In this chapter, we examine a very flexible and widely used test of significance called **the analysis of variance** (often abbreviated as **ANOVA**). This test is designed to be used with interval-ratio-level dependent variables and is a powerful tool for analyzing the most sophisticated and precise measurements you are likely to encounter.

It is perhaps easiest to think of ANOVA as an extension of the test for the significance of the difference between two sample means, which was presented in Chapter 8. Those tests can be used only in situations in which our independent variable has exactly two categories (e.g., Protestants and Catholics). On the other hand, the analysis of variance is appropriate for independent variables with more than two categories (e.g., Protestants, Catholics, Jews, people with no religious affiliation, and so forth).

To illustrate, suppose we were interested in analyzing support for capital punishment. Why does opinion on this issue vary from person to person? Could there be a relationship between religion (the independent variable) and support for capital punishment (the dependent variable)? The death penalty has an obvious moral dimension and may well be affected by a person's religious background.

Suppose we administered a scale that measures support for capital punishment at the interval-ratio level to a randomly selected sample that includes Protestants, Catholics, Jews, people with no religious affiliation ("Nones"), and people from other religions ("Others"). We have five categories of subjects, and we want to see if opinion varies significantly by the category (religion) into which a person is classified. We will also want to raise other issues: Which religion shows the most support for capital punishment? Are Protestants significantly more supportive than Catholics or Jews? How do people with no religious affiliation compare to other people? The analysis of variance provides a very useful statistical context in which the questions can be addressed.

The Logic of the Analysis of Variance

For ANOVA, the null hypothesis is that the populations from which the samples are drawn have the same score on the dependent variable. As applied to our problem, the null hypothesis could be phrased as "People from different religious affiliations do not vary in their support for the death penalty" or, symbolically, as $\mu_1 = \mu_2 = \mu_3 = \ldots \mu_k$. (Note that this is an extended version of the null hypothesis for the two-sample *t* test.) As usual, the researcher will normally want to reject the null hypothesis and, in this case, show that support is related to religion.

If the null hypothesis of "no difference" in the populations is true, then any means calculated from randomly selected samples should be roughly equal in value. The average score for the Protestant sample should be about the same as the average score for the Catholics, the Jews, and so forth. Note that the averages are unlikely to be exactly the same value even if the null hypothesis really is true because we will always encounter some error or chance fluctuations in the measurement process. We are *not* asking: "Are there differences between the samples (or, in our example, the religions)?" Rather, we are asking: "Are the differences between the samples large enough to reject the null hypothesis and justify the conclusion that the populations are different?"

TABLE 9.1 Support for Capital Punishment by Religion (Fictitious Data)

	Protestant	Catholic	Jew	None	Other
Mean =	10.3	11.0	10.1	9.9	10.5
Standard deviation =	2.4	1.9	2.2	1.7	2.0

TABLE 9.2 Support for Capital Punishment by Religion (Fictitious Data)

	Protestant	Catholic	Jew	None	Other
Mean =	14.7	11.3	5.7	8.3	7.1
Standard deviation =	0.9	0.8	1.0	1.1	0.7

Now consider what kinds of outcomes we might encounter if we actually administered a "Support for Capital Punishment Scale" and organized the scores by religion. Of the infinite variety of possibilities, let us focus on the two extreme outcomes presented in Tables 9.1 and 9.2. In the first set of hypothetical results (Table 9.1), we see that the means and standard deviations of the groups are quite similar. The average scores are about the same, and all five groups exhibit about the same dispersion. These results are consistent with the null hypothesis of no difference between the populations on support for capital punishment. Neither the average score nor the dispersion of the scores varies in any important way by religion.

Now consider another set of fictitious results, as displayed in Table 9.2. Here, we see substantial differences in average score, with Jews showing the lowest support and Protestants showing the highest. Also, the standard deviations are low and similar from category to category, indicating that there is not much variation within the religions. Table 9.2 shows marked differences *between* religions combined with homogeneity (or low standard deviations) *within* religions. In other words, the religions are different from each other and there is not much variation within each religion. These results would contradict the null hypothesis and support the notion that support for the death penalty does vary by religion

ANOVA proceeds by making these kinds of comparisons. The test compares the amount of variation between categories (for example, from Protestants to Catholics to Jews to "Nones" to "Others") with the amount of variation within categories (among Protestants, among Catholics, and so forth). The greater the difference *between* categories (as measured by the means) relative to the differences *within* categories (as measured by the standard deviations), the more likely that the null hypothesis of "no difference" is false and can be rejected. If support for capital punishment truly varies by religion, then the sample mean for each religion should be quite different from the others and dispersion within the categories should be relatively low.

