THREE BIVARIATE HYPOTHESIS TESTS

We now turn to three specific bivariate hypothesis tests. In each case, we are testing for whether there is a relationship between X and Y. We are doing this with sample data, and then, based on what we find, making inferences about the underlying population.

7.4.1 Example 1: Tabular Analysis

Tabular presentations of data on two variables are still used quite widely. In the more recent political science literature, scholars use them as steppingstones on the way to multivariate analyses. It is worth noting at this point in the process that, in tables, most of the time the dependent variable is displayed in the rows whereas the independent variable is displayed in the columns. Any time that you see a table, it is very important to take some time to make sure that you understand what is being conveyed. We can break this into the following three-step process:

1. Figure out what the variables are that define the rows and columns of the table.
2. Figure out what the individual cell values represent. Sometimes they will be the number of cases that take on the particular row and column values; other times the cell values will be proportions (ranging from 0 to 1.0) or percentages (ranging from 0 to 100). If this is the case, it is critical that you figure out whether the researcher calculated the percentages or proportions for the entire table or for each column or row.
3. Figure out what, if any, general patterns you see in the table.

Table 7.2. Union households and vote in the 2008 U.S. presidential election

Candidate	Not from a union household	From a union household	Total
McCain	47.1	33.4	45.0
Obama	52.9	66.6	55.0
Total	100.0	100.0	100.0

Note: Cell entries are column percentages.

Let's go through these steps with Table 7.2. In this table we are testing the theory that affiliation with trade unions makes people more likely to support left-leaning candidates.[4] We can tell from the title and the column and row headings that this table is comparing the votes of people from union households with those not from union households in the 2008 U.S. presidential election. We can use the information in this table to test the hypothesis that voters from union households were more likely to support Democratic Party presidential candidate Barack Obama.[5] As the first step in reading this table, we determine that the columns indicate values for the independent variable (whether or not the individual was from a union household) and that the rows indicate values for the dependent variable (presidential vote). The second step is fairly straightforward; the table contains a footnote that tells us that the "cell entries are column percentages." This is the easiest format for pursuing step 3, because the column percentages correspond to the comparison that we want to make. We want to compare the presidential votes of people from union households with the presidential votes of people not from union households. The pattern is fairly clear: People from the union households overwhelmingly supported Obama (66.6 for Obama and 33.4 for McCain), whereas people from the nonunion households only marginally favored Obama (52.9 for Obama and 47.1 for McCain). If we think in terms of independent (X) and dependent (Y) variables, the comparison that we have made is between the distribution of the dependent variable ($Y =$ Presidential Vote) across values of the independent variable ($X =$ Type of Household).

[4] Take a moment to assess this theory in terms of the first two of the four hurdles that we discussed in Chapter 3. The causal mechanism is that left-leaning candidates tend to support policies favored by trade unions. Is this credible? What about hurdle 2? Can we rule out the possibility that support for left-leaning candidates make one more likely to be affiliated with a trade union?

[5] What do you think about the operationalization of these two variables? How well does it stand up to what we discussed in Chapter 5?

Table 7.3. Gender and vote in the 2008 U.S. presidential election: Hypothetical scenario

Candidate	Male	Female	Row total
McCain	?	?	45.0
Obama	?	?	55.0
Column total	100.0	100.0	100.0

Note: Cell entries are column percentages.

In Table 7.2, we follow the simple convention of placing the values of the independent variable in the columns and the values of the dependent variable in the rows. Then, by calculating column percentages for the cell values, this makes comparing across the columns straightforward. It is wise to adhere to these norms, because it is the easiest way to make the comparison that we want, and because it is the way many readers will expect to see the information.

In our next example we are going to go step-by-step through a bivariate test of the hypothesis that gender (X) is related to vote (Y) in U.S. presidential elections. To test this hypothesis about gender and presidential vote, we are going to use data from the 2008 National Annenberg Election Survey (NAES from here on). This is an appropriate set of data for testing this hypothesis because these data are from a randomly selected sample of cases from the underlying population of interest (U.S. adults). Before we look at results obtained by using actual data, think briefly about the measurement of the variables of interest and what we would expect to find if there was no relationship between the two variables.

Table 7.3 shows partial information from a hypothetical example in which we know that 45.0% of our sample respondents report having voted for John McCain and 55.0% of our sample respondents report having voted for Barack Obama. But, as the question marks inside this table indicate, we do not know how voting breaks down in terms of gender. If there were no relationship between gender and presidential voting in 2008, consider what we would expect to see given what we know from Table 7.3. In other words, what values should replace the question marks in Table 7.3 if there were no relationship between our independent variable (X) and dependent variable (Y)?

If there is not a relationship between gender and presidential vote, then we should expect to see no major differences between males and females in terms of how they voted for John McCain and Barack Obama. Because we know that 45.0% of our cases voted for McCain and 55.0% for Obama,

Table 7.4. Gender and vote in the 2008 U.S. presidential election: Expectations for hypothetical scenario if there were no relationship

Candidate	Male	Female	Row total
McCain	45.0	45.0	45.0
Obama	55.0	55.0	55.0
Column total	100.0	100.0	100.0

Note: Cell entries are column percentages.

Table 7.5. Gender and vote in the 2008 U.S. presidential election

Candidate	Male	Female	Row total
McCain	?	?	1,434
Obama	?	?	1,755
Column total	1,379	1,810	3,189

Note: Cell entries are number of respondents.

what should we expect to see for males and for females? We should expect to see the same proportions of males and females voting for each candidate. In other words, we should expect to see the question marks replaced with the values in Table 7.4. This table displays the expected cell values for the null hypothesis that there is no relationship between gender and presidential vote.

