

## Chapter 12

# Comparing Groups: Analysis of Variance (ANOVA) Methods

Chapter 7 presented methods for comparing the means of two groups. We next see how those methods extend for comparing means of *several* groups.

Chapter 8 presented methods for analyzing association between two categorical variables. Chapters 9 and 11 presented regression methods for analyzing association between quantitative variables. Methods for comparing means for several groups relate to the association between a quantitative response variable and categorical explanatory variable. The mean of the quantitative response variable is compared among groups that are categories of the explanatory variable. For example, for a comparison of mean annual income among blacks, whites, and Hispanics, the quantitative response variable is annual income and the categorical explanatory variable is racial-ethnic status.

The inferential method for comparing several means is called the *analysis of variance*, abbreviated by **ANOVA**. Section 12.1 shows that the name refers to the way the significance test focuses on two types of variability in the data. Section 12.2 presents confidence intervals comparing group means. Section 12.3 shows that the inferences are special cases of a multiple regression analysis. Sections 12.4 and 12.5 extend the methods to incorporate additional explanatory variables, for example to compare mean income both across categories of racial-ethnic status and of gender.

Sections 12.1–12.5 present analyses for *independent samples*. As Section 7.1 explained, when each sample has the same subjects rather than unmatched samples, the samples are *dependent* and different methods apply. Sections 12.6 and 12.7 present such methods.

## 12.1 Comparing Several Means: The Analysis of Variance

The great British statistician R. A. Fisher developed the analysis of variance method in the 1920s. The heart of this analysis is a significance test, using his  $F$  distribution, for detecting differences among a set of population means.

### Assumptions for the $F$ Test Comparing Means

Let  $g$  denote the number of groups to compare. The means of the response variable for the corresponding populations are  $\mu_1, \mu_2, \dots, \mu_g$ . The sample means are  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_g$ .

The analysis of variance (ANOVA) is an  $F$  test of

$$H_0: \mu_1 = \mu_2 = \dots = \mu_g \text{ against} \\ H_a: \text{at least two of the population means are unequal.}$$

If  $H_0$  is false, perhaps all the population means differ, perhaps some differ, or perhaps merely one mean differs from the others. The test analyzes whether the differences observed among the sample means could have reasonably occurred by chance, if  $H_0$  were true.

The assumptions for the test are as follows:

### 12.1. COMPARING SEVERAL MEANS: THE ANALYSIS OF VARIANCE 497

- For each group, the population distribution of the response variable  $Y$  is normal.
- The standard deviation of the population distribution is the same for each group. Denote the common value by  $\sigma$ .
- The samples from the populations are *independent* random samples.

Figure 12.1 portrays the first two assumptions. Under these assumptions, the null hypothesis states that the population distribution does not depend on the group to which a subject belongs. The ANOVA test is a *test of independence* between the quantitative response variable and the categorical explanatory variable.

Figure 12.1: Assumptions About Population Distributions: Normal with Equal Standard Deviations,  $\sigma$

((Fig. 12.1 in 3e))

The assumptions about the population distributions are stringent ones that are never satisfied exactly in practice. As usual, the random sampling assumption is the most important one. The last section of this chapter discusses the effects of violating assumptions.

#### **Example 12.1 Political Ideology by Political Party Identification**

Table 12.1 summarizes observations on political ideology for three groups, based on data from subjects of age 18-30 in the 2004 General Social Survey. The three groups are the (Democrat, Independent, Republican) categories of the explanatory variable, political party identification (ID). Political ideology, the response variable, is measured on a seven-point scale, ranging from extremely liberal (1) to extremely conservative (7). For each party ID, Table 12.1 shows the number of subjects who made each response. For instance, of 91 Democrats, 9 responded extremely liberal, 20 responded liberal, . . . , 0 responded extremely conservative.

Since Table 12.1 displays the data as counts in a contingency table, we could use methods for categorical data (Chapter 8). The chi-squared test treats both variables as nominal, however, whereas political ideology is ordinal. That test is not directed toward detecting whether responses have a higher or lower mean in some groups than others. Likewise, the ordinal methods of Chapter 8 (such as the gamma measure of association) are inappropriate, because they require both variables to be ordinal. Here, the groups, which are the categories of political party ID, are nominal.

When an ordinal response has several categories, in practice it is common to assign scores to its levels and treat it as a quantitative variable. This is a reasonable strategy when we would like to focus on a measure of center such as the mean rather than on the proportions in particular categories. For Table 12.1, for instance, interest might

Table 12.1: Political Ideology by Political Party Identification (ID), for those of Age 18-30

Group (Party ID)	Political Ideology							Sample Size	Mean	Standard Deviation
	1	2	3	4	5	6	7			
Democrat	9	20	17	36	4	5	0	91	3.23	1.28
Independent	7	11	17	48	12	11	5	111	3.90	1.43
Republican	0	2	7	23	23	17	2	74	4.70	1.10

Note: 1 = extremely liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = extremely conservative.

focus on how liberal or conservative the responses tend to be for each group, in some average sense, rather than on the proportions falling in each category. We analyze these data by assigning the scores (1, 2, 3, 4, 5, 6, 7) to the levels of political ideology and then comparing means. The higher the mean score, the more conservative the group's responses tended to be.

For these scores, Table 12.1 also shows the mean and standard deviation for each group. The overall sample mean is  $\bar{y} = 3.89$ , quite near the score of 4.0 corresponding to moderate ideology. We shall test whether the three populations have equal means. The null hypothesis is  $H_0: \mu_1 = \mu_2 = \mu_3$ , where  $\mu_1$  is the population mean ideology for Democrats,  $\mu_2$  for Independent, and  $\mu_3$  for Republicans.

□

## Variability Between and Within Groups

Why is a method for comparing population *means* called an analysis of *variance*? The reason is that the test statistic compares two estimates of variance. One uses the variability *between* each sample mean  $\bar{y}_i$  and the overall mean  $\bar{y}$ . The other uses the variability *within* each group of the sample observations about their separate means—the observations from the first group about  $\bar{y}_1$ , the observations from the second group about  $\bar{y}_2$ , and so forth.

To illustrate, suppose randomly sampled observations from three groups are as shown in Figure 12.2a. It seems clear that the means of the populations these samples represent are unequal. The basis for this conclusion is that the variability *between* sample means is large and the variability of the observations *within* each sample is small.

By contrast, look at Figure 12.2b. It has the same sample means as in Figure 12.2a, so the variability *between* sample means is the same. But, in Figure 12.2b the variability *within* the groups is much larger than in Figure 12.2a. The sample standard deviation for each group is much larger in Figure 12.2b. Now it is not clear whether the population means differ. Generally, the greater the variability between sample

means and the smaller the variability within each group, the stronger the evidence against the null hypothesis of equal population means.

Figure 12.2: Two Samples: The means are the same in each case, so variability *between* groups is the same, but variability *within* groups is larger in the second set.

((Fig. 12.2 in 3e))

### The $F$ Test Statistic is a Ratio of Two Variance Estimates

For testing equal population means ( $H_0: \mu_1 = \mu_2 = \dots = \mu_g$ ), the test statistic is the ratio of two estimates of the population variance,  $\sigma^2$ , for each group. One estimate, which uses the variability *within* each sample, is called the ***within-groups estimate***. The second estimate, which uses the variability *between* each sample mean  $\bar{y}_i$  and the overall sample mean  $\bar{y}$ , is called the ***between-groups estimate***.

The  $F$  test statistic has the form

$$F = \frac{\text{Between-groups estimate of variance}}{\text{Within-groups estimate of variance}}.$$

This is called the ***analysis of variance  $F$  statistic***, or ***ANOVA  $F$  statistic*** for short. We'll leave the computational details of the variance estimates for later in the section.

The within-groups estimate is an unbiased estimate of  $\sigma^2$  regardless of whether  $H_0$  is true. If  $H_0$  is true, the between-groups estimate is also unbiased, taking about the same value as the within-groups estimate, apart from sampling error. If  $H_0$  is false, however, the between-groups estimate tends to overestimate  $\sigma^2$ . It then tends to be larger than the within-groups estimate, and the  $F$  test statistic tends to be larger than 1.0, moreso with larger samples.

When  $H_0$  is true, this  $F$  test statistic has an  $F$  sampling distribution. As in  $F$  tests for multiple regression parameters (e.g., Section 11.4), the  $P$ -value is the right-tail probability that the  $F$  test statistic exceeds the observed  $F$  value. The larger the  $F$  test statistic, the smaller the  $P$ -value.

#### Example 12.2 $F$ Test Comparing Mean Political Ideology by Party ID

Software displays the results of ANOVA  $F$  tests in a table similar to the one used to display sums of squares in regression analysis. This table is called an ***ANOVA table***. Table 12.2 shows the basic format, illustrating for the  $F$  test of  $H_0: \mu_1 = \mu_2 = \mu_3$  comparing population mean political ideology for three party IDs with Table 12.1.

In the ANOVA table:

- The two ‘mean squares’ are the between-groups and within-groups estimates of the population variance  $\sigma^2$ .
- The  $F$  test statistic is the ratio of the two mean squares.

From the Mean Square column of Table 12.2., the between-groups estimate of the variance is 44.21 and the within-groups estimate is 1.68. The  $F$  test statistic is  $F = 44.21/1.68 = 26.3$ . In other words, the between-groups estimate is more than 25 times the within-groups estimate. Recall that if  $H_0$  is true, we expect similar variance estimates and values of  $F$  near 1.0, apart from sampling error. So, this test statistic gives strong evidence against  $H_0: \mu_1 = \mu_2 = \mu_3$ . Table 12.2 reports that the  $P$ -value is 0.0000, rounded to four decimal places. We conclude that a difference exists among the population mean political ideology values for the three political parties.

Table 12.2: ANOVA Table for Result of  $F$  Test for Table 12.1. The  $F$  test statistic is the ratio of mean squares.

Source	Sum of Squares	df	Mean Square	$F$	Sig
Between-Groups (Party ID)	88.43	2	44.21	26.3	.0000
Within-Groups (Error)	459.52	273	1.68		
Total	547.95	275			

□

## Within-Groups Estimate of Variance

Now we’ll see how to construct the variance estimates that form the  $F$  statistic. Each estimate is a measure of variability divided by a degrees of freedom term.

The within-groups estimate of the population variance  $\sigma^2$  pools together the sums of squares of the observations about their means. Now, for the  $n_1$  observations from the first group,  $\sum(y - \bar{y}_1)^2$  is the sum of squares of the observations about their mean. This sum of squares has  $n_1 - 1$  degrees of freedom, the denominator of this sum for forming the sample variance  $s_1^2$  for group 1. Similarly, for the  $n_2$  observations from the second group,  $\sum(y - \bar{y}_2)^2$  is the sum of squares of the observations about their sample mean, with  $n_2 - 1$  degrees of freedom. The sum of these sum of squares terms for all the samples is called the **within-groups sum of squares**, since the sums of squares are calculated *within* each sample.

The within-groups sum of squares has degrees of freedom equal to the sum of the  $df$  values of the component parts:

$$\begin{aligned} df &= (n_1 - 1) + (n_2 - 1) + \cdots + (n_g - 1) = (n_1 + n_2 + \cdots + n_g) - g \\ &= N - g = \text{Total sample size} - \text{number of groups}, \end{aligned}$$

where  $N$  denotes the total sample size. The ratio

$$s^2 = \frac{\text{Within-groups sum of squares}}{df} = \frac{\text{Within-groups SS}}{N - g}$$

is the within-groups estimate of the population variance  $\sigma^2$  for the  $g$  groups.

This estimate summarizes information about variability from the separate samples. The estimate of  $\sigma^2$  using only the first group is

$$s_1^2 = \frac{\sum (y - \bar{y}_1)^2}{n_1 - 1}.$$

In Table 12.2, for example, this is the square of the reported standard deviation,  $s_1 = 1.28$ . Similarly, the sample variance for the second group is  $s_2^2 = \sum (y - \bar{y}_2)^2 / (n_2 - 1)$ , and so forth for the remaining groups. Under the assumption that the population variances are identical, these terms all estimate the same parameter,  $\sigma^2$ . The numerator and denominator of  $s^2$  pool the information from these estimates by adding their numerators and adding their denominators. The resulting estimate relates to the separate sample variances by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_g - 1)s_g^2}{N - g}.$$

This estimate is a weighted average of the separate sample variances, with greater weight given to larger samples. With equal sample sizes,  $s^2$  is the mean of the  $g$  sample variances.

For the political ideology data in Table 12.1, we could calculate  $s^2$  by calculating the within sum of squares from the raw data. Since Table 12.1 reports the standard deviations, however, it is simpler to use the formula just given, or better yet, use computer software. The sample sizes for the groups are  $n_1 = 91$ ,  $n_2 = 111$ ,  $n_3 = 74$ , for a total sample size of  $N = 276$ . From Table 12.1,

$$\begin{aligned} s^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{N - 3} \\ &= \frac{(91 - 1)(1.28)^2 + (111 - 1)(1.43)^2 + (74 - 1)(1.10)^2}{276 - 3} = \frac{459.5}{273} = 1.68. \end{aligned}$$

In summary, the within-groups sum of squares equals 459.5, with  $df = 273$ , providing a within-groups variance estimate of 1.68. The standard deviation estimate of  $s = \sqrt{1.68} = 1.30$  summarizes the three sample standard deviations from Table 12.1.