The Computation of Anova

Even though we have been thinking of ANOVA as a test for the significance of the difference between sample means, the computational routine actually involves developing two separate estimates of the population variance σ^2

(hence, the name *analysis of variance*). Recall from Chapter 4 that the variance is the standard deviation squared. One estimate of the population variance is based on the amount of variation *within* each of the categories of the independent variable and the other is based on the amount of variation *between* categories.

Before constructing these estimates, we need to introduce some new concepts and statistics. The first new concept is the total variation of the scores, which is measured by a quantity called the **total sum of squares**, or ***SST***:

FORMULA 9.1

$$SST = \sum X_i^2 - N\overline{X}^2$$

To solve this formula, first find the sum of the squared scores (in other words, square each score and then add up the squared scores). Next, square the mean of all scores, multiply that value by the total number of cases in the sample (N), and the subtract that quantity from the sum of the squared scores.

Formula 9.1 may seem vaguely familiar. A similar expression—$\Sigma(X_i - \overline{X})^2$—appears in the formula for the standard deviation and variance (see Chapter 4). All three statistics incorporate information about the variation of the scores (or, in the case of *SST*, the squared scores) around the mean (or, in the case of *SST*, the square of the mean multiplied by *N*). In other words, all three statistics are measures of the variation, or dispersion, of the scores.

To construct the two separate estimates of the population variance, the total variation (*SST*) is divided into two components. One component reflects the pattern of variation *within* each of the categories and is called the **sum of squares within (*SSW*)**. In our example problem, *SSW* would measure the amount of variety in support for the death penalty within each of the religions.

The other component is based on the variation *between* categories and is called the **sum of squares between (*SSB*)**. Again using our example to illustrate, *SSB* measures how different people in each religion are from each other in their support for capital punishment. *SSW* and *SSB* are components of *SST*, as reflected in Formula 9.2:

FORMULA 9.2

$$SST = SSB + SSW$$

Let us with the computation of *SSB*—our measure of the variation between categories. We use the category means as summary statistics to determine the size of the difference from category to category. In other words, we compare the average support for the death penalty for each religion with the average support for all religions combined to determine *SSB*. The formula for the sum of squares between (*SSB*) is:

FORMULA 9.3

$$SSB = \sum N_k(\overline{X}_k - \overline{X})^2$$

where:

SSB = the sum of squares between the categories
N_k = the number of cases in a category
$\overline{X}_k$ = the mean of a category

To find *SSB*, subtract the overall mean of all scores $(\overline{X})$ from each category mean $(\overline{X}_k)$, square the difference, multiply by the number of cases in the category, and then add the results across all the categories.

The second estimate of the population variance (*SSW*) is based on the amount of variation within the categories. Formula 9.2 shows that the total sum of squares (*SST*) is equal to the addition of *SSW* and *SSB*. This relationship provides us with an easy method for finding *SSW* by simple subtraction. Formula 9.4 rearranges the symbols in Formula 9.2.

FORMULA 9.4

$$SSW = SST - SSB$$

Let us pause for a second to remember what we are after here. If the null hypothesis is *true*, then there should not be much variation from category to category (see Table 9.1) and *SSW* and *SSB* should be roughly equal. If the null hypothesis is *not true*, there will be large differences between categories (see Table 9.2) relative to the differences within categories and *SSB* should be much larger than *SSW*. *SSB* will increase as the differences *between* category means increases, especially when there is not much variation *within* the categories (*SSW*). The larger *SSB* is compared to *SSW*, the more likely we are to reject the null hypothesis.