Table 7.5 shows the total number of respondents who fit into each column and row from the 2008 NAES. If we do the calculations, we can see that the numbers in the rightmost column of Table 7.5 correspond with the percentages from Table 7.3. We can now combine the information from Table 7.5 with our expectations from Table 7.4 to calculate the number of respondents that we would expect to see in each cell if gender and presidential vote were unrelated. We display these calculations in Table 7.6. In Table 7.7, we see the actual number of respondents that fell into each of the four cells.

Finally, in Table 7.8, we compare the observed number of cases in each cell (O) with the number of cases that we would expect to see if there were no relationship between our independent and dependent variables (E).

We can see a pattern. Among males, the proportion observed voting for Obama is lower than what we would expect if there were no relationship

Table 7.6. Gender and vote in the 2008 U.S. presidential election: Calculating the expected cell values if gender and presidential vote are unrelated

Candidate	Male	Female
McCain	(45% of 1,379) $= 0.45 \times 1,379 = 620.55$	(45% of 1,810) $= 0.45 \times 1,810 = 814.5$
Obama	(55% of 1,379) $= 0.55 \times 1,379 = 758.45$	(55% of 1,810) $= 0.55 \times 1,810 = 995.5$

Note: Cell entries are expectation calculations if these two variables are unrelated.

Table 7.7. Gender and vote in the 2008 U.S. presidential election

Candidate	Male	Female	Row total
McCain	682	752	1,434
Obama	697	1,058	1,755
Column total	1,379	1,810	3,189

Note: Cell entries are number of respondents.

Table 7.8. Gender and vote in the 2008 U.S. presidential election

Candidate	Male	Female
McCain	$O = 682; E = 620.55$	$O = 752; E = 814.5$
Obama	$O = 697; E = 758.45$	$O = 1,058; E = 995.5$

Note: Cell entries are the number observed (O); the number expected if there were no relationship (E).

between the two variables. Also, among men, the proportion voting for McCain is higher than what we would expect if there were no relationship. For females this pattern is reversed – the proportion voting for Obama (McCain) is higher (lower) than we would expect if there were no relationship between gender and vote for U.S. president. The pattern of these differences is in line with the theory that women support Democratic Party candidates more than men do. Although these differences are present, we have not yet determined that they are of such a magnitude that we should now have increased confidence in our theory. In other words, we want to know whether or not these differences are statistically significant.

To answer this question, we turn to the **chi-squared (χ^2) test for tabular association**. Karl Pearson originally developed this test when he was testing theories about the influence of nature versus nurture at the beginning of the 20th century. His formula for the χ^2 statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

The summation sign in this formula signifies that we sum over each cell in the table; so a 2×2 table would have four cells to add up. If we think about an individual cell's contribution to this formula, we can see the underlying logic of the χ^2 test. If the value observed, O, is exactly equal to the expected value if there were no relationship between the two variables, E, then we would get a contribution of zero from that cell to the overall formula (because $O-E$ would be zero). Thus, if all observed values were exactly equal to the values that we expect if there were no relationship between the two variables, then $\chi^2 = 0$. The more the O values differ from the E values, the greater the value will be for χ^2. Because the numerator on the right-hand side of the χ^2 formula $(O-E)$ is squared, any difference between O and E will contribute positively to the overall χ^2 value.

Here are the calculations for χ^2 made with the values in Table 7.8:

$$\begin{aligned}\chi^2 &= \sum \frac{(O-E)^2}{E}\\ &= \frac{(682-620.55)^2}{620.55} + \frac{(752-814.5)^2}{814.5} + \frac{(697-758.45)^2}{758.45} + \frac{(1{,}058-995.5)^2}{995.5}\\ &= \frac{3{,}776.1}{620.55} + \frac{3{,}906.25}{814.5} + \frac{3{,}776.1}{758.45} + \frac{3906.25}{995.5}\\ &= 6.09 + 4.8 + 4.98 + 3.92 = 19.79.\end{aligned}$$

So our calculated value of χ^2 is 19.79 based on the observed data. What do we do with this? We need to compare that 19.79 with some predetermined standard, called a **critical value**, of χ^2. If our calculated value is greater than the critical value, then we conclude that there is a relationship between the two variables; and if the calculated value is less than the critical value, we cannot make such a conclusion.

How do we obtain this critical value? First, we need a piece of information known as the **degrees of freedom** (df) for our test.[6] In this case, the df calculation is very simple: $\mathrm{df} = (r-1)(c-1)$, where r is the number of rows

[6] We define degrees of freedom in the next section.

in the table, and c is the number of columns in the table. In the example in Table 7.8, there are two rows and two columns, so $(2 - 1)(2 - 1) = 1$.

You can find a table with critical values of χ^2 in Appendix A. If we adopt the standard p-value of .05, we see that the critical value of χ^2 for df $= 1$ is 3.841. Therefore a calculated χ^2 value of 19.79 is well over the minimum value needed to achieve a p-value of .05. In fact, continuing out in this table, we can see that we have exceeded the critical value needed to achieve a p-value of .001.