### Between-Groups Estimate of Variance

The estimate of  $\sigma^2$  based on variability between each sample mean and the overall sample mean equals

$$\frac{\sum_i n_i (\bar{y}_i - \bar{y})^2}{g - 1} = \frac{n_1 (\bar{y}_1 - \bar{y})^2 + \cdots + n_g (\bar{y}_g - \bar{y})^2}{g - 1}.$$

Exercise 73 motivates this formula. Since this estimate describes variability among  $g$  means, its

$$df = g - 1 = \text{Number of groups} - 1,$$

which is the denominator of the estimate.

The numerator of this estimate is called the **between-groups sum of squares**. The squared difference between each sample mean and the overall mean is weighted by the sample size upon which it is based. When the population means are unequal, the  $\bar{y}_i$  values tend to be more variable than if the population means are equal. The farther the population means fall from the null hypothesis case,  $H_0: \mu_1 = \dots = \mu_g$ , the larger the between-groups SS (sum of squares), the between-groups estimate, and the  $F$  test statistic value tend to be.

For Table 12.1, the overall mean  $\bar{y} = 3.89$ . The between-groups sum of squares equals

$$\begin{aligned} \text{Between-groups SS} &= n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + n_3(\bar{y}_3 - \bar{y})^2 \\ &= 91(3.23 - 3.89)^2 + 111(3.90 - 3.89)^2 + 74(4.70 - 3.89)^2 \\ &= 88.43. \end{aligned}$$

Since  $g = \text{number of groups} = 3$ , this sum of squares has  $df = g - 1 = 3 - 1 = 2$ . The between-groups estimate of the variance is

$$\frac{\text{Between-groups SS}}{g - 1} = \frac{88.43}{2} = 44.2.$$

## Sums of Squares in ANOVA Tables

Let's now look again at the ANOVA table (Table 12.2) for political ideology and political party ID. For the  $F$  test,

$$df_1 = g - 1 = \text{No. groups} - 1 \quad \text{and} \quad df_2 = N - g = \text{Total sample size} - \text{no. groups}.$$

These are reported in the  $df$  column of the table. For these data,  $df_1 = g - 1 = 3 - 1 = 2$  and  $df_2 = N - g = 943 - 3 = 940$ .

In the 'Between-groups' row of the ANOVA table, the between-groups SS divided by  $df_1$  gives a mean square,  $88.43/2 = 44.21$ . In the 'Within-groups' row, the within-groups SS divided by  $df_2$  gives the other mean square,  $459.52/273 = 1.68$ . The  $F$  test statistic for  $H_0: \mu_1 = \mu_2 = \mu_3$  is the ratio of the variance estimates, which is the ratio of the two mean squares,  $F = 44.21/1.68 = 26.3$ . The two  $df$  terms for the test are the denominators of the two estimates of the variance.

The sum of the within-groups and between-groups sums of squares is called the **total sum of squares**. In fact, this equals

$$TSS = \sum (y - \bar{y})^2 = \text{Between-groups SS} + \text{Within-groups SS},$$

the sum of squares of the combined sample of  $N$  observations about the overall mean,  $\bar{y}$ . Table 12.2 shows that  $TSS = 547.95 = 88.43 + 459.52$ .

The ANOVA partitions the total variability about the overall mean, TSS, into two independent parts. One part, the between-groups SS, is the portion of the total explained by the difference between each group mean and the overall mean. This is also called the *group sum of squares*, and most software replaces the ‘Between-groups’ label in Table 12.2 by the name of the group variable (e.g., PARTY ID). The other part, the within-groups SS, is the portion of the total variability that cannot be explained by the differences among the groups. It represents the variability that remains after classifying the observations into separate groups. The within-groups sum of squares is also called the *error sum of squares*, and most software replaces the ‘Within-groups’ label in Table 12.2 by ‘Error.’ Section 12.3 explains the analogy between these sums of squares and the sums of squares in regression analysis.

### The $F$ Test vs. Several $t$ Tests

With two groups, Section 7.3 showed how a  $t$  test can compare the means under the assumption of equal population standard deviations. In fact, if we apply the ANOVA  $F$  test to data from  $g = 2$  groups, the  $F$  test statistic equals the square of the  $t$  test statistic. The  $P$ -value for the  $F$  test is exactly the same as the two-sided  $P$ -value for the  $t$  test. We can use either test to conduct the analysis.

With several groups, instead of using the  $F$  test, why not use a  $t$  test to compare each pair of means? One reason is that using a single  $F$  test rather than multiple  $t$  tests enables us to control the overall probability of type I error. With an  $\alpha$ -level of 0.05 in the  $F$  test, the probability of incorrectly rejecting a true  $H_0$  is fixed at 0.05. By contrast, when we do a separate  $t$  test for each pair of means, a type I error probability applies for *each* comparison. We do not then control the overall type I error rate for all the comparisons. The next section shows how these same considerations are relevant for confidence intervals.

## 12.2 Multiple Comparisons of Means

The analysis of variance  $F$  test of  $H_0: \mu_1 = \mu_2 = \dots = \mu_g$  is a global test of independence of the response and explanatory variables, just like the  $F$  test of  $H_0: \beta_1 = \dots = \beta_k = 0$  for multiple regression models or the chi-squared test for contingency tables. When the  $P$ -value is small, this does not indicate which means are different or how different they are. Confidence intervals help us to determine this. Even if the  $P$ -value is not small, it still is informative to estimate just how large differences in means could plausibly be.

### Confidence Intervals Comparing Means

In practice, it is more informative to estimate the population means than merely to test whether they are all equal. We can construct a confidence interval for each mean or for each difference between a pair of means.

- A confidence interval for  $\mu_i$  is

$$\bar{y}_i \pm t \frac{s}{\sqrt{n_i}}.$$

In this formula,  $s$  is the square root of the within-groups estimate of  $\sigma^2$  that is the denominator of the ANOVA  $F$  test statistic. The  $t$ -value for the chosen confidence level is based on  $df$  for that estimate,  $df = N - g$ .

- A confidence interval for  $\mu_i - \mu_j$  is

$$(\bar{y}_i - \bar{y}_j) \pm ts \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

In this formula also,  $df = N - g$  for the tabled  $t$ -value. Evidence exists of a difference between  $\mu_i$  and  $\mu_j$  when the interval does not contain 0. (For  $g = 2$  groups,  $df = N - g = (n_1 + n_2 - 2)$ ; this confidence interval then simplifies to the one Section 7.3 introduced for  $\mu_2 - \mu_1$  using a pooled standard deviation estimate of a common population standard deviation.)

### Example 12.3 Comparing Mean Ideology of Democrats and Republicans

For Table 12.1, let's compare population mean ideology of Democrats (group 1) and Republicans (group 3). From Table 12.1,  $\bar{y}_1 = 3.23$  for  $n_1 = 91$  Democrats and  $\bar{y}_3 = 4.70$  for  $n_3 = 74$  Republicans. From Table 12.2, the estimate of the population standard deviation is  $s = \sqrt{1.68} = 1.30$ , with  $df = 273$ . For a 95% confidence interval with  $df = 273$ , the  $t$ -score is  $t_{.025} = 1.97$  (essentially the  $z_{.025}$ -score). The confidence interval for  $\mu_3 - \mu_1$  is

$$\begin{aligned} (\bar{y}_3 - \bar{y}_1) \pm t_{.025} s \sqrt{\frac{1}{n_1} + \frac{1}{n_3}} &= (4.70 - 3.23) \pm 1.97(1.30) \sqrt{\frac{1}{91} + \frac{1}{74}} \\ &= 1.47 \pm 0.40 \quad \text{or} \quad (1.07, 1.87). \end{aligned}$$

We infer that population mean ideology was between 1.07 and 1.87 units higher for Republicans than for Democrats. Since the interval contains only positive numbers, we conclude that  $\mu_3 - \mu_1 > 0$ ; that is,  $\mu_3$  exceeds  $\mu_1$ . On the average, Republicans were more conservative than Democrats, with difference about 1 to 2 categories on the 7-category scale.

□

### Error Rates with Large Numbers of Confidence Intervals

With  $g$  groups, there are  $g(g - 1)/2$  pairs of groups to compare. When  $g$  is relatively large, the number of comparisons can be very large. Confidence intervals for some pairs of means may suggest they are different *even if all of the population means are equal*.

When  $g = 10$ , for example, there are  $g(g - 1)/2 = 45$  pairs of means. Suppose we form a 95% confidence interval for the difference between each pair. The error probability of 0.05 applies for each comparison. For the 45 comparisons, we'd expect that  $45(0.05) = 2.25$  of the intervals would not contain the true differences of means.

For 95% confidence intervals, the error probability of 0.05 is the probability that any particular confidence interval will not contain the true difference in population means. When we form a large number of confidence intervals, the probability that *at least* one confidence interval will be in error is much larger than the error probability for any particular interval. The larger the number of groups to compare, the greater is the chance of at least one incorrect inference.

### Bonferroni Multiple Comparisons of Means

When we plan many comparisons, methods are available that control the probability that *all* intervals will contain the true differences. Such methods are called **multiple comparison** methods. For them, all intervals contain the true parameter values *simultaneously* with an overall fixed probability.

For example, with a multiple comparison method applied with  $g = 10$  means and 95% confidence, the probability equals 0.95 that *all* 45 of the intervals will contain the pairwise differences  $\mu_i - \mu_j$ . Equivalently, the probability that *at least one* interval is in error equals 0.05. This probability is called the **multiple comparison error rate**.

We first present the **Bonferroni multiple comparison** method, since it is simple and applies to a wide variety of situations. The Bonferroni method uses the same formulas for confidence intervals introduced at the beginning of this section. However, it uses a more stringent confidence level for each interval, to ensure that the overall confidence level is sufficiently high.

To illustrate, suppose we'd like a multiple comparison error rate of 0.10, that is, a probability of 0.90 that all confidence intervals are simultaneously correct. If we plan four comparisons of means, then the Bonferroni method uses error probability  $0.10/4 = 0.025$  for each one. That is, it uses a 97.5% confidence level for each interval. This approach is somewhat conservative: It ensures that the actual overall error rate is *at most* 0.10 and that the overall confidence level is *at least* 0.90. The method is based on an inequality shown by the Italian probabilist Carlo Bonferroni in 1935. It states that the probability that at least one of a set of events occurs can be no greater than the sum of the separate probabilities of the events. For instance, if the probability of an error equals 0.025 for each of four confidence intervals, then the probability that at least one of the four intervals will be in error is no greater than  $(0.025 + 0.025 + 0.025 + 0.025) = 0.10$ .

#### Example 12.4 Bonferroni Intervals for Political Ideology Comparisons

We refer again to Table 12.1. For the  $g = 3$  groups, we compare  $\mu_1$  with  $\mu_2$ ,  $\mu_1$  with  $\mu_3$ , and  $\mu_2$  with  $\mu_3$ . We construct confidence intervals having overall confidence level at least 0.95. For a multiple comparison error rate of 0.05 with three comparisons,

the Bonferroni method uses error probability  $0.05/3 = 0.0167$  for each interval. These use the  $t$ -score with two-tail probability 0.0167, or single-tail probability 0.0083. For the large  $df$  value here ( $df = 273$ ), this equals 2.41, close to the  $z$ -score of 2.39. Recall also that  $s = 1.30$ .

The interval for  $\mu_3 - \mu_1$ , the difference between the population mean ideology of Republicans and Democrats, is

$$\begin{aligned} (\bar{y}_2 - \bar{y}_1) \pm ts \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &= (4.70 - 3.23) \pm 2.41(1.30) \sqrt{\frac{1}{91} + \frac{1}{74}} \\ &= 1.47 \pm 0.49 \quad \text{or} \quad (0.98, 1.96). \end{aligned}$$

We construct the intervals for the other two pairs of means in a similar way. Table 12.3 displays them. All three confidence intervals hold with overall confidence level at least 0.95. The probability that one or more of them does not contain the true difference is at most 0.05.

Table 12.3: Bonferroni and Tukey 95% Multiple Comparisons of Mean Political Ideology for Three Political Party ID Groups. The 95% confidence applies to the entire set of three intervals, rather than each individual interval.

Groups	Difference Of Means $\mu_i - \mu_j$	Estimated Difference $\bar{y}_i - \bar{y}_j$	Bonferroni 95% CI	Tukey 95% CI
(Independent, Democrat)	$\mu_2 - \mu_1$	0.67	(0.23, 1.11)*	(0.24, 1.10)*
(Republican, Democrat)	$\mu_3 - \mu_1$	1.47	(0.98, 1.96)*	(0.99, 1.95)*
(Republican, Independent)	$\mu_3 - \mu_2$	0.80	(0.33, 1.27)*	(0.34, 1.26)*

Note: An asterisk \* indicates a significant difference.

None of the intervals contain 0. They show significant evidence of a difference between each pair of population means.

□

The Bonferroni 95% multiple comparison confidence intervals are wider than separate 95% confidence intervals. For instance, the ordinary 95% confidence interval comparing Republicans and Democrats is (1.07, 1.87), whereas the Bonferroni interval is (0.98, 1.96). This is because the multiple comparison method uses a higher confidence level for each separate interval to ensure achieving the overall confidence level that applies to the entire set of comparisons.

## Tukey Multiple Comparisons of Means

Of the other methods available for multiple comparisons, we recommend *Tukey's method*. Proposed by the great statistician John Tukey, who also developed exploratory data analysis methods such as box plots and stem-and-leaf plots as well

as terminology such as *software*, this method has intervals that are slightly narrower than the Bonferroni intervals. This is because they are designed to *approximate* the nominal confidence level rather than to have *at least* that level. The Tukey method uses a probability distribution (the *Studentized range*) that refers to the difference between the largest and smallest sample means. We do not present this distribution in this text, so we rely on software rather than a formula for the Tukey intervals.