The next step in the computational routine is to construct the estimates of the population variance. To do this, we divide each sum of squares by its respective degrees of freedom. To find the degrees of freedom associated with *SSW*, subtract the number of categories (k) from the number of cases (N). The degrees of freedom associated with *SSB* are the number of categories minus 1. In summary:

FORMULA 9.5

$$dfw = N - k$$

where:

dfw = degrees of freedom associated with *SSW*
N = total number of cases
k = number of categories

FORMULA 9.6

$$dfb = k - 1$$

where:

dfb = degrees of freedom associated with *SSB*
k = number of categories

The actual estimates of the population variance—called the **mean square estimates**—are calculated by dividing each sum of squares by its respective degrees of freedom:

FORMULA 9.7

$$\text{Mean square within} = \frac{SSW}{dfw}$$

FORMULA 9.8

$$\text{Mean square between} = \frac{SSB}{dfb}$$

The test statistic calculated in step 4 of the five-step model is called the *F ratio,* and its value is determined by this formula:

FORMULA 9.9

$$F = \frac{\text{Mean square between}}{\text{Mean square within}}$$

ONE STEP AT A TIME **Computing ANOVA**

To compute ANOVA, we will use Formulas 9.1, 9.3, and 9.4 to find *SST*, *SSB*, and *SSW*. Then, we will calculate the degrees of freedom, mean square estimates of the population variance, and the obtained *F* ratio. I strongly recommend you use a computing table like Table 9.3 to organize the computations.

Step ***Operation***

To Find SST by Using Formula 9.1

1. Find ΣX_i^2 by squaring each score and adding all the squared scores together.
2. Find $N\overline{X}_i^2$ by squaring the value of the mean and then multiplying by *N*.
3. Subtract the quantity you found in step 2 from the quantity you found in step 1.

To Find SSB by Using Formula 9.3

1. Subtract the mean of all scores $(\overline{X})$ from the mean of each category $(\overline{X}_k)$ and then square each difference.
2. Multiply each of the squared differences you found in step 1 by the number of cases in the category N_k.
3. Add the quantities you found in step 2.

To Find SSW by Using Formula 9.4

1. Subtract the value of *SSB* from the value of *SST*.

To Calculate Degrees of Freedom

1. For *dfw*, subtract the number of categories (*k*) from the number of cases (*N*).
2. For *dfb*, subtract 1 from the number of categories (*k*).

To Construct the Two Mean Square Estimates of the Population Variance

1. To find the *MSW* estimate, divide *SSW* by *dfw*.
2. To find the *MSB* estimate, divide *SSB* by *dfb*.

To Find the Obtained F Ratio

1. Divide the *MSB* estimate by the *MSW* estimate.

As you can see, the value of the *F* ratio will be a function of the amount of variation between categories (based on *SSB*) to the amount of variation within the categories (based on *SSW*). The greater the variation between the categories relative to the variation within, the higher the value of the *F* ratio and the more likely we will reject the null hypothesis.

An Example of Computing the Analysis of Variance Assume we have administered our Support for Capital Punishment Scale to a sample of 20 individuals who are equally divided into the five religions. (Obviously, this sample is much too small for any serious research and is intended solely for purposes of illustration.) All scores are reported in Table 9.3, along with the other quantities needed to complete the computations. The scores (X_i) are listed for each of the five religions, and a column has been added for the squared scores (X_i^2). The sums of X_i and X_i^2 are reported at the bottom of each column. The category means $(\overline{X}_k)$ show that the four Protestants averaged 12.5 on the Support for Capital Punishment Scale, the four Catholics averaged 21.0, and so forth. Finally, the overall mean (sometimes called

the *grand mean*) is reported in the bottom row of the table. This shows that all 20 respondents averaged 16.6 on the scale.

To organize our computations, we will follow the routine summarized in the "One Step at a Time" box. We begin by finding *SST* by means of Formula 9.1:

$$SST = \sum X^2 - N\overline{X}^2$$
$$SST = (666 + 1898 + 1{,}078 + 1794 + 712) - (20)(16.6)^2$$
$$SST = (6{,}148) - (20)(275.56)$$
$$SST = 6{,}148 - 5{,}511.2$$
$$SST = 636.80$$

The sum of squares between (*SSB*) is found by means of Formula 9.3:

$$SSB = \sum N_k(\overline{X}_k - \overline{X})^2$$
$$SSB = 4(12.5 - 16.6)^2 + 4(21.0 - 16.6)^2 + 4(16.0 - 16.6)^2 + 4(20.5 - 16.6)^2 + 4(13.0 - 16.6)^2$$
$$SSB = 67.24 + 77.44 + 1.44 + 60.84 + 51.84$$
$$SSB = 258.80$$