At this point, we have established that the relationship between our two variables meets a conventionally accepted standard of statistical significance (i.e., $p < .05$). Although this result is supportive of our hypothesis, we have not yet established a causal relationship between gender and presidential voting. To see this, think back to the four hurdles along the route to establishing causal relationships that we discussed in Chapter 3. Thus far, we have cleared the third hurdle, by demonstrating that X (gender) and Y (vote) covary. From what we know about politics, we can easily cross hurdle 1, "Is there a credible causal mechanism that links X to Y?" Women might be more likely to vote for candidates like Obama because, among other things, women depend on the social safety net of the welfare state more than men do. If we turn to hurdle 2, "Can we rule out the possibility that Y could cause X?," we can pretty easily see that we have met this standard through basic logic. We know with confidence that changing one's vote does not lead to a change in one's gender. We hit the most serious bump in the road to establishing causality for this relationship when we encounter hurdle 4, "Have we controlled for all confounding variables Z that might make the association between X and Y spurious?" Unfortunately, our answer here is that we do not yet know. In fact, with a bivariate analysis, we cannot know whether some other variable Z is relevant because, by definition, there are only two variables in such an analysis. So, until we see evidence that Z variables have been controlled for, our scorecard for this causal claim is [*y y y n*].

7.4.2 Example 2: Difference of Means

In our second example, we examine a situation in which we have a continuous dependent variable and a categorical independent variable. In this type of bivariate hypothesis test, we are looking to see if the means are different across the values of the independent variable. We follow the basic logic of hypothesis testing: comparing our real-world data with what we would expect to find if there were no relationship between our independent and dependent variables. We use the sample means and standard deviations to make inferences about the unobserved population.

Our theory in this section will come from the study of parliamentary governments. When political scientists study phenomena across different forms of government, one of the fundamental distinctions that they draw between different types of democracies is whether the regime is parliamentary or not. A democratic regime is labeled "parliamentary" when the lower house of the legislature is the most powerful branch of government and directly selects the head of the government.[7] One of the interesting features of most parliamentary regimes is that a vote in the lower house of the legislature can remove the government from power. As a result, political scientists have been very interested in the determinants of how long parliamentary governments last when the possibility of such a vote exists.

One factor that is an important difference across parliamentary democracies is whether the party or parties that are in government occupy a majority of the seats in the legislature.[8] By definition, the opposition can vote out of office a minority government, because it does not control a majority of the seats in the legislature. Thus a pretty reasonable theory about government duration is that majority governments will last longer than minority governments.

We can move from this theory to a hypothesis test by using a data set produced by Michael D. McDonald and Silvia M. Mendes titled "Governments, 1950–1995." Their data set covers governments from 21 Western countries. For the sake of comparability, we will limit our sample to those governments that were formed after an election.[9] Our independent variable, "Government Type," takes on one of two values: "majority government" or

[7] An important part of research design is determining which cases are and are not covered by our theory. In this case, our theory, which we will introduce shortly, is going to apply to only parliamentary democracies. As an example, consider whether or not the United States and the United Kingdom fit this description at the beginning of 2007. In the United States in 2007, the head of government was President George W. Bush. Because Bush was selected by a presidential election and not by the lower branch of government, we can already see that the United States at the beginning of 2007 is not covered by our theory. In the United Kingdom, we might be tempted at first to say that the head of government at the beginning of 2007 was Queen Elizabeth II. But, if we consider that British queens and kings have been mostly ceremonial in UK politics for some time now, we then realize that the head of government was the prime minister, Tony Blair, who was selected from the lower house of the legislature, the House of Commons. If we further consider the relative power of the House of Commons compared with the other branches of government at the beginning of 2007, we can see that the United Kingdom met our criteria for being classified as parliamentary.

[8] Researchers usually define a party as being in government if its members occupy one or more cabinet posts, whereas parties not in government are in opposition.

[9] We have also limited the analyses to cases in which the governments had a legal maximum of four years before they must call for new elections. These limitations mean that, strictly speaking, we are only able to make inferences about the population of cases that also fit these criteria.

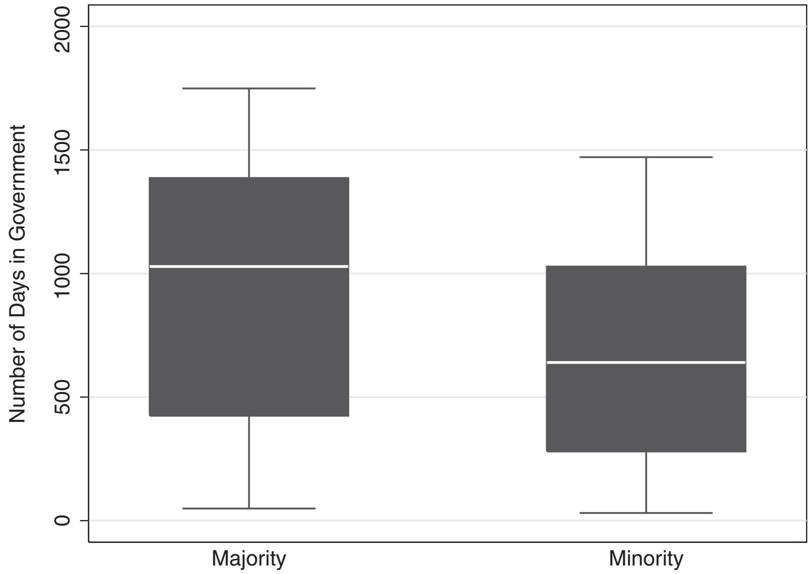

Figure 7.1. Box-whisker plot of Government Duration for majority and minority governments.