Table 12.3 shows Tukey intervals for the political ideology data. For practical purposes, they provide the same conclusions as the Bonferroni intervals.

## 12.3 Performing ANOVA by Regression Modeling

Chapter 11 used multiple regression to model the relationship between the mean of a quantitative response variable and a collection of *quantitative* explanatory variables. The analysis of variance (ANOVA) models the relationship between the mean of a quantitative response variable and a *categorical* explanatory variable, the categories of which are the groups compared. In fact, ANOVA is a special case of multiple regression. Artificial explanatory variables in a regression model can represent the groups.

### Regression with Dummy Variables

We set up an artificial variable to equal 1 if an observation comes from a particular group and 0 otherwise. With three groups, as in the political ideology example, we use two artificial variables. The first, denoted by  $z_1$ , equals 1 for observations from the first group and equals 0 otherwise. The second, denoted by  $z_2$ , equals 1 for observations from the second group and equals 0 otherwise. That is,

$$\begin{aligned} z_1 = 1 \text{ and } z_2 = 0: & \text{ observations from group 1} \\ z_1 = 0 \text{ and } z_2 = 1: & \text{ observations from group 2} \\ z_1 = 0 \text{ and } z_2 = 0: & \text{ observations from group 3} \end{aligned}$$

It is unnecessary and redundant to create a variable for the last (third) group, because values of 0 for  $z_1$  and  $z_2$  identify observations from it.

The artificial variables  $z_1$  and  $z_2$  are called ***dummy variables***. They identify the group for an observation. That is, they give a classification, not a magnitude, for the categorical predictor. Table 12.4 summarizes the dummy variables for three groups.

For the dummy variables just defined, consider the multiple regression equation

$$E(Y) = \alpha + \beta_1 z_1 + \beta_2 z_2.$$

For observations from group 3,  $z_1 = z_2 = 0$ . The equation then simplifies to

$$E(Y) = \alpha + \beta_1(0) + \beta_2(0) = \alpha.$$

Table 12.4: The Two  
Dummy Variables for  
Three Groups

Group	$z_1$	$z_2$
1	1	0
2	0	1
3	0	0

So,  $\alpha$  represents the population mean  $\mu_3$  of  $Y$  for the last group. For observations from group 1,  $z_1 = 1$  and  $z_2 = 0$ , so

$$E(Y) = \alpha + \beta_1(1) + \beta_2(0) = \alpha + \beta_1$$

equals the population mean  $\mu_1$  for that group. Similarly,  $\alpha + \beta_2$  equals the population mean  $\mu_2$  for group 2 (let  $z_1 = 0$  and  $z_2 = 1$ ).

Since  $\alpha + \beta_1 = \mu_1$  and  $\alpha = \mu_3$ ,  $\beta_1$  represents the difference  $\mu_1 - \mu_3$ . Similarly,  $\beta_2 = \mu_2 - \mu_3$ . Table 12.5 summarizes the parameters of the regression model and their correspondence with the population means. The  $\beta$  coefficient of a dummy variable represents the difference between the mean for the group that dummy variable represents and the mean of the group not having its own dummy variable.

Table 12.5: Interpretation of Coefficients of Dummy  
Variables in Model  $E(Y) = \alpha + \beta_1 z_1 + \beta_2 z_2$

Group	$z_1$	$z_2$	Mean of $Y$	Interpretation of $\beta$
1	1	0	$\mu_1 = \alpha + \beta_1$	$\beta_1 = \mu_1 - \mu_3$
2	0	1	$\mu_2 = \alpha + \beta_2$	$\beta_2 = \mu_2 - \mu_3$
3	0	0	$\mu_3 = \alpha$	

Dummy variable coding works because it allows the population means to take arbitrary values, with no assumed distances between groups. Using a single artificial variable with coding such as  $z = 1$  for group 1,  $z = 2$  for group 2, and  $z = 3$  for group 3 would not work. The model  $E(Y) = \alpha + \beta z$  would then assume an ordering as well as equal distances between groups. It treats the categorical variable as if it were quantitative, which is improper. Whereas it takes only one term in a regression model to represent the linear effect of a quantitative explanatory variable, it requires  $g - 1$  terms to represent the  $g$  categories of a categorical variable.

### Example 12.5 Regression Model for Political Ideology and Party ID

For Table 12.1, the group variable (Party ID) has three categories. The regression model for the ANOVA procedure with  $Y =$  political ideology is

$$E(Y) = \alpha + \beta_1 z_1 + \beta_2 z_2.$$

The dummy variables satisfy  $z_1 = 1$  only for Democrats,  $z_2 = 1$  only for Independents, and  $z_1 = z_2 = 0$  for Republicans. Table 12.6 shows a portion of a printout for fitting this regression model. No dummy variable estimate appears in the table for Party 3 (Republicans), because it is redundant to include a dummy variable for the last group.

Table 12.6: Printout for Fitting Regression Model  $E(Y) = \alpha + \beta_1 z_1 + \beta_2 z_2$  to Data on  $Y =$  Political Ideology with Dummy Variables  $z_1$  and  $z_2$  for Party ID

Dependent Variable: IDEOLOGY

Parameter		Estimate	Std Error	t	Sig
(Constant)		4.534	0.0759	59.73	0.0001
PARTY	1	-0.717	0.1033	-6.94	0.0001
	2	-0.541	0.1054	-5.13	0.0001
	3	0.000	0.	0.	0.

The prediction equation is  $\hat{y} = 4.53 - 0.72z_1 - 0.54z_2$ . The coefficients in the prediction equation relate to the sample means in the same manner that the regression parameters relate to the population means. Just as  $\alpha = \mu_3$ , so does its estimate  $4.53 = \bar{y}_3$ , the sample mean for Republicans. Similarly, the coefficient of  $z_1$  is  $-0.72 = \bar{y}_1 - \bar{y}_3$  and the coefficient of  $z_2$  is  $-0.54 = \bar{y}_2 - \bar{y}_3$ .

□

### Regression for ANOVA Test Comparing Means

For three groups, the null hypothesis in the ANOVA  $F$  test is  $H_0: \mu_1 = \mu_2 = \mu_3$ . If  $H_0$  is true, then  $\mu_1 - \mu_3 = 0$  and  $\mu_2 - \mu_3 = 0$ . Since  $\mu_1 - \mu_3 = \beta_1$  and  $\mu_2 - \mu_3 = \beta_2$  in the multiple regression model  $E(Y) = \alpha + \beta_1 z_1 + \beta_2 z_2$  with dummy variables, the ANOVA hypothesis is equivalent to  $H_0: \beta_1 = \beta_2 = 0$  in that model. If all  $\beta$ -values in the model equal 0, then the mean of the response variable equals  $\alpha$  for each group. By setting up dummy variables, then, we can perform the ANOVA test using the  $F$  test of  $H_0: \beta_1 = \beta_2 = 0$  for this regression model.

The assumption from regression analysis that the conditional distributions of  $Y$  about the regression equation are normal with constant standard deviation implies here that the population distributions for the groups are normal, with the same standard deviation for each group. These are precisely the assumptions for the ANOVA  $F$  test.

**Example 12.6 Regression for Comparing Political Ideology Means by Party ID**

Table 12.7 shows the sums of squares for fitting the regression model with dummy variables to the data on political ideology and party ID. Notice the similarity between this and the ANOVA table in Table 12.2. The ‘between-groups sum of squares’ in ANOVA is the ‘regression sum of squares’ in the regression analysis. The ‘within-groups sum of squares’ in ANOVA is the ‘residual sum of squares’ (also called ‘sum of squared errors’) and denoted by SSE. This is the variability within the groups unexplained by including parameters in the model to account for the differences between the means. The sum of squared errors divided by its degrees of freedom is the mean square error (MSE), which is the within-groups estimate  $s^2 = 1.67$  of the variance of observations for each group. The regression mean square is the between-groups estimate.

Table 12.7: Printout Showing Sums of Squares for Regression Model  $E(Y) = \alpha + \beta_1 z_1 + \beta_2 z_2$  for Modeling Political Ideology in Terms of Party ID

---

	Sum of Squares	df	Mean Square	F Value	Sig
Regression	85.382	2	42.691	25.55	0.0001
Residual	1570.837	940	1.671		
Total	1656.218	942			

---

The ratio of the regression mean square to the mean square error is the  $F$  test statistic ( $F = 25.55$ ), with  $df_1 = 2$  and  $df_2 = 940$ , for testing  $H_0: \beta_1 = \beta_2 = 0$ . This hypothesis is equivalent to  $H_0: \mu_1 = \mu_2 = \mu_3$  for the three party IDs. The regression analysis provides the same  $F$  statistic as ANOVA did in Section 12.1.

□

## 12.4 Two-Way Analysis of Variance

We’ve seen how to compare means for groups that are categories of a categorical explanatory variable. Sometimes the groups refer to two (or more) categorical variables. For example, the groups (white men, white women, black men, black women) result from cross-classifying race and gender. The method for comparing the mean of a quantitative response variable across categories of each of two categorical variables is called a *two-way ANOVA*.

The ANOVA discussed so far, having a single explanatory variable, is called *one-way ANOVA*. It ignores other variables. Chapters 10 and 11 showed that such analyses are usually not as informative as multivariate analyses that control other

variables. The rest of this chapter deals with two-way ANOVA and more complex methods for categorical explanatory and control variables.

## Main Effect Hypotheses in Two-Way ANOVA

Two-way ANOVA compares population means across categories of two explanatory variables. Each null hypothesis states that the population means are identical across categories of one categorical variable, controlling for the other one.

To illustrate two-way ANOVA, we analyze mean political ideology using explanatory variables party ID and gender. Six means result from the  $2 \times 3 = 6$  combinations of their categories, as Table 12.8 shows. Let  $\mu_{ij}$  denote the population mean political ideology for gender  $i$  and party ID  $j$ . For example,  $\mu_{23}$  denotes mean political ideology for row 2 and column 3, males who are Republicans.

Table 12.8: A Two-Way Classification of Population Mean Political Ideology by Party Identification and Gender

Gender	Party Identification		
	Democrat	Independent	Republican
Female	$\mu_{11}$	$\mu_{12}$	$\mu_{13}$
Male	$\mu_{21}$	$\mu_{22}$	$\mu_{23}$

One possible analysis compares the mean political ideology for the three party IDs, controlling for gender. For females, we compare the means  $\mu_{11}$ ,  $\mu_{12}$ , and  $\mu_{13}$  for the three party IDs; for males, we compare  $\mu_{21}$ ,  $\mu_{22}$ , and  $\mu_{23}$  for the three party IDs. Another possible analysis compares the mean political ideology for males and females, controlling for party ID, by comparing means within each column of the table.

Table 12.9a displays a set of population means satisfying the null hypothesis that mean political ideology is identical for the three party IDs, controlling for gender. Table 12.9b displays a set of population means satisfying the null hypothesis that mean political ideology is identical for the two genders, controlling for party ID. The effects of individual predictors tested in these two null hypotheses are called *main effects*.

## F Tests about Main Effects

The  $F$  tests for two-way ANOVA have the same assumptions as the  $F$  test for one-way ANOVA: Randomization, a normal population distribution for each group, with the same standard deviation for each group. Now, the groups are formed by the cells of the cross-classification of the two explanatory variables.

The test statistics for two-way ANOVA have complex formulas except when the sample sizes in all cells are equal. We'll rely on software. As in one-way ANOVA, the test for a predictor effect uses two estimates of the variance for each group. These

Table 12.9: Population Mean Political Ideology Satisfying Main Effect Null Hypotheses: (a) No Effect of Party Identification, (b) No Effect of Gender

Table	Gender	Party Identification		
		Democrat	Independent	Republican
(a)	Female	3.0	3.0	3.0
	Male	5.0	5.0	5.0
(b)	Female	3.0	4.0	5.0
	Male	3.0	4.0	5.0

estimates appear in the mean square (MS) column of the ANOVA table. For testing the main effect for a predictor, the test statistic is the ratio of mean squares,

$$F = \frac{\text{MS for the predictor}}{\text{MS error}}.$$

The MS for the predictor is a variance estimate based on between-groups variation for that predictor. The MS error is a within-groups variance estimate that is always unbiased.

### Example 12.7 Two-Way ANOVA for Political Ideology by Party ID and Gender

Table 12.10 shows GSS data for political ideology by party ID and gender (with no restrictions on age). The table also shows the sample means and standard deviations of political ideology, based on scores (1, 2, 3, 4, 5, 6, 7).

Table 12.10: GSS Data on Political Ideology by Party Identification and Gender

Party ID	Gender	Political Ideology							Sample Size	Mean	Std. Dev.
		1	2	3	4	5	6	7			
Democrat	Female	5	30	35	98	20	24	3	215	3.85	1.26
	Male	6	20	25	41	15	15	3	125	3.77	1.43
Independent	Female	4	17	27	83	16	17	5	169	3.95	1.24
	Male	4	16	20	59	21	23	1	144	4.04	1.30
Republican	Female	2	10	17	63	32	33	5	162	4.43	1.26
	Male	0	9	13	36	33	28	9	128	4.66	1.31

Software reports Table 12.11 for summarizing the analyses. The mean square error (MSE) estimates the population variance within each cell. For testing main effects, it equals

$$s^2 = \frac{\text{SSE}}{df} = \frac{1569.53}{939} = 1.67.$$

Table 12.11: ANOVA Table for Two-Way Analysis of Main Effects of Party Identification and Gender on Mean Political Ideology

---

Dependent Variable: IDEOLOGY

Source	Sum of Squares	df	Mean Square	F	Sig
Model	86.693	3	28.898	17.29	0.0001
Error	1569.525	939	1.671		
Total	1656.218	942			

  

Source	Type III SS	df	Mean Square	F	Sig
PARTY	84.2516	2	42.1258	25.20	0.0001
GENDER	1.3110	1	1.3110	0.78	0.3760

---

We test each null hypothesis by comparing this to another estimate of the variance that tends to be inflated when  $H_0$  is not true. The variance estimates, listed in Table 12.11 in the Mean Square column, divide each sum of squares by its  $df$  value. The  $F$  test statistics are the ratios of these estimates to  $s^2$ . The degrees of freedom for the  $F$  statistics are  $df_1 = df$  for the numerator estimate, and  $df_2 = df$  for  $s^2$  (939, in this case). As usual, the  $P$ -value is the right-tail probability.