Now *SSW* can be found by subtraction (Formula 9.4):

$$SSW = SST - SSB$$
$$SSW = 636.8 - 258.80$$
$$SSW = 378.00$$

To find the degrees of freedom for the two sums of squares, we use Formulas 9.5 and 9.6:

$$dfb = N - k = 20 - 5 = 15$$
$$dfb = k - 1 = 5 - 1 = 4$$

Finally, we are ready to construct the mean square estimates of the population variance. For the estimate based on *SSW*, we use Formula 9.7:

$$\text{Mean square within} = \frac{SSW}{dfw} = \frac{378.00}{15} = 25.20$$

TABLE 9.3 Support for Capital Punishment by Religion (Fictitious Data)

Protestant		Catholic		Jewish		None		Other	
X_i	X_i^2	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2
8	64	12	144	12	144	15	225	10	100
12	144	20	400	13	169	16	256	18	324
13	169	25	625	18	324	23	529	12	144
17	289	27	729	21	441	28	784	12	144
50	666	84	1,898	64	1,078	82	1,794	52	712
$\overline{X}_k = 12.5$		$\overline{X}_k = 21.0$		$\overline{X}_k = 16.0$		$\overline{X}_k = 20.5$		$\overline{X}_k = 13.0$	

$\overline{X} = 16.6$

For the between estimate, we use Formula 9.8:

$$\text{Mean square between} = \frac{SSB}{dfw} = \frac{258.80}{4} = 64.70$$

The test statistic, or F ratio, is found by means of Formula 9.9:

$$F = \frac{\text{Mean square between}}{\text{Mean square within}} = \frac{64.70}{25.20} = 2.57$$

This statistic must still be evaluated for its significance. *(Solve any of the end-of-chapter problems to practice computing these quantities and solving these formulas.)*

A Test of Significance for ANOVA

Now we will see how to test an F ratio for significance. We will also take a look at some of the assumptions underlying the ANOVA test. As usual, we will follow the five-step model as a convenient way of organizing the decision-making process.

Step 1. Making assumptions and meeting test requirements.

Model: Independent random samples
Level of measurement is interval-ratio
Populations are normally distributed
Population variances are equal

The model assumptions are quite strict and underscore the fact that ANOVA should be used only with dependent variables that have been carefully and precisely measured. However, as long as the categories are roughly equal in size, ANOVA can tolerate some violation of the model assumptions. In situations where you are uncertain or have samples of very different size, it is probably advisable to use an alternative test. (Chi square in Chapter 10 is one option.)

Step 2. Stating the null hypothesis. For ANOVA, the null hypothesis always states that the means of the populations from which the samples were drawn are equal. For our example problem, we are concerned with five different populations, or categories, so our null hypothesis would be:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

where μ_1 represents the mean for Protestants, μ_2 the mean for Catholics, and so forth.

The alternative hypothesis states simply that at least one of the population means is different. The wording here is important. If we reject the null hypothesis, ANOVA does not identify which mean or means are significantly different.

(H_1: At least one of the population means is different.)

Step 3. Selecting the sampling distribution and establishing the critical region. The sampling distribution for ANOVA is the F distribution, which is summarized in Appendix D. Note that there are separate tables for alphas of .05 and .01.

As with the *t* table, the value of the critical *F* score will vary by degrees of freedom. For ANOVA, there are two separate degrees of freedom—one for each estimate of the population variance. The numbers across the top of the table are the degrees of freedom associated with the between estimate (*dfb*), and the numbers down the side of the table are those associated with the within estimate (*dfw*). In our example, *dfb* is $(k - 1)$, or 4, and *dfw* is $(N - k)$, or 15 (see Formulas 9.5 and 9.6). Thus, if we set alpha at .05, our critical *F* score will be 3.06.

Summarizing these considerations:

$$\begin{aligned} \text{Sampling distribution} &= F \text{ distribution} \\ \text{Alpha} &= 0.05 \\ \text{Degrees of freedom within } (dfw) &= (N - k) = 15 \\ \text{Degrees of freedom between } (dfb) &= (k - 1) = 4 \\ F(\text{critical}) &= 3.06 \end{aligned}$$

Take a moment to inspect the two *F* tables and you will notice that all the values are greater than 1.00. This is because ANOVA is a one-tailed test and we are concerned only with outcomes in which there is more variance between categories than within categories. *F* values of less than 1.00 would indicate that the between estimate was lower in value than the within estimate, and because we would always fail to reject the null in such cases, we simply ignore this class of outcomes.

Step 4. Computing the test statistic. This was done in the previous section, where we found an obtained *F* ratio of 2.57.