"minority government." Our dependent variable, "Government Duration," is a continuous variable measuring the number of days that each government lasted in office. Although this variable has a hypothetical range from 1 day to 1461 days, the actual data vary from an Italian government that lasted for 31 days in 1953 to a Dutch government that lasted for 1749 days in the late 1980s and early 1990s.

To get a better idea of the data that we are comparing, we can turn to two graphs that we introduced in Chapter 5 for viewing the distribution of continuous variables. Figure 7.1 presents a box-whisker plot of government duration for minority and majority governments, and Figure 7.2 presents a kernel density plot of government duration for minority and majority governments. From both of these plots, it appears that majority governments last longer than minority governments.

To determine whether the differences from these figures are statistically significant, we turn to a **difference of means test**. In this test we compare what we have seen in the two figures with what we would expect if there were no relationship between Government Type and Government Duration. If there were no relationship between these two variables, then the world would be such that the duration of governments of both types were drawn from the same underlying distribution. If this were the case, the mean or average value of Government Duration would be the same for minority and majority governments.

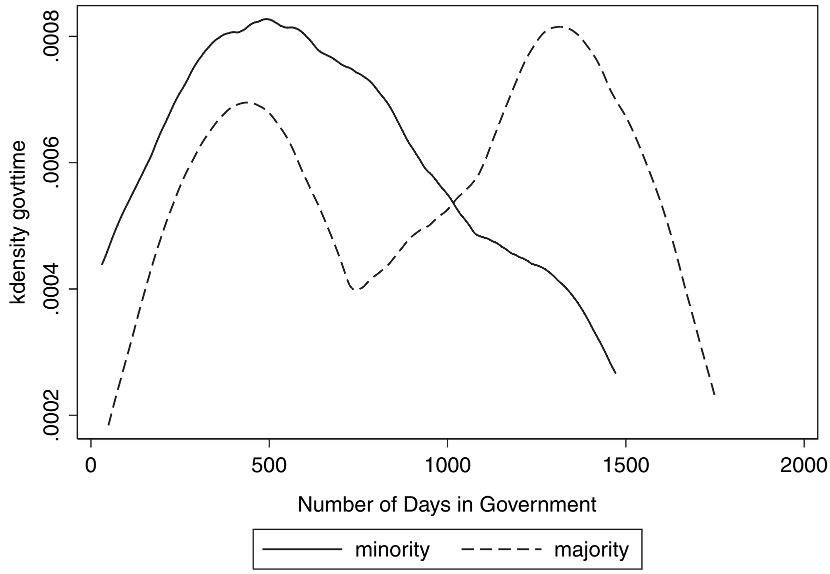

Figure 7.2. Kernel density plot of Government Duration for majority and minority governments.

To test the hypothesis that these means are drawn from the same underlying distribution, we use another test developed by Karl Pearson for these purposes. The test statistic for this is known as a t-test because it follows the t-distribution. The formula for this particular t-test is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\mathrm{se}(\bar{Y}_1 - \bar{Y}_2)},$$

where $\bar{Y}_1$ is the mean of the dependent variable for the first value of the independent variable and $\bar{Y}_2$ is the mean of the dependent variable for the second value of the independent variable. We can see from this formula that the greater the difference between the mean value of the dependent variable across the two values of the independent variable, the further the value of t will be from zero.

In Chapter 6 we introduced the notion of a standard error, which is a measure of uncertainty about a statistical estimate. The basic logic of a standard error is that the larger it is, the more uncertainty (or less confidence) we have in our ability to make precise statements. Similarly, the smaller the standard error, the greater our confidence about our ability to make precise statements.

To better understand the contribution of the top and bottom parts of the t-calculation for a difference of means, look again at Figures 7.1 and 7.2. The further apart the two means are and the less dispersed the distributions

Table 7.9. Government type and government duration

Government type	Number of observations	Mean duration	Standard deviation
Majority	124	930.5	466.1
Minority	53	674.4	421.4
Combined	177	853.8	467.1

(as measured by the standard deviations s_1 and s_2), the greater confidence we have that $\bar{Y}_1$ and $\bar{Y}_2$ are different from each other.

Table 7.9 presents the descriptive statistics for government duration by government type. From the values displayed in this table we can calculate the t-test statistic for our hypothesis test. The standard error of the difference between two means ($\bar{Y}_1$ and $\bar{Y}_2$), $\text{se}(\bar{Y}_1 - \bar{Y}_2)$, is calculated from the following formula:

$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)} \times \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

where n_1 and n_2 are the sample sizes, and s_1^2 and s_2^2 are the sample variances. If we label the number of days in government for majority governments Y_1 and the number of days in government for minority governments Y_2, then we can calculate the standard error as

$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\left(\frac{(124-1)(466.1)^2 + (53-1)(421.4)^2}{124 + 77 - 2}\right)} \times \sqrt{\left(\frac{1}{124} + \frac{1}{53}\right)}$$

$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = 74.39.$$

Now that we have the standard error, we can calculate the t-statistic:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{se}(\bar{Y}_1 - \bar{Y}_2)} = \frac{930.5 - 674.4}{74.39} = \frac{256.1}{74.39} = 3.44.$$