For the null hypothesis of no difference in mean political ideology for the three party IDs, controlling for gender, Table 12.11 shows that the  $F$  test statistic is

$$F = \frac{\text{Party ID mean square}}{\text{Mean square error}} = \frac{42.13}{1.67} = 25.2,$$

with  $df_1 = 2$  and  $df_2 = 939$ . The  $P$ -value is 0.0001. Very strong evidence exists of a difference in mean political ideology among the three party IDs, controlling for gender.

For the null hypothesis of no difference in mean political ideology between females and males, controlling for party ID, the  $F$  test statistic is

$$F = \frac{\text{Gender mean square}}{\text{Mean square error}} = \frac{1.31}{1.67} = 0.78,$$

with  $df_1 = 1$  and  $df_2 = 939$ . The  $P$ -value is  $P = 0.38$ . There is negligible evidence that mean political ideology varies by gender, within each party ID.

□

## Interaction in Two-Way ANOVA

In practice, before conducting the main effects tests just described, we would first test another null hypothesis. Sections 10.3 and 11.5 showed that the study of *interaction* is

Figure 12.3: Mean Political Ideology, by Party Identification and Gender, Displaying No Interaction

((Fig. 12.3 in 3e))

important whenever we analyze multivariate relationships. An absence of interaction between two explanatory variables means that the effect of either variable on the response variable (in the population) does not change for different levels of the other.

Suppose that no interaction exists between gender and party ID in their effects on political ideology. Then, the difference between each pair of party IDs in population mean political ideology is the same for males and females. Also, the difference between females and males in population mean political ideology is the same for each party ID.

Table 12.12 shows population means satisfying a lack of interaction. The difference between males and females in mean political ideology is 1.0 for each party. Similarly, the difference between each pair of parties in mean political ideology is the same for each gender. The difference between Republicans and Democrats, for example, equals 2.0 both for females and for males. Figure 12.3 plots the means for the party ID categories, within each gender. The ordering of categories on the horizontal axis is unimportant, since party ID is nominal. The absence of interaction is indicated by the parallel sequences of points.

Table 12.12: Population Means for a Two-Way Classification, Displaying No Interaction

Gender	Democrat	Independent	Republican
Female	3.0	3.5	5.0
Male	4.0	4.5	6.0

By contrast, Table 12.13 and Figure 12.4 show population means displaying interaction. The difference between females and males in mean political ideology is  $-2$  for Democrats,  $0$  for Independents, and  $+2$  for Republicans. Here, the difference in means between females and males depends on the party ID. Similarly, the party ID effect on ideology differs for females and males. For females, Republicans are the most conservative, whereas for males, Democrats are the most conservative.

In Table 12.13, suppose the numbers of males and females are equal, for each party ID. Then the overall mean political ideology, ignoring gender, is  $4.0$  for each party. The overall difference in means between any two party IDs equals  $0$ . In a one-way

Table 12.13: Population Means for a Two-Way Classification, Displaying Interaction

Gender	Democrat	Independent	Republican	
Females	3.0	4.0	5.0	predictor
Males	5.0	4.0	3.0	

Figure 12.4: Mean Political Ideology, by Party Identification and Gender, Displaying Interaction

((Fig. 12.4 in 3e))

comparison of mean political ideology by party ID, party ID has no effect. However, in a two-way comparison, the interaction implies differing party ID effects for males and females. As in other multivariate analyses, the effect of a predictor can change dramatically when we control for another variable.

### **$F$ Test of $H_0$ : No Interaction**

Besides the main effects hypotheses, the other relevant null hypothesis in two-way ANOVA is  $H_0$ : no interaction. This  $H_0$  is true if the difference between population means for two categories of one predictor is the same for each category of the other predictor, as in Table 12.12. A small  $P$ -value in this test suggests that each categorical predictor has an effect on the response, but the size of effect varies according to the category of the other predictor.

When interaction exists, it is not meaningful to test the main effects hypotheses. When we reject  $H_0$ : no interaction, we conclude that each variable has an effect, but the nature of that effect changes according to the category of the other variable. For that reason, in two-way ANOVA, we first test  $H_0$ : no interaction. If the evidence of interaction is not strong (i.e., if the  $P$ -value is not small), we then test the two main effect hypotheses. On the other hand, if important evidence of interaction exists, it's better to compare the means for one predictor separately within categories of the other.

Table 12.10 showed the sample mean political ideology for the six combinations of party ID and gender. The means show no obvious evidence of interaction. For each gender the sample mean conservatism is higher for Republicans than Democrats. In the next section, we'll see that the test of  $H_0$ : no interaction has  $F = 1.06$  and a  $P$ -value of  $P = 0.34$ . So, a lack of interaction is plausible, and the main effect tests are valid.

Next, it's natural to use confidence intervals to find out more about the party ID effect, controlling for gender. To do this, it's helpful to study two-way ANOVA in the context of regression modelling. As in one-way ANOVA, the  $F$  tests in two-way ANOVA are equivalently tests about parameters in a regression model. The next section shows how to conduct the analyses using multiple regression models with dummy variables for the categorical predictors.

## 12.5 Two-Way ANOVA and Regression

To conduct the  $F$  tests as special cases of tests about parameters of a multiple regression model, we set up dummy variables for each predictor. We illustrate with Table 12.10, on political ideology with the predictors party identification and gender (sex).

We use the symbol  $p$  for dummy variables for party ID and  $s$  as a dummy variable for sex (whether a subject is female). That is,

$$\begin{aligned} p_1 &= \begin{cases} 1 & \text{if subject is Democrat} \\ 0 & \text{otherwise} \end{cases} \\ p_2 &= \begin{cases} 1 & \text{if subject is Independent} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Both  $p_1$  and  $p_2$  equal 0 when the subject is Republican. Also

$$s = \begin{cases} 1 & \text{if subject is female} \\ 0 & \text{if subject is male} \end{cases}$$

It is redundant to include dummy variables for the final categories.

### Regression Model Assuming No Interaction

For simplicity, we'll first assume there's no interaction. In practice, we'd check this assumption first, as we discussed in the previous section. The regression model is

$$E(Y) = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s.$$

To find the correspondence between the population means and the regression parameters, we substitute the possible values for the dummy variables. To illustrate, for Republicans ( $p_1 = p_2 = 0$ ) who are female ( $s = 1$ ), the mean political ideology is

$$\mu = \alpha + \beta_1(0) + \beta_2(0) + \beta_3(1) = \alpha + \beta_3.$$

For the six combinations of party ID and gender, Table 12.14 expresses the population means in terms of the regression parameters. For each party ID, the difference in means between females and males equals  $\beta_3$ . That is, the coefficient  $\beta_3$  of the dummy variable  $s$  for sex equals the difference between females and males in mean political ideology, controlling for party ID. The null hypothesis of no difference between females and males in the mean, controlling for party ID, is  $H_0 : \beta_3 = 0$ .

Table 12.14: Population Means of Political Ideology for the Two-Way Classification of Party ID and Gender (Sex), Assuming No Interaction

Gender	Party Identification	Dummy Variables			Population Mean of $Y$
		$p_1$	$p_2$	$s$	$\alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s$
Female	Democrat	1	0	1	$\alpha + \beta_1 + \beta_3$
	Independent	0	1	1	$\alpha + \beta_2 + \beta_3$
	Republican	0	0	1	$\alpha + \beta_3$
Male	Democrat	1	0	0	$\alpha + \beta_1$
	Independent	0	1	0	$\alpha + \beta_2$
	Republican	0	0	0	$\alpha$

The test of  $H_0: \beta_3 = 0$  of no gender effect compares the complete model

$$E(Y) = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s$$

to the reduced model,

$$E(Y) = \alpha + \beta_1 p_1 + \beta_2 p_2.$$

The reduced model lacks the gender effect. In the ANOVA table shown in Table 12.11, the gender sum of squares is the amount of the variation accounted for by introducing the term  $\beta_3 s$  into the model, once the other terms are already there. It represents the difference between the sums of squared errors (SSE) when these terms are omitted and when they are included. There is a difference of one parameter in the two models, so this sum of squares has  $df = 1$ . As usual, the  $df$  value for SSE equals the total sample size minus the number of parameters in the regression model. For the complete model, there are 4 parameters, so  $df = 943 - 4 = 939$ . Table 12.11 showed that the  $F$  test statistic equals 0.78, with  $P$ -value  $P = 0.38$ .

In the complete model,  $\beta_1$  is the difference between the means for Democrats and Republicans, and  $\beta_2$  is the difference between the means for Independents and Republicans, controlling for gender. The interpretations are similar to the  $\beta$ -values for the regression model for one-way ANOVA, except that here we also control for gender. The null hypothesis of no differences among the parties in mean political ideology, controlling for gender, is  $H_0: \beta_1 = \beta_2 = 0$ . We observed in Table 12.11 that the  $F$  test statistic equals 25.2, with a  $P$ -value of 0.0001.

Table 12.15 shows some output for fitting the complete regression model. The prediction equation is

$$\hat{y} = 4.58 - 0.71p_1 - 0.54p_2 - 0.08s.$$

The coefficient of  $s$  is  $-0.08$ . This is the estimated difference between females and males in mean political ideology, for each party ID. The test of the gender main effect indicated that this difference is not statistically significant. The coefficient of  $p_1$  is

Table 12.15: Fit of Regression Model for Two-Way Analysis of Mean Political Ideology by Party Identification and Gender, Assuming No Interaction. The estimate is 0 at the last level of each predictor, because a dummy variable for that level is not needed and would be redundant.

---

Dependent Variable: IDEOLOGY

Parameter		B	Std. Error	t	Sig
Intercept		4.5768	0.0897	51.02	0.0001
PARTY	1	-0.7112	0.1035	-6.87	0.0001
	2	-0.5423	0.1054	-5.15	0.0001
	3	0	.	.	.
GENDER	1	-0.0758	0.0856	-0.89	0.3760
	2	0	.	.	.

---

-0.71. This is the estimated difference between Democrats and Republicans in mean political ideology, for each gender. The coefficient of  $p_2$  is -0.54. This is the estimated difference between Independents and Republicans, for each gender. The estimated difference between Democrats and Independents is  $(-0.71) - (-0.54) = -0.17$ , for each gender.

Substituting dummy variable values into the prediction equation yields estimated means that satisfy the no interaction model. For instance, for female Republicans,  $p_1 = p_2 = 0$  and  $s = 1$ , so  $\hat{y} = 4.58 - 0.71(0) + 0.54(0) - 0.08(1) = 4.50$ .

## Regression Model with Interaction

The model considered so far is inadequate when there is interaction. Section 11.5 showed that cross-product terms in a multiple regression model can represent interaction. Here, we take cross-products of dummy variables to obtain a regression model that allows interaction effects.

The interaction model for the two-way classification of party ID and gender is

$$E(Y) = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s + \beta_4 (p_1 \times s) + \beta_5 (p_2 \times s).$$

The last two terms use cross-products for the interaction. It is not necessary to take cross-products of dummy variables from categories of the same categorical predictor, such as  $p_1 p_2$ . This is because no more than one dummy variable for a given predictor can be nonzero for any observation, since an observation cannot fall in more than one category. All such cross-products would equal 0.

Table 12.16 shows an ANOVA table for the model that allows interaction. The sum of squares for interaction, shown in the row with the product label PARTY \*

GENDER, is the amount of variability explained by the two interaction terms. It equals the difference between SSE without and with these terms in the model. The interaction mean square is an estimate of  $\sigma^2$  based on

$$\frac{\text{Interaction SS}}{df} = \frac{3.64}{2} = 1.82.$$

We test  $H_0$ : no interaction, that is,  $H_0: \beta_4 = \beta_5 = 0$ , using

$$F = \frac{\text{Interaction mean square}}{\text{Mean square error}} = \frac{1.82}{1.67} = 1.09.$$

From the  $F$  distribution with  $df_1 = 2$  and  $df_2 = 937$ , the  $P$ -value is  $P = 0.34$ .

Table 12.16: ANOVA Table for Two-Way Analysis of Mean Political Ideology by Party Identification and Gender, Allowing Interaction

Dependent Variable: IDEOLOGY

Source	Sum of Squares	df	Mean Square	F	Sig
Model	90.332	5	18.066	10.81	0.0001
Error	1565.886	937	1.671		
Total	1656.218	942			

  

Source	Type III SS	df	Mean Square	F	Sig
PARTY	87.795	2	43.898	26.27	0.0001
GENDER	1.488	1	1.488	0.89	0.3456
PARTY*GENDER	3.640	2	1.820	1.09	0.3370

There is not much evidence of interaction. It is sensible to remove the cross-product terms and use the simpler model discussed previously. Since an absence of interaction is plausible, the main effect tests presented in Table 12.11 for party ID and gender are valid. When there is not significant evidence of interaction, it is better to use the model without interaction terms in testing the main effects of the predictors and in constructing confidence intervals for the effects.

