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.

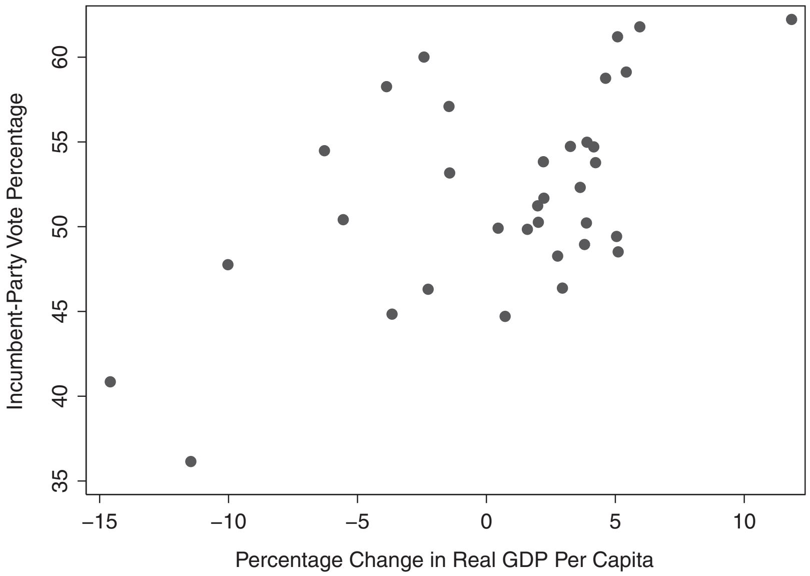

Figure 7.3. Scatter plot of change in GDP and incumbent-party vote share.

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

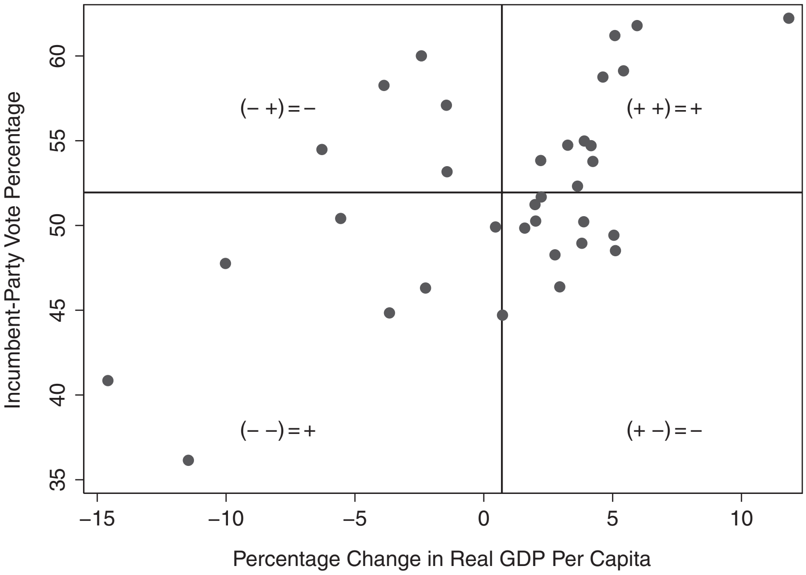

Figure 7.4. Scatter plot of change in GDP and incumbent-party vote share with mean-delimited quadrants.

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?