Now that we have calculated this t-statistic, we need one more piece of information before we can get to our p-value. This is called the **degrees of freedom** (df). Degrees of freedom reflect the basic idea that we will gain confidence in an observed pattern as the amount of data on which that pattern is based increases. In other words, as our sample size increases, we become more confident about our ability to say things about the underlying population. If we turn to Appendix B, which is a table of critical values for t, we can see that it reflects this logic. This table also follows the same basic logic as the χ^2 table. The way to read such a table is that the columns are defined by targeted p-values, and, to achieve a particular target p-value, you need to obtain a particular value of t. The rows in the t-table indicate the number of degrees of freedom. As the number of degrees of freedom goes up, the t-statistic we need to obtain a particular p-value goes down. We calculate the degrees of freedom for a difference of means t-statistic based on the sum of total sample size minus two. Thus our degrees of freedom is

$$n_1 + n_2 - 2 = 124 + 53 - 2 = 175.$$

From the p-value, we can look across the row for which df = 100 and see the minimum t-value needed to achieve each targeted value of p.[10] In the second column of the t-table, we can see that, to have a p-value of .10 (meaning that there is a 10%, or 1 in 10, chance that we would see this relationship randomly in our sample if there were no relationship between X and Y in the underlying population), we must have a t-statistic greater than or equal to 1.29. Because $3.44 > 1.29$, we can proceed to the next column for $p = .05$ and see that 3.44 is also greater than 1.66. In fact, if we go all the way to the end of the row for df = 100, we can see that our t-statistic is greater than 3.174, which is the t-value needed to achieve $p = .001$ (meaning that there is a 0.1%, or 1 in 1000, chance that we would see this relationship randomly in our sample if there were no relationship between X and Y in the underlying population). This indicates that we have very confidently cleared the third hurdle in our assessment of whether or not there is a causal relationship between majority status and government duration.

7.4.3 Example 3: Correlation Coefficient

In our final example of bivariate hypothesis testing we look at a situation in which both the independent variable and the dependent variable are continuous. We test the hypothesis that there is a positive relationship between economic growth and incumbent-party fortunes in U.S. presidential elections.

In Chapter 5 we discussed the variation (or variance) of a single variable, and in Chapter 1 we introduced the concept of covariation. In the three examples that we have looked at so far, we have found there to be covariation between being from a union household and presidential vote, gender and presidential vote, and government type and government duration. All of these examples used at least one categorical variable. When we

[10] Although our degrees of freedom equal 175, we are using the row for df = 100 to get a rough idea of the p-value. With a computer program, we can calculate an exact p-value.

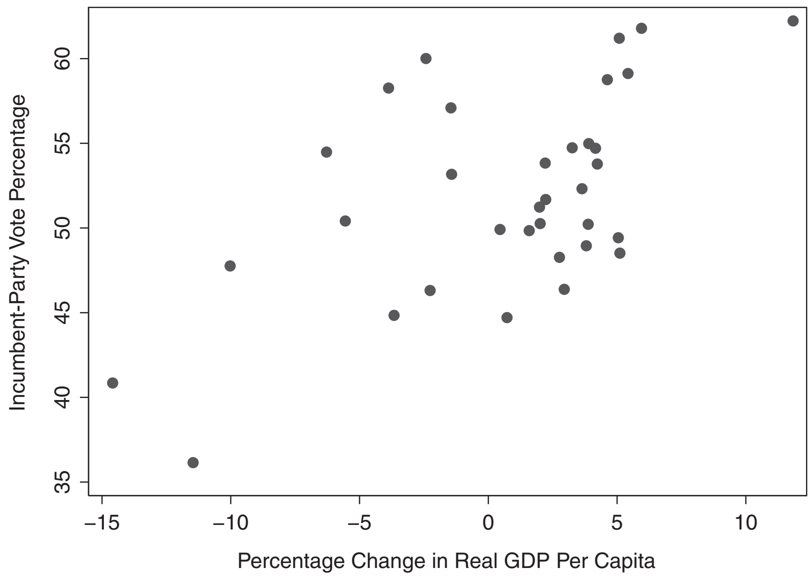

Figure 7.3. Scatter plot of change in GDP and incumbent-party vote share.

have an independent variable and a dependent variable that are both continuous, we can visually detect covariation pretty easily in graphs. Consider the graph in Figure 7.3, which shows a scatter plot of incumbent vote and economic growth. Scatter plots are useful for getting an initial look at the relationship between two continuous variables. Any time that you examine a scatter plot, you should figure out what are the axes and then what each point in the scatter plot represents. In these plots, the dependent variable (in this case incumbent vote) should be displayed on the vertical axis while the independent variable (in this case economic growth) should be displayed on the horizontal axis. Each point in the scatter plot should represent the values for the two variables for an individual case. So, in Figure 7.3, we are looking at the values of incumbent vote and economic growth for each U.S. presidential election year on which we have data for both variables.

When we look at this graph, we want to assess whether or not we see a pattern. Since our theory implies that the independent variable causes the dependent variable, we should move from left to right on the horizontal axis (representing increasing values of the independent variable) and see whether there is a corresponding increase or decrease in the values of the dependent variable. In the case of Figure 7.3, as we move from left to right, we generally see a pattern of increasing values on the vertical axis. This indicates that, as expected by our hypothesis, when the economy is doing better (more rightward values on the horizontal axis), we also tend to see higher vote percentages for the incumbent party in U.S. presidential elections (higher values on the vertical axis).