## Partial Sums of Squares

The sums of squares for party, gender, and their interaction in Tables 12.11 and 12.16 are called *partial sums of squares* (See Exercise 60 in Chapter 11). Some software labels them as *Type III sums of squares*. They represent the variability in  $Y$  explained by those terms, once the other terms are already in the model. This equals the difference between the SSE values for the model without those terms and the model with them.

Suppose the predictors are independent, in the sense that the numbers of observations in the cells of the cross classification satisfy independence. An example is when the same number of subjects occur at each combination of party ID and gender. Then, these sums of squares explain completely separate portions of the variability in  $Y$ . They then sum to the regression sum of squares. Recall that the regression sum of squares equals  $TSS - SSE$ . It represents the variability explained together by all the terms in the model.

Designed experiments often have independent predictors. For example, this happens when an experiment assigns the same number of subjects to each cell of a two-way classification of two predictors, such as a driving simulation experiment that observes reaction time to a red light for 20 females using cell phones, 20 females not using cell phones, 20 males using cell phones, and 20 males not using cell phones. For survey research with observational data, predictors are rarely independent. Gender is somewhat associated with party ID, for instance. Because of this, the partial sum of squares for party ID overlaps somewhat with the partial sum of squares for gender. Consequently, the Type III sums of squares listed in Tables 12.11 and 12.16 do not add up exactly to the regression sum of squares.

When the categorical predictors are associated, the partial sum of squares explained by a predictor depends on whether the interaction terms are in the model. The partial sums of squares for party and for gender differ slightly between Tables 12.11 and 12.16.

## Multiple Comparisons Following Two-Way ANOVA

In practice, we follow-up  $F$  tests with confidence intervals to estimate the sizes of effects. Suppose  $H_0$ : no interaction seems plausible. Then, we can treat the difference in population means between two categories for one predictor as the same at each category of the other. So, we construct a single set of comparisons, rather than a separate set at each category of the other variable. In Example 12.6, we estimate the differences in the mean political ideology between each pair of party IDs, controlling for gender, a total of three comparisons. We can do this using ordinary confidence intervals for regression parameters. The form is the usual one of estimate plus and minus a  $t$ -score times the standard error, with  $df$  for  $t$  being the  $df$  value for the mean square error.

For instance, for the model assuming no interaction,  $\hat{\beta}_1 = -0.71$  is the estimated difference between Democrats and Republicans in mean political ideology, controlling for gender. The standard error of this estimate, reported in Table 12.11, is 0.104. A 95% confidence interval is  $-0.71 \pm 1.96(0.104)$ , or  $(-0.9, -0.5)$ . Democrats are less conservative, on the average, for each gender.

The Bonferroni approach (Section 12.2) for one-way ANOVA extends to higher-way ANOVA. A comparison of all three pairs of party IDs with a multiple comparison error rate of 0.05 uses error probability  $0.05/3 = 0.0167$  in determining the  $t$  score for each interval. For these data, we obtain similar intervals to those shown in Table 12.3 following the one-way ANOVA.

When a practically significant degree of interaction exists, it is not appropriate to

make summary comparisons of categories of one variable, controlling for the other. Instead, compare the pairs of rows separately within each column and/or compare the pairs of columns separately within each row.

## Factorial ANOVA

The methods of two-way ANOVA extend to several predictors. Categorical explanatory variables in ANOVA are often called **factors**. A multi-factor ANOVA with observations from all the combinations of the factors is called **factorial ANOVA**.

For instance, with three factors, **three-way ANOVA** considers possible interactions as well as main effects for those factors. For factors denoted by A, B, and C, the full model contains a main effect for each factor, A×B, A×C, and B×C two-factor interactions, and the A×B×C three-factor interaction. The regression model has a set of dummy variables for each factor, cross-products of pairs of dummy variables for the two-factor interactions, and three-way products of dummy variables from all three factors for the three-factor interaction.

In three-way ANOVA, we first test the three-factor interaction. If the *P*-value is small, we compare pairs of categories for one variable at each combination of categories of the other two. Otherwise, it's better to drop the three-factor term from the model and test the two-factor interactions. Suppose, for instance, that the *P*-value is small for the A×B interaction but not for the others. After re-fitting the model with the main effects and the A×B interaction, we can test the C main effect and compare pairs of means for various pairs of categories of C. Because of the A×B interaction, we compare means from categories of A separately at each category of B, and we compare means from categories of B separately at each category of A.

When you have two or more factors, why not instead perform separate one-way ANOVAs? For instance, you could compare the mean political ideology for females and males using a one-way ANOVA, ignoring the information about race. Likewise, you could perform a separate one-way ANOVA to compare the means for blacks and whites, ignoring the information about gender. The main reason is that with factorial ANOVA we learn whether there is interaction. When there is, it is more informative to compare levels of one factor separately at each level of the other factor. This enables us to investigate how the effect depends on that other factor.

Yet another benefit of factorial ANOVA is that the residual variability, which affects the MS error and the denominators of the *F* test statistics, tends to decrease. When we use two or more factors to predict a response variable, we usually tend to get better predictions (that is, less residual variability) than when we use one factor. With less residual (within-groups) variability, we get larger test statistics, and hence greater power for rejecting false null hypotheses.

## 12.6 Repeated Measures Analysis of Variance

The methods presented so far assume that the samples in the groups are *independent*, each group having a separate sample of subjects. In many studies, however, each

group has the same subjects. Most commonly this happens when there is *repeated measurement* of the subjects over time or on several related response variables. The samples are then *dependent*, and the analysis must take this into account.

### Example 12.8 Positive and Negative Influences on Children

A recent General Social Survey asked subjects to respond to the following: “Children are exposed to many influences in their daily lives. What kind of influence does each of the following have on children? 1. Movies, 2. Programs on network television, 3. Rock music.” The possible responses were (very negative, negative, neutral, positive, very positive). Table 12.17 shows responses for 12 of the sampled subjects, using scores  $(-2, -1, 0, 1, 2)$  for the possible responses. This is part of a much larger data file for more than 1000 respondents. We analyze only this small sample here to help explain the concepts and so you can easily use your own software to try to replicate results.

□

Table 12.17: Opinions About Three Influences on Children. The scores represent  $-2 =$  very negative,  $-1 =$  negative,  $0 =$  neutral,  $1 =$  positive,  $2 =$  very positive.

Subject	Influence		
	Movies	TV	Rock
1	-1	0	-1
2	1	0	0
3	0	1	-2
4	2	0	1
5	0	-1	-1
6	-2	-2	-2
7	-1	-1	0
8	0	1	-1
9	-1	-1	-1
10	1	0	1
11	1	1	-1
12	-1	-1	-2
Mean	-0.08	-0.25	-0.75

## One-Way ANOVA with Repeated Measurement

For Table 12.17,  $H_0$  is the same as in ordinary one-way ANOVA: Equal population means for several groups. Is there much evidence that the population means differ for the three influences? Ordinary ANOVA is inappropriate because the three samples for the categories of influence are not independent. Each sample has the same subjects.

Suppose we regard the rows of Table 12.17, like the columns, as a factor. Then, the data layout resembles a two-way ANOVA. Each cell cross classifies a subject with an influence. A regression model could express the expected response as a function of 2 dummy variables for the 3 influences and 11 dummy variables for the 12 subjects. The test comparing population means for the three influences is then the main effect test for the column variable in the two-way ANOVA. In fact, this is the appropriate test for data of this sort.

Table 12.18 shows the ANOVA table for a two-way ANOVA. Consider

$H_0$ : Equal population means for the three influences.

The  $F$  test statistic is the mean square for influence divided by the mean square error, which is  $F = 1.44/0.57 = 2.55$ . The  $df$  values for the mean squares are  $df_1 = 2$  and  $df_2 = 22$ . The  $P$ -value equals  $P = 0.10$ . The evidence against  $H_0$  is not strong. But, with only 12 subjects, if  $H_0$  is false the power is probably low.

Table 12.18: ANOVA Table for Repeated Measures ANOVA of Opinion Response by Influence Type. This shows results of using software for two-way ANOVA, treating the subject as a second factor.

---

Source	Sum of Squares	df	Mean Square	F	Sig
Model	27.861	13	2.143	3.79	0.003
Error	12.444	22	0.566		
Total	40.306	35			

  

Source	Type III SS	df	Mean Square	F	Sig
INFLUENC	2.889	2	1.444	2.55	0.101
SUBJECT	24.972	11	2.270	4.01	0.003

---

Table 12.19 shows that we get similar results if we use specialized software for repeated measures, such as in SPSS by using the *Repeated Measures* option after selecting *General Linear Model* in the *Analyze* menu.

## The Sphericity Assumption

The traditional repeated measures ANOVA assumes *sphericity*. This is satisfied if the correlation is identical between responses for each pair of observations, a condition called *compound symmetry*. If this assumption is badly violated, the  $P$ -value

Table 12.19: Partial SPSS Output for Repeated Measures ANOVA of Opinion Response by Influence Type

Source	Test of Within-Subjects Effects				
	Type III Sum of Squares	df	Mean Square	F	Sig.
Sphericity assumed					
Influence	2.889	2	1.444	2.55	.101
Error	12.444	22	.566		

tends to be too small. Most software provides a formal significance test (Mauchly's test) of the sphericity assumption. When the data strongly contradict that assumption, an approximate test adjusts the degrees of freedom downward for the usual  $F$  test statistic, using an adjustment due to Greenhouse and Geisser. The technical details for these tests and adjustments are beyond the scope of this text, but standard software reports these results.

Using a repeated measurement design can improve precision of estimation. Having the same subjects in each group helps to eliminate extraneous sources of error. For instance, other variables that affect the response have the same values for each group, so differences between group means cannot reflect differences between groups on those variables. Controlling for possibly confounding factors by keeping them fixed in each row of the data file is referred to as **blocking**. For details on the linkage of analysis of variance procedures with experimental designs, see Howell (2006), Kirk (1982), and Winer et al. (1991).

### Confidence Intervals Comparing Dependent Samples

As usual, we learn more from estimating parameters. Table 12.17 showed the sample means for the three influences. Since the sample size is small, we weaken the multiple comparison confidence level a bit so that the intervals are not overly wide. The 90% Bonferroni confidence intervals use error probability  $0.10/3 = 0.0333$  for each interval. The error  $df = 22$ , and the  $t$ -score with probability  $0.0333/2 = 0.0167$  in each tail is  $t = 2.27$ . The square root of the mean square error equals  $s = \sqrt{0.566} = 0.75$ . Each group has 12 observations, so the margin of error for each confidence interval is

$$ts\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = 2.27(0.75)\sqrt{\frac{1}{12} + \frac{1}{12}} = 0.70.$$

For instance, the confidence interval for the difference between the mean on movies and the mean on rock music is  $(-0.08) - (-0.75) \pm 0.70$ , or  $(-0.03, 1.37)$ . It is plausible that the means are equal, but also plausible that the mean for movies is much more in the positive direction than the mean for rock music. Table 12.20 shows all three Bonferroni comparisons. It is also plausible that the mean for TV is much more in

the positive direction than the mean for rock music. Confidence intervals can convey useful information even if the overall test statistic is not significant.

Table 12.20: Bonferroni Multiple Comparison 90% Confidence Intervals for Comparing Mean Responses for Three Influences

Influences	Difference of Means	Confidence Interval
Movies, TV	0.17	(-0.53, 0.87)
Movies, Rock	0.67	(-0.03, 1.37)
TV, Rock	0.50	(-0.20, 1.20)

Suppose there are only two groups, with the same subjects in each sample. Section 7.4 showed that inference then uses the  $t$  distribution with difference scores between the two samples. For testing equality of means, the  $F$  statistic from ANOVA then simplifies to the square of the  $t$  statistic from that matched-pairs  $t$  test.

## Fixed Effects and Random Effects

The regression model for the previous analysis is

$$E(Y) = \alpha + \beta_1 m + \beta_2 t + \gamma_1 s_1 + \gamma_2 s_2 + \cdots + \gamma_{11} s_{11}.$$

Here,  $Y$  is the response on the five-category rating of an influence,  $m$  is a dummy variable for movies (i.e.,  $m = 1$  for a response on movies, 0 otherwise),  $t$  is a dummy variable for TV ( $t = 1$  for a response on TV, 0 otherwise), and  $m = t = 0$  for a response on rock music. Similarly,  $s_1$  is a dummy variable for subject 1, equaling 1 for that subject's three responses and 0 otherwise, and likewise for 10 other subject dummy variables. We use  $\gamma$  (gamma) instead of  $\beta$  for the coefficients of these terms for convenience, so the index of the parameter agrees with the index of the dummy variable. As usual, each factor has one fewer dummy variable than its number of categories.

A short-hand way of writing this regression model is

$$E(Y) = \alpha + \beta_j + \gamma_i.$$

Here,  $\beta_j$  denotes the effect for influence  $j$  and  $\gamma_i$  is the effect for subject  $i$ , where  $\beta_3 = 0$  and  $\gamma_{12} = 0$  for the final category of each variable. This equation expresses the expected response in the cell in row  $i$  and column  $j$  in terms of a row main effect and a column main effect. Testing equality of the population mean of  $Y$  for the three influences corresponds to testing  $H_0 : \beta_1 = \beta_2 = 0$ .