Step 5. Making a decision and interpreting the results of the test. Compare the test statistic with the critical value:

$$\begin{aligned} F(\text{critical}) &= 3.06 \\ F(\text{obtained}) &= 2.57 \end{aligned}$$

Because the test statistic does not fall into the critical region, our decision would be to fail to reject the null. Support for capital punishment does not differ significantly by religion, and the variation we observed in the sample means is unimportant.

Applying Statistics 9.1 The Analysis of Variance

An experiment in teaching introductory biology was recently conducted at a large university. One section was taught by the traditional lecture-lab method, a second was taught by an all-lab/demonstration approach with no lectures, and a third was taught entirely by a series of recorded lectures and demonstrations that the students were free to view at any time and as often as they wanted. Students were randomly assigned to each of the three sections, and at the end of the semester, random samples of final exam scores were collected from each section. Is there a significant difference in student performance by teaching method?

Applying Statistics 9.1 *(continued)*

Final Exam Scores by Teaching Method

Lecture		Demonstration		Recording	
X	X^2	X	X^2	X	X^2
55	3,025	56	3,136	50	2,500
57	3,249	60	3,600	52	2,704
60	3,600	62	3,844	60	3,600
63	3,969	67	4,489	61	3,721
72	5,184	70	4,900	63	3,969
73	5,329	71	5,041	69	4,761
79	6,241	82	6,724	71	5,041
85	7,225	88	7,744	80	6,400
92	8,464	95	9,025	82	6,724
$\Sigma X = 636$		651		588	
$\Sigma X^2 =$	46,286		48,503		39,420
$\bar{X}_k = 70.67$		72.33		65.33	

$$\bar{X} = 1{,}875/27 = 69.44$$

We can see by inspection that the "Recording" group had the lowest average score and that the "Demonstration" group had the highest average score. The ANOVA test will tell us if these differences are large enough to justify the conclusion that they did not occur by chance alone. We can organize the computations following the steps described in the "One Step at a Time" box about computing ANOVA (see page 229):

$$SST = \Sigma X^2 - N\bar{X}^2$$
$$SST = (46{,}286 + 48{,}503 + 39{,}420) - 27(69.44)^2$$
$$SST = 134{,}209 - 130{,}191.67$$
$$SST = 4017.33$$

$$SSB = \Sigma N_k(\bar{X}_k - \bar{X})^2$$
$$SSB = (9)(70.67 - 69.44)^2 + (9)(72.33 - 69.44)^2 + (9)(65.33 - 69.44)^2$$
$$SSB = 13.62 + 75.17 + 152.03$$
$$SSB = 240.82$$

$$SSW = SST - SSB$$
$$SSW = 4017.33 - 240.82$$
$$SSW = 3776.51$$

$$dfw = N - k = 27 - 3 = 24$$
$$dfb = k - 1 = 3 - 1 = 2$$

$$\text{Mean square within} = \frac{SSW}{dfw} = \frac{3776.51}{24} = 157.36$$

$$\text{Mean square between} = \frac{SSB}{dfb} = \frac{240.82}{2} = 120.41$$

$$F = \frac{\text{Mean square between}}{\text{Mean square within}}$$
$$F = \frac{120.41}{157.36}$$
$$F = 0.77$$

We can now conduct the test of significance.

Step 1. Making Assumptions and Meeting Test Requirements.

Model: Independent random sample
Level of measurement is interval-ratio
Populations are normally distributed
Population variances are equal

Step 2. Stating the Null Hypothesis.

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: At least one of the population means is different,

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

$$\text{Sampling distribution} = F \text{ distribution}$$
$$\text{Alpha} = 0.05$$
$$\text{Degrees of freedom (within)} = (N - k)$$
$$= (27 - 3) = 24$$
$$\text{Degrees of freedom (between)} = (\text{k} - 1)$$
$$= (3 - 1) = 2$$
$$F(\text{critical}) = 3.40$$

Step 4. Computing the Test Statistics. We found an obtained F ratio of 0.77.

Step 5. Making a Decision and Interpreting the Results of the Test. Compare the test statistic with the critical value:

$$F(\text{critical}) = 3.40$$
$$F(\text{obtained}) = 0.77$$

We would clearly fail to reject the null hypothesis and conclude that the observed differences among the category means were the results of random chance. Student performance in this course does not vary significantly by teaching method.