Covariance is a statistical way of summarizing the general pattern of association (or the lack thereof) between two continuous variables. The formula for covariance between two variables X and Y is

$$\text{cov}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}.$$

To better understand the intuition behind the covariance formula, it is helpful to think of individual cases in terms of their values relative to the mean of X ($\bar{X}$) and the mean of Y ($\bar{Y}$). If an individual case has a value for the independent variable that is greater than the mean of X ($X_i - \bar{X} > 0$) and its value for the dependent variable is greater than the mean of Y ($Y_i - \bar{Y} > 0$), that case's contribution to the numerator in the covariance equation will be positive. If an individual case has a value for the independent variable that is less than the mean of X ($X_i - \bar{X} < 0$) and a value of the dependent variable that is less than the mean of Y ($Y_i - \bar{Y} < 0$), that case's contribution to the numerator in the covariance equation will also be positive, because multiplying two negative numbers yields a positive product. If a case has a combination of one value greater than the mean and one value less than the mean, its contribution to the numerator in the covariance equation will be negative because multiplying a positive number by a negative number yields a negative product. Figure 7.4 illustrates this; we see the same plot of

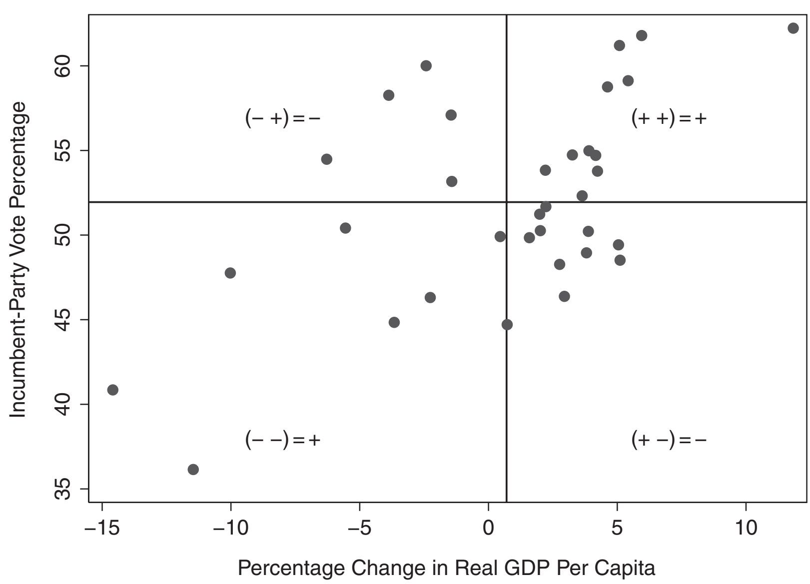

Figure 7.4. Scatter plot of change in GDP and incumbent-party vote share with mean-delimited quadrants.

growth versus incumbent vote, but with the addition of lines showing the mean value of each variable. In each of these mean-delimited quadrants we can see the contribution of the cases to the numerator. If a plot contains cases in mostly the upper-right and lower-left quadrants, the covariance will tend to be positive. On the other hand, if a plot contains cases in mostly the lower-right and upper-left quadrants, the covariance will tend to be negative. If a plot contains a balance of cases in all four quadrants, the covariance calculation will be close to zero because the positive and negative values will cancel out each other. When the covariance between two variables is positive, we describe this situation as a positive relationship between the variables, and when the covariation between two variables is negative, we describe this situation as a negative relationship.

Table 7.10 presents the calculations for each year in the covariance formula for the data that we presented in Figure 7.4. For each year, we have started out by calculating the difference between each X and $\bar{X}$ and the difference between each Y and $\bar{Y}$. If we begin with the year 1876, we can see that the value for growth (X_{1876}) was 5.11 and the value for vote (Y_{1876}) was 48.516. The value for growth is greater than the mean and the value for vote is less than the mean, $X_{1876} - \bar{X} = 5.11 - 0.7025294 = 4.407471$ and $Y_{1876} - \bar{Y} = 48.516 - 51.94718 = -3.431181$. In Figure 7.4, the dot for 1876 is in the lower-right quadrant. When we multiply these two mean deviations together, we get $(X_{1876} - \bar{X})(Y_{1876} - \bar{Y}) = -15.12283$.

We repeat this same calculation for every case (presidential election year). Each negative calculation like this contributes evidence that the overall relationship between X and Y is negative, whereas each positive calculation contributes evidence that the overall relationship between X and Y is positive. The sum across all 34 years of data in Table 7.10 is 616.59088, indicating that the positive values have outweighed the negative values. When we divide this by 34, we have the sample covariance, which equals 18.6846. This tells us that we have a positive relationship, but it does not tell us how confident we can be that this relationship is different from what we would see if our independent and dependent variables were not related in our underlying population of interest. To see this, we turn to a third test developed by Karl Pearson, Pearson's correlation coefficient. This is also known as **Pearson's *r***, the formula for which is

$$r = \frac{\text{cov}_{XY}}{\sqrt{\text{var}_X \text{var}_Y}}.$$

Table 7.11 is a covariance table. In a covariance table, the cells across the main diagonal (from upper-left to lower-right) are cells for which the column and the row reference the same variable. In this case the cell entry is the variance for the referenced variable. Each of the cells off of the main

Table 7.10. Contributions of individual election years to the covariance calculation