In this model, the main focus is comparing the influence parameters  $\{\beta_j\}$ , not the subject parameters  $\{\gamma_i\}$ . The subject parameters depend on which subjects are chosen

Figure 12.5: Box Plots for Weights of Anorexic Girls, by Treatment and Time of Measurement

((Fig. 12.5 in 3e))

for the sample. The subject effect is called a *random effect*, because the categories of that factor represent a random sample of all the possible ones. By contrast, the factor that defines the groups, influence type, is called a *fixed effect*. The analyses use *all* the categories of interest of a fixed effect, rather than a random sample of them. Models studied in earlier sections of this chapter contained only fixed effects.

When the classification variables are a mixture of random and fixed effects, such as in this example, the model is called a *mixed model*. For more complex mixed models than this one, the test statistics differ for some tests from their form when all classification variables are fixed effects. The next section discusses an important case.

## 12.7 Two-Way ANOVA with Repeated Measures on a Factor\*

Repeated measurement data sets often have more than one fixed effect. A second fixed effect may refer to groups to be compared on the repeated response. Most commonly those groups have independent samples. The following example illustrates:

### Example 12.9 Comparing Three Treatments for Anorexia

For 72 young girls suffering from anorexia, Table 12.21 shows their weights before and after an experimental period. The girls were randomly assigned to receive one of three therapies during this period. One group, a control group, received the standard therapy. The study analyzed whether one treatment is better than the others, the girls tending to gain more weight under that treatment.

Figure 12.5 shows box plots, graphically describing the response distributions before and after the experimental period for each treatment. Table 12.22 shows the summary sample means. The three treatments have similar distributions originally. This is not surprising, because subjects were randomly allocated to the three groups at that time. There is some evidence of a greater mean weight gain for the family therapy group, though there are a few low outlying weight values.

□

### Repeated Measures on One of Two Fixed Effects

Tables 12.21 and 12.22 have two fixed effects. One of them, ‘Treatment,’ has categories (CB = cognitive behavioral, FT = family therapy, C = control). It defines

12.7. TWO-WAY ANOVA WITH REPEATED MEASURES ON A FACTOR\*527

Table 12.21: Weights of Anorexic Girls, in Pounds, Before and After Receiving One of Three Treatments

Cognitive Behavioral		Family Therapy		Control	
Weight Before	Weight After	Weight Before	Weight After	Weight Before	Weight After
80.5	82.2	83.8	95.2	80.7	80.2
84.9	85.6	83.3	94.3	89.4	80.1
81.5	81.4	86.0	91.5	91.8	86.4
82.6	81.9	82.5	91.9	74.0	86.3
79.9	76.4	86.7	100.3	78.1	76.1
88.7	103.6	79.6	76.7	88.3	78.1
94.9	98.4	76.9	76.8	87.3	75.1
76.3	93.4	94.2	101.6	75.1	86.7
81.0	73.4	73.4	94.9	80.6	73.5
80.5	82.1	80.5	75.2	78.4	84.6
85.0	96.7	81.6	77.8	77.6	77.4
89.2	95.3	82.1	95.5	88.7	79.5
81.3	82.4	77.6	90.7	81.3	89.6
76.5	72.5	83.5	92.5	78.1	81.4
70.0	90.9	89.9	93.8	70.5	81.8
80.4	71.3	86.0	91.7	77.3	77.3
83.3	85.4	87.3	98.0	85.2	84.2
83.0	81.6			86.0	75.4
87.7	89.1			84.1	79.5
84.2	83.9			79.7	73.0
86.4	82.7			85.5	88.3
76.5	75.7			84.4	84.7
80.2	82.6			79.6	81.4
87.8	100.4			77.5	81.2
83.3	85.2			72.3	88.2
79.7	83.6			89.0	78.8
84.5	84.6				
80.8	96.2				
87.4	86.7				

Source: Thanks to Prof. Brian Everitt, Institute of Psychiatry, London, for these data.

Table 12.22: Sample Mean Weight, by Treatment and Time of Measurement, in Anorexia Study

Treatment	Time	
	Before	After
Cognitive Behavioral (CB)	82.7	85.7
Family Therapy (FT)	83.2	90.5
Control (C)	81.6	81.1

three groups of girls, represented by three independent samples. The second, ‘Time,’ consists of the two times for observations, (before, after). Each time has the same subjects, so the samples at its levels are dependent. Time is called a *within-subjects factor*, because comparisons of its categories use repeated measurements on subjects. Treatment is called a *between-subjects factor*, because comparisons of its categories use different subjects.

Although the two factors (treatment and time) are fixed effects, the analysis differs from ordinary two-way ANOVA. This is because the repeated measurements on the within-subjects factor (time) creates a third effect, a random effect for subjects. Each subject is measured at every category of time. Subjects are said to be *crossed* with the time factor. Each subject occurs at only one category of the between-subjects factor (treatment). Subjects are said to be *nested* within the treatment factor.

As in ordinary two-way ANOVA, we can test each main effect as well as their interaction. However, tests about the within-subjects factor (both its main effect and its interaction with the other fixed effect) use a different error term than the test about the between-subjects main effect. The ordinary sum of squared error term is partitioned into two parts. One part uses the variability between mean scores of subjects. It forms an error term for testing the between-subjects factor. The other part is based on how the pattern of within-subject scores varies among subjects. It forms an error term for any test involving the within-subjects factor.

Figure 12.6 shows the partitioning of the total sum of squares for a two-way ANOVA with repeated measures on one factor. Software automatically performs this partitioning and creates  $F$  statistics in the proper way for testing the main effects and interaction.

Figure 12.6: Partitioning of Variability in Two-Way ANOVA with Treatment and Time Factors and Repeated Measures on Time. Tests involving the within-subjects factor (time) use a separate error term.

((Fig 12.6 in 3e))

**Example 12.10 ANOVA  $F$  Tests Comparing Anorexia Treatments**

Table 12.23 shows a SPSS printout for the analysis of the anorexia study data. Each sum of squares summarizes the variation its terms explain in a corresponding regression model. This is the reduction in SSE when we add terms to the model for that effect. Since treatment has three categories, it has two dummy variables in the regression model, and its sum of squares has  $df = 2$ . Since time has two levels, it has one dummy variable, and its sum of squares has  $df = 1$ . The interaction between these effects has two terms in the model, based on the cross-product of the two dummy variables for treatment with the dummy variable for time, so its  $df = 2$ .

Table 12.23: Printout for Two-Way Analysis of Variance of Table 12.21 with Treatment and Time Fixed Effects and Repeated Measures on Time

---

Tests of Within-Subject Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig
TIME	366.04	1	366.04	12.92	0.001
TIME*TREATMENT	307.32	2	153.66	5.42	0.006
Error(TIME)	1955.37	69	28.34		

  

Tests of Between-Subjects Effects.					
Source	Type III Sum of Squares	df	Mean Square	F	Sig
TREATMENT	644.23	2	322.12	6.20	0.003
Error	3584.03	69	51.94		

---

The error term for the between-subjects part of the table uses the variability among subjects' means within each group. Its  $df = 69$ , based on 28 dummy variables for the 29 subjects receiving CB therapy, 16 dummy variables for the 17 subjects receiving FT, and 25 dummy variables for the 26 subjects in group C ( $28 + 16 + 25 = 69$ ). The remaining variability, not accounted for by this error term or by the main effects and interaction terms, is the error sum of squares for testing the within-subjects effects.

The 'TIME\*TREATMENT' row of the ANOVA table indicates that the interaction is highly significant. The  $P$ -value = 0.006. The difference between population means for the two times differs according to the treatment, and the difference between population means for a pair of treatments varies according to the time. Because of the significant interaction, we do not test the main effects. We need, instead, to use confidence intervals to describe the interaction.

□

### Follow-up Confidence Intervals

Table 12.22 showed the sample means for the six combinations of the two factors. The evidence of interaction is clear. The sample means for the three treatments are similar at the initial time. At the second time, by contrast, the mean for the control group is similar to its mean at the initial time, but the mean is somewhat larger for the other two treatments than their initial means, particularly for the FT group.

To construct confidence intervals comparing means at the two times, for each treatment, the appropriate common standard deviation estimate is the square root of the mean square error from the within-subjects analysis. From Table 12.23, this equals  $\sqrt{28.34} = 5.3$ , with  $df = 69$ . We illustrate by constructing a 95% confidence interval comparing the two means for family therapy (FT), which 17 girls received at each time. The  $t$ -score for 95% confidence when  $df = 69$  equals 1.99. The confidence interval has margin of error equal to this  $t$ -score times the root mean square error times the square root factor involving the inverse of each sample size (17 for each time). This interval equals

$$(90.5 - 83.2) \pm 1.99(5.3)\sqrt{\frac{1}{17} + \frac{1}{17}}, \text{ or } 7.3 \pm 3.6, \text{ or } (3.6, 10.9).$$

For the FT therapy, we conclude that the population mean weight is between 3.6 and 10.9 pounds higher following the treatment period. Similarly, a 95% confidence interval comparing the two means equals (0.2, 5.8) for the CB therapy and (-3.4, 2.5) for the control group. There is evidence of an increase, albeit a small one, for the CB therapy, but no evidence of change for the control group.

To make between-subjects comparisons of treatments, for each time, one cannot use the root mean square error from the between-subjects analysis. The reason is that these separate comparisons involve both the treatment main effect and the interaction, and these two sources of variation have different error terms in the repeated measures ANOVA. At a particular time, however, the subjects in the three treatments are independent samples. Thus, we can compare three means at a given time using a one-way ANOVA  $F$  test or using confidence intervals for those data alone.

For instance, for the 72 observations at time = after, the  $F$  test statistic for the one-way ANOVA comparing the three means is 8.6, with  $df_1 = 2$  and  $df_2 = 69$ . The  $P$ -value is 0.0004, very strong evidence of a difference among the treatment means. For this one-way ANOVA, the square root of MSE equals  $s = 7.3$ . The 95% confidence interval for the difference of means between the FT and the CB treatments, based on the 17 + 29 observations for the two groups, equals

$$(90.5 - 85.7) \pm 1.99(7.3)\sqrt{\frac{1}{17} + \frac{1}{29}}, \text{ or } 4.8 \pm 4.4, \text{ or } (0.4, 9.2).$$

We conclude that, at the follow-up time, the mean weight is between 0.4 and 9.2 pounds higher with the FT treatment than with the CB treatment. The true means

may be essentially equal, but if they differ, the advantage could be quite noticeable for the family therapy. Table 12.24 shows the confidence intervals for each pair of treatments.

Table 12.24: 95% Confidence Intervals Comparing Treatment Means After the Treatment Period

Treatments Compared	Difference of Sample Means	Confidence Interval	Bonferroni Interval
FT - CB	4.8	(0.4, 9.2)	(-0.7, 10.3)
FT - C	9.4	(4.9, 13.9)	(3.8, 15.0)
CB - C	4.6	(0.7, 8.5)	(-0.2, 9.4)

In summary, there is evidence that the mean weight increases during the experimental period for both noncontrol treatments. There is marginal evidence that the mean is higher after the experiment for the FT treatment than for the CB treatment.

At this stage, further interest may relate to whether the change in means, between time = after and time = before, differed for the two noncontrol treatments. That is, do the difference scores for the FT treatment have a significantly higher mean than the difference scores for the CB treatment? The difference scores have a mean of  $(90.5 - 83.2) = 7.3$  for the FT treatment and  $(85.7 - 82.7) = 3.0$  for the CB treatment, and we used these as the basis of separate confidence intervals for the mean change, above. Since the two groups are independent samples, the variance of the difference of these means is the sum of the variances. Thus, a 95% confidence interval for the difference between the mean changes in weight is

$$(7.3 - 3.0) \pm 1.99(5.3) \sqrt{\frac{1}{17} + \frac{1}{17} + \frac{1}{29} + \frac{1}{29}}, \text{ or } 4.3 \pm 4.6, \text{ or } (-0.3, 8.8).$$

Although the mean change could be considerably larger for the FT treatment, it is also plausible that the mean changes could be identical.

### Bonferroni Multiple Comparisons of Treatments

As usual, we can control the overall error rate for several comparisons using the Bonferroni multiple comparison method. Suppose we use three confidence intervals to compare treatments at time = after, and three intervals to compare times within the treatments. To ensure at least 90% confidence for the entire set, since  $0.10/6 = 0.0167$ , we use a 98.33% confidence interval for each individual comparison.

Such intervals are wider than the ones just reported, since they use a  $t$ -score of 2.45 instead of 1.99. Table 12.24 shows them for the pairwise comparisons of treatments at time = after. With this more conservative approach, only the difference between the FT and C treatments is significant, with the interval not containing 0.

### More Complex Repeated Measures Analyses

Two-way ANOVA with repeated measures on a factor extends to more complex designs. Suppose, for example, that a study has three factors, A, B, and C, with repeated measures on C. That is, subjects are crossed with C but nested within combinations of levels of A and B. The between-subjects effects, namely, the A and B main effects and the A×B interaction, are tested with a mean square error based on variability between subjects. All effects involving the within-subjects factor C, namely the C main effect, the A×C interaction, the B×C interaction, and the A×B×C interaction, are tested with a separate mean square error.

In some studies with two fixed effects, repeated measures occur on both factors. For instance, we may observe the same subjects for each treatment at each of several times. Then, subjects (a random effect) are crossed with both factors (fixed effects), and an observation occurs for every subject at every combination of factor levels. As in ordinary two-way ANOVA, the effects of interest refer to the fixed effects—their main effects and interaction. The complicating factor is that each test requires a separate mean square error, but software can easily conduct this analysis.

### Repeated Measures at More than Two Times

In Example 12.8 the repeated measurements occur at two times. When observations are made at several times, the repeated measures ANOVA is more complex. In particular, the results depend on assumptions about the correlation structure of the repeated measurements. The standard test for the within-subject effect assumes *sphericity*, as in 1-way repeated measures ANOVA. Tests of the between-subjects effects are not affected by violation of the sphericity assumption, so no adjustment is needed for that  $F$  test.

Alternative *multivariate* ANOVA approaches to testing the within-subjects effects (such as the *Wilks lambda likelihood-ratio test*) make fewer assumptions but often have weaker power. Other recently developed methods allow more varied types of random effects modeling of the correlation structure for the repeated responses, providing options other than sphericity. We'll discuss such approaches in Section 16.1.

## 12.8 Effects of Violations of ANOVA Assumptions

Each ANOVA method presented in this chapter assumes (besides randomization) that the groups have population distributions that are normal with identical standard deviations. These are stringent assumptions. As in ordinary regression assumptions, they are never exactly satisfied in practice.

### Robustness of $F$ Tests

Moderate departures from normality of the population distributions can be tolerated. The  $F$  distribution still provides a good approximation to the actual sampling distri-

bution of a  $F$  test statistic. This is particularly true for larger sample sizes, since the sampling distributions then have weaker dependence on the shape of the population distribution. Moderate departures from equal standard deviations can also be tolerated. When the sample sizes are identical for the groups, the  $F$  test is very robust to violations of this assumption.

Constructing histograms for each sample data distribution helps to check for extreme deviations from these assumptions. Misleading results may occur in the  $F$  tests if the population distributions are highly skewed and the sample size is small, or if there are relatively large differences among the population standard deviations (say, the largest sample standard deviation is several times as large as the smallest one) and the sample sizes are unequal. Section 14.5 discusses alternative regression approaches for such gross violations. When the distributions are very highly skewed, the mean may not even be an appropriate summary measure.

Confidence intervals, like tests, are not highly dependent on the normality assumption. When the standard deviations are quite different, with the ratio of the largest to smallest exceeding about 2, it is preferable to use formulas for intervals based on separate standard deviations for the groups rather than a single pooled value. For instance, the first confidence interval method presented in Section 7.3 does not assume equal standard deviations.

As in other inferences, the quality of the sample is most crucial. In one-way ANOVA, for instance, conclusions may be invalid if the observations in the separate groups compared are not independent random samples. ANOVA procedures are not robust to violations of sampling assumptions.

Let's consider the validity of the assumptions for two-way ANOVA for the data in Table 12.10 on political ideology classified by party ID and race. The sample standard deviations are similar for the four groups (1.26, 1.43, 1.24, 1.30). Also, the sample sizes are large (215, 125, 169, and 144), so the normality assumption is not crucial. The full GSS sample was randomly obtained, so we may regard the four samples as independent random samples. ANOVA is suitable for these data.

## The Kruskal-Wallis Test: A Nonparametric Approach

The *Kruskal-Wallis test* is an alternative to one-way ANOVA for comparing several groups. It is a nonparametric method, not requiring the normality assumption. The test statistic uses only the ordinal information in the data. It ranks the observations and compares mean ranks for the various groups. The test statistic is larger when the differences among the mean ranks are larger. It has an approximate chi-squared distribution with  $df = g - 1$ .

This test is especially useful for small samples in which the effects of severe departures from normality may be influential. We shall not present the test statistic here. Its result is similar to that of a chi-squared test for the effect of a categorical predictor in a model for an ordinal response presented in Section 15.4. In practice, it is more informative to use a modeling approach, because the model parameter estimates give us information about the sizes of effects. In addition, the modeling strategy adapts better to multivariate analyses.

Nonparametric tests also exist for more complex analyses. For instance, *Friedman's test* is an alternative to the  $F$  test of Section 12.6 comparing groups when the same subjects occur in each. An advantage of the parametric methods is that they more easily generalize to multivariate modeling and to estimation of effects, which are more important than significance testing. Lehmann (1975) is a good source for details about nonparametric methods.

## 12.9 Chapter Summary

This chapter presented *analysis of variance* (ANOVA) methods for comparing several groups according to their means on a quantitative response variable.

- *One-way ANOVA* methods compare means for categories of a single explanatory variable.
- *Two-way ANOVA* methods compare means across categories of each of two explanatory variables. Assuming no interaction, the main effects describe the effect of each predictor while controlling for the other one.
- *Multiple comparison* methods provide confidence intervals for the difference between each pair of means, while controlling the overall error probability. The *Bonferroni* method does this using an error probability for each comparison that equals the desired overall error probability divided by the number of comparisons.
- Analysis of variance methods are special cases of multiple regression analyses. *Dummy variables* in the regression model represent categories that define the groups. Each dummy variable equals 1 for a particular category and 0 otherwise.

Ordinary ANOVA methods compare groups with *independent* random samples from the groups. For some studies, different samples instead have the same subjects, such as when an experiment observes subjects repeatedly over time. Methods for *repeated measures ANOVA* result from regression models with *random effects* that represent the effects of the random sample of observed subjects. Such methods treat *within-subjects* effects (for repeated measurements on subjects) differently from *between-subjects* effects (for independent samples of subjects).

Chapters 9 and 11 presented regression models for a quantitative response variable when the explanatory variables are also *quantitative*. This chapter has modeled a quantitative response variable as a function of *categorical* explanatory variables. Table 12.25 summarizes the statistical tests discussed in Sections 12.1, 12.4, and 12.6. Models of the next chapter include both quantitative and categorical explanatory variables.

## PROBLEMS

Table 12.25: ANOVA Tests for Comparing Several Groups on a Response Variable

Element of Test	One-Way ANOVA	Two-Way ANOVA	Repeated Measures ANOVA
1. Samples	Independent	Independent	Dependent
2. Hypotheses	$H_0$ : Identical means $H_a$ : At least two means not equal	$H_0$ : Identical row means $H_0$ : Identical col. means $H_0$ : No Interaction	$H_0$ : Identical means $H_a$ : At least two means not equal
3. Test stat.	$F = \frac{\text{Between MS}}{\text{Within MS}}$ $F$ distribution $df_1 = g - 1$ $df_2 = N - g$	$F = \frac{\text{Effect MS}}{\text{MS error}}$ $F$ distribution $df_1 = df$ for effect $df_2 = df$ for error	$F = \frac{\text{Effect MS}}{\text{MS error}}$ $F$ distribution $df_1 = df$ for effect $df_2 = df$ for error
4. $P$ -value	Right-tail prob.	Right-tail prob.	Right-tail prob.

### Practicing the Basics

- A General Social Survey asked subjects how many good friends they have. Is this associated with the respondent's astrological sign (the 12 symbols of the Zodiac)? The ANOVA table for the GSS data reports  $F = 0.61$  based on  $df_1 = 11$ ,  $df_2 = 813$ .

  - Introduce notation, and specify the null hypothesis and alternative hypothesis for the ANOVA.
  - Based on what you know about the  $F$  distribution, would you guess that the test statistic value of 0.61 provides strong evidence against the null hypothesis? Explain.
  - Software reports a  $P$ -value of 0.82. Explain how to interpret it.
- For GSS data comparing the reported number of good friends for those who are (married, widowed, divorced, separated, never married), an ANOVA table reports  $F = 0.80$ .

  - Introduce notation, and specify the null hypothesis and the alternative hypothesis for the ANOVA  $F$  test.
  - Based on what you know about the  $F$  distribution, would you guess that the test statistic value of 0.80 provides strong evidence against the null hypothesis? Explain.
  - Software reports a  $P$ -value of 0.53. Explain how to interpret it.
- A recent General Social Survey asked subjects "What is the ideal number of kids for a family?" Do responses tend to depend on the subjects' religious affiliation? Results of an ANOVA are shown in Table 12.26, for religion categories (Protestant, Catholic, Jewish, Other or none).

  - Define notation and specify the null hypothesis tested in this table.
  - Report the  $F$  test statistic value and the  $P$ -value for this test. Interpret the  $P$ -value.

- c) Based on (b), can you conclude that *every* pair of religious affiliations has different population means for ideal family size? Explain.

Table 12.26:

Source	SS	df	Mean Square	F	Sig
Religion	11.72	3	3.91	5.48	0.001
Error	922.82	1295	0.71		
Total	934.54	1298			

4. A recent GSS asked, “How often do you go to a bar or tavern?” Table 12.27 shows results of ANOVA for comparing the mean reported number of good friends at three levels of this variable.
- a) State the (i) hypotheses, (ii) test statistic value, (iii)  $P$ -value for the significance test displayed in this table. Interpret the  $P$ -value.
- b) Based on the assumptions for the method in (a), is there any aspect of the data summarized here that suggests that the ANOVA test may not be appropriate? Explain.

Table 12.27:

	How often go to bar or tavern?		
	Very often	Occasional	Never
Mean no. good friends	12.1	6.4	6.2
Standard deviation	21.3	10.2	14.0
Sample size	41	166	215

  

Source	Sum of Squares	df	Mean Square	F	Sig
Group	1116.8	2	558.4	3.03	0.049
Error	77171.8	419	184.2		
Total	78288.5	421			

5. Table 12.28 shows scores on the first quiz (maximum score 10 points) in a beginning French course. Students in the course are grouped as follows:
- Group A: Never studied foreign language before, but have good English skills
- Group B: Never studied foreign language before; have poor English skills
- Group C: Studied other foreign language

- a) Table 12.29 provides results of an ANOVA. Report the assumptions, hypotheses, test statistic, and  $P$ -value, and interpret.
- b) The sample means are quite different, but the  $P$ -value is not small. Name one important reason for this. (Hint: For given sample means, how do the results of the test depend on the sample sizes?)
- c) Use Bonferroni 85% multiple comparisons to construct confidence intervals comparing pairs of means. Interpret each interval. Indicate which means, if any, that are significantly different.

Table 12.28:

Group A	Group B	Group C
4	1	9
6	5	10
8		5

Table 12.29:

---

Source	Sum of Squares	df	Mean Square	F	Sig
Between Groups	30.000	2	15.000	2.5000	0.1768
Within Groups	30.000	5	6.00		
Total	60.00	7			

  

Group	n	Mean	Standard Deviation	Standard Error
Grp 1	3	6.0000	2.0000	1.1547
Grp 2	2	3.0000	2.8284	2.0000
Grp 3	3	8.0000	2.6458	1.5275

---

- 6. Use software to reproduce the ANOVA results reported in the previous exercise.
- 7. A consumer protection group compares three types of front bumpers for a brand of automobile. A test is conducted by driving an automobile into a brick wall at 15 miles per hour. The response is the amount of damage to the car, as measured by the repair costs, in hundreds of dollars. Due to the potentially large costs, the study conducts only two tests with each bumper type. Table 12.30 shows the results.
  - a) Report the sample means for the three bumpers.
  - b) Find the within-groups sum of squares and the associated variance estimate.
  - c) Find the between-groups sum of squares and its associated variance estimate.

- d) Test the hypothesis that the mean repair costs are the same for the three types of bumpers. Report the null and alternative hypotheses, test statistic,  $df$  values,  $P$ -value, and interpret.
- e) Construct an ANOVA table for displaying the results of this analysis.

Table 12.30:

Bumper A	Bumper B	Bumper C
1	2	11
3	4	15

8. Refer to the previous exercise.
- a) Construct a 95% confidence interval for the difference between the population means for bumpers A and B. Interpret.
- b) Construct 95% multiple comparison confidence intervals for the differences in mean repair costs for each pair of bumpers. Interpret the results, and indicate which types of bumpers are judged to be different in mean repair cost.
9. Refer to the previous two exercises.
- a) Suppose that the first observation in the second group was actually 9, not 1. Then, the standard deviations are the same as reported in the table, but the sample means are 6, 7, and 8 rather than 6, 3, and 8. Do you think the  $F$  test statistic would be larger, the same, or smaller? Explain your reasoning, without doing any calculations.
- b) Suppose you had the same means as shown in the table but the sample standard deviations were 1.0, 1.8, and 1.6, instead of 2.0, 2.8, and 2.6. Do you think the  $F$  test statistic would be larger, the same, or smaller? Explain your reasoning.
- c) Suppose you had the same means and standard deviations as shown in the table but the sample sizes were 30, 20, and 30, instead of 3, 2, and 3. Do you think the  $F$  test statistic would be larger, the same, or smaller? Explain your reasoning.
- d) In (a), (b), and (c), would the  $P$ -value be larger, the same, or smaller? Why?
10. In a study to compare customer satisfaction at service centers for PC technical support in San Jose, California, Toronto, Canada, and Bangalore, India, each center randomly sampled 100 people who called during a two-week period. Callers rated their satisfaction on a scale of 0 to 10, with higher scores representing greater satisfaction. The sample means were 7.6 for San Jose, 7.8 for Toronto, and 7.1 for Bangalore. Table 12.31 shows the results of conducting an ANOVA.
- a) Define notation and specify the null hypothesis tested in this table.
- b) Explain how to obtain the  $F$  test statistic value reported in the table from the mean square values shown, and report the values of  $df_1$  and  $df_2$  for the  $F$

distribution.

c) Interpret the  $P$ -value reported for this test. What conclusion would you make using a 0.05 significance level?