Year	Growth (X_i)	Vote (Y_i)	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1876	5.11	48.516	4.407471	–3.431181	–15.12283
1880	3.879	50.22	3.176471	–1.727179	–5.486332
1884	1.589	49.846	.8864706	–2.101179	–1.862634
1888	–5.553	50.414	–6.255529	–1.533179	9.590843
1892	2.763	48.268	2.060471	–3.679178	–7.580839
1896	–10.024	47.76	–10.72653	–4.187181	44.91393
1900	–1.425	53.171	–2.127529	1.223821	–2.603716
1904	–2.421	60.006	–3.123529	8.058821	–25.17196
1908	–6.281	54.483	–6.98353	2.535822	–17.70899
1912	4.164	54.708	3.461471	2.76082	9.556498
1916	2.229	51.682	1.526471	–.2651808	–.4047907
1920	–11.463	36.148	–12.16553	–15.79918	192.2054
1924	–3.872	58.263	–4.574529	6.315821	–28.89191
1928	4.623	58.756	3.920471	6.808821	26.69378
1932	–14.586	40.851	–15.28853	–11.09618	169.6442
1936	11.836	62.226	11.13347	10.27882	114.439
1940	3.901	54.983	3.198471	3.035822	9.709987
1944	4.233	53.778	3.53047	1.83082	6.463655
1948	3.638	52.319	2.935471	.3718202	1.091467
1952	.726	44.71	.0234706	–7.237181	–.169861
1956	–1.451	57.094	–2.153529	5.146822	–11.08383
1960	.455	49.913	–.2475294	–2.034182	.5035198
1964	5.087	61.203	4.38447	9.255819	40.58187
1968	5.049	49.425	4.34647	–2.522181	–10.96258
1972	5.949	61.791	5.24647	9.843821	51.64531
1976	3.806	48.951	3.103471	–2.99618	–9.298556
1980	–3.659	44.842	–4.361529	–7.105181	30.98945
1984	5.424	59.123	4.72147	7.175821	33.88043
1988	2.21	53.832	1.507471	1.884821	2.841312
1992	2.949	46.379	2.24647	–5.568178	–12.50875
1996	3.258	54.737	2.55547	2.789819	7.129301
2000	2.014	50.262	1.311471	–1.685179	–2.210063
2004	1.989	51.233	1.286471	–.7141783	–.9187693
2008	–2.26	46.311	–2.962529	–5.636179	16.69735
	$\bar{X} = 0.7025294$	$\bar{Y} = 51.94718$			$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 616.59088$

diagonal displays the covariance for a pair of variables. In covariance tables, the cells above the main diagonal are often left blank, because the values in these cells are a mirror image of the values in the corresponding cells below the main diagonal. For instance, in Table 7.11 the covariance between

Table 7.11 Covariance table for economic growth and incumbent-party presidential vote, 1880–2004

	Vote	Growth
Vote	35.4804	
Growth	18.6846	29.8997

growth and vote is the same as the covariance between vote and growth, so the upper-right cell in this table is left blank.

Using the entries in Table 7.11, we can calculate the correlation coefficient:

$$r = \frac{\text{cov}_{XY}}{\sqrt{\text{var}_X \, \text{var}_Y}},$$
$$r = \frac{18.6846}{\sqrt{35.4804 \times 29.8997}},$$
$$r = \frac{18.6846}{\sqrt{1060.853316}},$$
$$r = \frac{18.6846}{32.57074325},$$
$$r = 0.57366207.$$

There are a couple of points worth noting about the correlation coefficient. If all of the points in the plot line up perfectly on a straight, positively sloping line, the correlation coefficient will equal 1. If all of the points in the plot line up perfectly on a straight, negatively sloping line, the correlation coefficient will equal –1. Otherwise, the values will lie between positive one and negative one. This standardization of correlation coefficient values is a particularly useful improvement over the covariance calculation. Additionally, we can calculate a t-statistic for a correlation coefficient as

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

with $n-2$ degrees of freedom, where n is the number of cases. In this case, our degrees of freedom equal $34-2=32$.

For the current example,

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$
$$t_r = \frac{0.57366207\sqrt{34-2}}{\sqrt{1-(0.57366207)^2}},$$
$$t_r = \frac{0.57366207 \times 5.656854249}{\sqrt{1-(0.329088171)}},$$

$$t_r = \frac{3.245122719}{\sqrt{0.670911829}},$$
$$t_r = \frac{3.245122719}{0.819092076},$$
$$t_r = 3.961853391.$$

With the degrees of freedom equal to 34 ($n = 34$) minus two, or 32, we can now turn to the t-table in Appendix B. Looking across the row for df $= 30$, we can see that our calculated t of 3.96 is greater even than the critical t at the p-value of .001 (which is 3.385). This tells us that the probability of seeing this relationship due to random chance is less than .001 or 1 in 1000. When we estimate our correlation coefficient with a computer program, we get a more precise p-value of .0004. Thus we can be quite confident that there is covariation between economic growth and incumbent-party vote share and that our theory has successfully cleared our third causal hurdle.[11]

7.5 WRAPPING UP

We have introduced three methods to conduct bivariate hypothesis tests – tabular analysis, difference of means tests, and correlation coefficients. Which test is most appropriate in any given situation depends on the measurement metric of your independent and dependent variables. Table 7.1 should serve as a helpful reference for you on this front.

We have yet to introduce the final method for conducting bivariate hypothesis tests covered in this book, namely bivariate regression analysis. That is the topic of our next chapter, and it serves as the initial building block for multiple regression (which we will cover in Chapter 9).

CONCEPTS INTRODUCED IN THIS CHAPTER

- chi-squared (χ^2) test for tabular association – a statistical test for a relationship between two categorical variables.
- correlation coefficient – a measure of linear association between two continuous variables.
- covariance – an unstandardized statistical measure summarizing the general pattern of association (or the lack thereof) between two continuous variables.

[11] The first causal hurdle is pretty well cleared if we refer back to the discussion of the theory of economic voting in earlier chapters. The second causal hurdle also can be pretty well cleared logically by the timing of the measurement of each variable. Because economic growth is measured prior to incumbent vote, it is difficult to imagine that Y caused X.

- critical value – a predetermined standard for a statistical test such that if the calculated value is greater than the critical value, then we conclude that there is a relationship between the two variables; and if the calculated value is less than the critical value, we cannot make such a conclusion.
- degrees of freedom – the number of pieces of information we have beyond the minimum that we would need to make a particular inference.
- difference of means test – a method of bivariate hypothesis testing that is appropriate for a categorical independent variable and a continuous dependent variable.
- Pearson's r – the most commonly employed correlation coefficient.
- p-value – the probability that we would see the relationship that we are finding because of random chance.
- statistically significant relationship – a conclusion, based on the observed data, that the relationship between two variables is not due to random chance, and therefore exists in the broader population.
- tabular analysis – a type of bivariate analysis that is appropriate for two categorical variables.

EXERCISES

1. What form of bivariate hypothesis test would be appropriate for the following research questions:
   (a) You want to test the theory that being female causes lower salaries.
   (b) You want to test the theory that a state's percentage of college graduates is positively related to its turnout percentage.
   (c) You want to test the theory that individuals with higher incomes are more likely to vote.
2. Explain why each of the following statements is either true or false:
   (a) The computer program gave me a p-value of .000, so I know that my theory has been verified.
   (b) The computer program gave me a p-value of .02, so I know that I have found a very strong relationship.
   (c) The computer program gave me a p-value of .07, so I know that this relationship is due to random chance.
   (d) The computer program gave me a p-value of .50, so I know that there is only a 50% chance of this relationship being systematic.
3. Take a look at Figure 7.5. What is the dependent variable? What are the independent variables? What does this table tell us about politics?
4. What makes the table in Figure 7.5 so confusing?

MORAL VALUES – THE TRANSATLANTIC GULF

Q How often do you go to church?

	BRITAIN				US	
	All Voters	Labour voters	Tory voters	Lib Dem voters	Bush voters	Kerry voters
More than weekly	2%	2%	3%	1%	63%	35%
Weekly	10%	10%	13%	7%	58%	41%
Monthly	5%	6%	4%	6%	50%	50%
A few times a year	36%	36%	38%	40%	44%	55%
Never	47%	46%	43%	44%	34%	64%
Q Which of the following is closest to your view of what the law should say about abortion?						
Always legal: absolute right to choose	38%	45%	34%	46%	24%	74%
Mostly legal: some restrictions	36%	35%	40%	32%	37%	62%
Mostly illegal: only in exceptional circumstances	19%	14%	18%	17%	72%	27%
Always illegal	4%	4%	3%	3%	77%	22%
Q Which of the following is closest to your view of what the law should be towards same-sex couples?						
Legal right to marry	28%	33%	18%	31%	22%	77%
Legally civil union but not marriage	37%	37%	39%	47%	51%	48%
No legal recognition of same sex couples	29%	23%	39%	20%	69%	30%

Sources: for British figures, Populus poll for *The Times* (Nov 5–7); for American figures, exit polls conducted by National Election Poll (Nov 2)

Figure 7.5. What is wrong with this table?

5. Conduct a tabular analysis from the information presented in the following hypothetical discussion of polling results: "We did a survey of 800 respondents who were likely Democratic primary voters in the state. Among these respondents, 45% favored Obama whereas 55% favored Clinton. When we split the respondents in half at the median age of 40, we found some stark differences: Among the younger half of the sample respondents, we found that 72.2% favored Obama to be the nominee and among the older sample respondents, we found that 68.2% favored Clinton."

6. For the example in Exercise 5, test the theory that age is related to preference for a Democratic nominee.

7. A lot of people in the United States think that the Watergate scandal in 1972 caused a sea change in terms of U.S. citizens' views toward incumbent politicians. Use the data in Table 7.12 to produce a difference of means test of the null hypothesis that average reelection rates were the same before and after the Watergate scandal. Because of the timing of the elections and the scandal, 1972 should be coded as a pre-scandal case. Do this test once each for the House and the Senate. Show all of your work.

Table 7.12. Incumbent reelection rates in U.S. congressional elections, 1964–2006

Year	House	Senate
1964	87	85
1966	88	88
1968	97	71
1970	85	77
1972	94	74
1974	88	85
1976	96	64
1978	94	60
1980	91	55
1982	90	93
1984	95	90
1986	98	75
1988	98	85
1990	96	96
1992	88	83
1994	90	92
1996	94	91
1998	98	90
2000	98	79
2002	96	86
2004	98	96
2006	94	79

8. Using the data set "BES2005 Subset," produce a table that shows the combination values for the variables "LabourVote" (Y) and "IraqWarApprovalDich" (X). Read the descriptions of these two variables and write about what this table tells you about politics in the United Kingdom in 2005. Compute a χ^2 hypothesis test for these two variables. Write about what this tells you about politics in the United Kingdom in 2005.
9. Using the data set "BES2005 Subset," test the hypothesis that values for "BlairFeelings" (Y) are different across different values of "IraqWarApprovalDich" (X). Read the descriptions of these two variables and write about what this table tells you about politics in the United Kingdom in 2005.
10. Using the data set "BES2005 Subset," produce a scatter plot of the values for "BlairFeelings" (Y) and "SelfLR" (X). Calculate a correlation coefficient and p-value for the hypothesis that these two variables are related to each other. Read the descriptions of these two variables and write about what this table tells you about politics in the United Kingdom in 2005.