Table 12.31:

Source	Sum of Squares	df	Mean Square	F	Sig
Group	26.00	2	13.00	27.6	0.000
Error	140.00	297	0.47		
Total	60.00	299			

11. Refer to the previous exercise.
  - a) Explain why the margin of error for separate 95% confidence intervals is the same for comparing the population means for each pair of cities. Show that this margin of error is 0.19.
  - b) The margin of error for Bonferroni or for Tukey 95% multiple comparison confidence intervals is 0.23. Why is it different than in (a)? What is an advantage of using the multiple comparison intervals? c) Construct and interpret the multiple comparison confidence intervals.
  - c) With dummy variables to represent the three service centers, the prediction equation is  $\hat{y} = 7.1 + 0.5z_1 + 0.7z_2$ . Show how the terms in this equation relate to the sample means of 7.6, 7.8, and 7.1.
  
12. The 2000 General Social Survey asked 1829 subjects how many hours per day they watched TV, on the average. The sample means were 2.75 for whites ( $n = 1435$ ), 4.14 for blacks ( $n = 305$ ), and 2.61 for other ( $n = 89$ ). In a one-way ANOVA, the between-groups estimate of the variance is 249.93 and the within-groups estimate is 6.41.
  - a) Conduct the ANOVA test and make a decision using 0.05 significance level.
  - b) The 95% confidence interval comparing the population means is (1.1, 1.7) for blacks and whites, (-0.4, 0.7) for whites and the other category, and (0.9, 2.1) for blacks and the other category. Based on the three confidence intervals, indicate which pairs of means are significantly different. Interpret.
  - c) Based on the information given, show how to construct the confidence interval that compares the population mean TV watching for blacks and whites.
  
13. The General Social Survey asks “Would you say that you are very happy, pretty happy, or not too happy?” (variable HAPPY). A recent GSS also asked, “About how many good friends do you have?” (variable NUMFRIEND). Table 12.32 summarizes results.

Table 12.32:

	Very happy	Pretty happy	Not too happy
Mean	10.4	7.4	8.3
Standard deviation	17.8	13.6	15.6
Sample size	276	468	87

  

Source	Sum of Squares	df	Mean Square	F	Sig
Group	1626.8	2	813.4	3.47	0.032
Error	193900.9	828	234.2		
Total	195527.7	830			

- a) Interpret the result of the  $F$  test. Why is this insufficient to tell you how the number of friends depends on happiness?
- b) The 95% confidence interval comparing means for the very happy and pretty happy categories is (0.7, 5.3). Interpret.
- c) For the Bonferroni method with overall confidence level 95%, the interval comparing means for the very happy and pretty happy categories is (0.2, 5.8). Why is it wider than the interval in (b)?
- d) Software reports Tukey 95% confidence intervals of (0.3, 5.7) comparing very happy and pretty happy, (-2.3, 6.5) comparing very happy and not too happy, and (-5.1, 3.3) comparing pretty happy and not too happy. Interpret, and indicate which groups are significantly different.
14. Refer to Table 12.21.
- a) Using software with the data at the text website, conduct a one-way ANOVA for the 72 observations at time = after. Verify that the  $F$  test statistic equals 8.65, with  $df_1 = 2$  and  $df_2 = 69$ , for which  $P = 0.0004$ . Interpret.
- b) Show how to obtain the same results by conducting a regression analysis.
15. A geographer compares residential lot sizes in four quadrants of a city. To do this, he randomly samples 300 records from a city file on home residences and records the lot sizes (in thousands of square feet) by quadrant. The ANOVA table (Table 12.33) refers to a comparison of mean lot sizes for the northeast (NE), northwest (NW), southwest (SW), and southeast (SE) quadrants of the city.
- a) Find the mean square error, and interpret its square root.
- b) Find the  $F$  test statistic, and make a conclusion using  $\alpha$ -level = 0.05.
- c) The sample mean lot sizes for the NE, NW, SW, and SE quadrants are 8, 15, 11, and 9, with  $n_1 = 100$ ,  $n_2 = 100$ ,  $n_3 = 50$ ,  $n_4 = 50$ . Illustrate multiple comparison 94% confidence intervals by constructing the Bonferroni interval

comparing the NE and NW quadrants.

**d)** Would *separate* 94% confidence intervals for the differences between each pair of means be wider, or narrower, than the multiple comparison intervals? Explain.

Table 12.33:

	Sum of Squares	df	Mean Square	$F$	Prob > $F$
Quadrant	2700	3	—	—	—
Error	1480	296	—		
Total	4180	299			

16. For  $g$  groups with large sample sizes, we plan to compare all pairs of population means. We want the probability to equal at least 0.80 that the entire set of confidence intervals contain the true differences.
- If  $g = 10$ , for the Bonferroni method, which tabled  $t$ -score should we use for each interval? (Assume the sample size is large enough that the  $t$ -score is well approximated by a  $z$ -score.)
  - Which  $t$ -score should we use if  $g = 5$ ?
  - Describe how the  $t$ -score depends on the number of groups, and explain the implication regarding width of the intervals.
17. A psychologist compares the mean amount of time of REM sleep for subjects under three conditions. She uses three groups of subjects, with four subjects in each group. Table 12.34 shows a SAS printout for the analysis.
- Interpret the  $P$ -value reported for the  $F$  test.
  - Explain how to obtain the margin of error of 13.96 for the 95% Bonferroni confidence intervals.
  - According to the Bonferroni method, can you conclude that any pair of the population means differ? Explain.
18. Refer to the previous exercise.
- Set up dummy variables for a regression model so that an  $F$  test for the regression parameters is equivalent to the ANOVA test. Express the null hypothesis both in terms of population means and regression parameters.
  - Report the prediction equation obtained in fitting this regression equation to the REM sleep data.
19. A study compares the mean level of contributions to political campaigns in Pennsylvania by registered Democrats, registered Republicans, and unaffiliated voters.
- Write a regression equation for this analysis, and interpret the parameters in the model.

Table 12.34:

---

Dependent Variable: TIME

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	72.000	36.000	0.79	0.4813
Error	9	408.000	45.333		
Total	11	480.000			

Bonferroni T tests for variable: TIME

Alpha= 0.05 df= 9 MSE= 45.33333

Critical Value of T= 2.93

Minimum Significant Difference= 13.965

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	GROUP
A	18.000	4	3
A			
A	15.000	4	2
A			
A	12.000	4	1

---

- b) Explain how one can use the regression model to test the null hypothesis of equal mean contributions for the three groups.
20. Exercise 7 showed an ANOVA for an experiment comparing three bumpers, with sample mean damages of 2 for Bumper A, 3 for Bumper B, and 13 for Bumper C (in hundreds of dollars).
- Set up the regression model with dummy variables for the ANOVA.
  - Show the correspondence between the null hypothesis for the means in ANOVA and the null hypothesis for the regression parameters.
  - Report the prediction equation you would obtain for the model in (a).
21. For the “house selling price” data file at the text website, whether a house is new is a dummy variable. Using software, put this as the sole predictor of selling price in a regression analysis.
- Conduct the  $t$  test for the effect of this variable in the regression analysis (i.e., test  $H_0: \beta = 0$ ). Interpret.
  - Conduct the  $F$  test for the analysis of variance comparing the mean selling prices of new and existing homes.
  - Explain the connection between the value of  $t$  in (a) and the value of  $F$  in (b).
22. The 25 women faculty in the humanities division of a college have a mean salary of \$66,000, whereas the five in the science division have a mean salary of \$80,000. On the other hand, the 20 men in the humanities division have a mean salary of \$65,000, and the 30 men in the science division have a mean salary of \$79,000.
- Construct a table of sample mean incomes for the  $2 \times 2$  cross-classification of gender and division of college. Find the overall means for men and women. Interpret.
  - Discuss how the results of a one-way comparison of mean incomes by gender would differ from the results of a two-way comparison of mean incomes by gender, controlling for division of college. (Note: This reversal of which gender has the higher mean salary, according to whether one controls division of college, illustrates *Simpson's paradox*. See Exercise 16 in Chapter 10.)
23. When we use the 2000 GSS to evaluate how the mean number of hours a day watching TV depends on gender and race, we get the results shown in the ANOVA table.
- Is there a significant (i) gender effect, (ii) race effect? Explain.
  - The sample means were 2.71 for white females, 2.79 for white males, 4.13 for black females, and 4.16 for black males. Explain how these results are compatible with the results of the tests discussed in (a).
24. The GSS asks, “What is the ideal number of kids for a family?” Table 12.36 shows results of using a recent GSS to evaluate the effects of gender and race.
- Explain how to interpret the results of the  $F$  tests for the gender and race main effects.

Table 12.35:

Source	SS	df	MS	F	Sig
Gender	2.22	1	2.22	0.35	0.555
Race	489.65	1	489.65	76.62	0.000
Error	11094.16	1737	6.39		
Total	11583.81	1739			

b) Let  $s = 1$  for females and 0 for males, and let  $r = 1$  for blacks and 0 for whites. The no interaction model has  $\hat{y} = 2.42 + 0.04s + 0.37r$ . Interpret the coefficient of  $s$ . What is the practical implication of this estimate being so close to 0?

c) Use the prediction equation in (b) to find the estimated mean for each of the four combinations of gender and race. Explain how these means satisfy ‘no interaction.’

Table 12.36:

Source	SS	df	MS	F	P-value
Gender	0.25	1	0.25	0.36	0.550
Race	16.98	1	16.98	24.36	0.000
Error	868.67	1246	0.70		
Total	886.12	1248			

25. In 2000 in the U.S., the population mean hourly wage for males was \$22 for white-collar jobs, \$11 for service jobs, and \$14 for blue-collar jobs. For females the means were \$15 for white-collar jobs, \$8 for service jobs, and \$10 for blue-collar jobs.

a) Identify the response variable and the two factors.

b) Show these means in a two-way classification of the two factors.

c) Compare the differences between males and females for (i) white-collar jobs (ii) blue-collar jobs. Explain why there is interaction, and describe it.

26. In 2000 in the U.S., the U.S. Census Bureau<sup>1</sup> stated the population median income to be \$29,661 for white females, \$25,736 for black females, \$40,350 for white males, \$30,886 for black males.

<sup>1</sup>[www.census.gov/hhes/www/income00.html](http://www.census.gov/hhes/www/income00.html)

- a) Identify the response variable and the two factors, and show these medians in a two-way classification of the factors.
- b) Compare the differences between males and females for (i) white subjects (ii) black subjects. Explain why there is interaction in terms of the median summary of center. Describe its nature.
- c) Show a set of four population median incomes that would satisfy  $H_0$ : no interaction.
27. Table 12.37 shows results of an ANOVA on  $Y =$  depression index and the predictors gender and marital status (married, never married, divorced). State the sample size and fill in the blanks in the ANOVA table. Interpret results.

Table 12.37:

Source	Sum of Squares	df	Mean Square	$F$	Sig
Gender	100	—	—	—	—
Marital status	200	—	—	—	—
Interaction	100	—	—	—	—
Error	—	—	—	—	—
Total	4000	205	—	—	—

28. Table 12.15 gave the prediction equation  $\hat{y} = 4.58 - 0.71p_1 - 0.54p_2 - 0.08s$  for a model relating political ideology to party and to sex. Find the estimated means for each of the six cells, and show that they satisfy a lack of interaction.
29. A regression analysis of college faculty salaries (M. Bellas, *American Sociological Review*, Vol. 59, 1994, p. 807) included several predictors, including a dummy variable for gender (male = 1) and a dummy variable for race (nonwhite = 1). For annual income measured in thousands of dollars, the estimated coefficients were 0.76 for gender and 0.62 for race.
- a) Interpret these coefficients.
- b) At particular settings of the other predictors, the estimated mean salary for white females was 30.2 thousand. Find the estimated means for the other three groups.
30. When we use the 2002 GSS and regress  $Y =$  number of hours per day watching TV on  $s =$  sex (1 = male, 0 = female) and religious affiliation ( $r_1 = 1$  for Protestant,  $r_2 = 1$  for Catholic,  $r_3 = 1$  for Jewish,  $r_1 = r_2 = r_3 = 0$  for none or other), we get  $\hat{y} = 2.4 + 0.2s + 0.5r_1 + 0.8r_2 - 0.1r_3$ .
- a) Interpret the sex effect.
- b) Interpret the coefficient of  $r_1$ .
- c) State a corresponding model for the population, and indicate which parameters would need to equal zero for the response variable to be independent of religious affiliation, for each sex.

31. Table 12.38 summarizes responses on political ideology in the 2000 General Social Survey by religion and gender. The  $P$ -value is 0.03 for testing  $H_0$ : no interaction. Explain what this means in the context of this example, and indicate one place in the table that may be responsible for the small  $P$ -value.

Table 12.38:

---

Religion		Political Ideology		Political Ideology		
		Mean	Std Dev.	Mean	Std. Dev.	
Protestant	Female	4.23	1.34	Male	4.44	1.41
Catholic	Female	4.10	1.33	Male	4.22	1.42
Jewish	Female	2.59	1.22	Male	3.82	1.68
None	Female	3.34	1.34	Male	3.58	1.42

---

32. Use software with Table 12.10, analyzed in Section 12.5.
- Fit the no interaction model, and verify the results given there.
  - Fit the interaction model. Compare SSE for this model to SSE for the no interaction model. Show how the difference between these values relates to the numerator sum of squares for testing  $H_0$ : no interaction.
  - Using the prediction equation for the interaction model, find the six estimated cell means, and compare them to the sample means. (Note: The model uses six parameters to summarize six means, so it has a perfect fit.)
33. Consider the regression model  $E(Y) = \alpha + \beta_1 s + \beta_2 r + \beta_3(s \times r)$ , where  $Y$  = annual income (thousands of dollars),  $s$  = sex ( $s = 1$  for men,  $s = 0$  for women), and  $r$  = race ( $r = 1$  for whites,  $r = 0$  for blacks).
- Suppose that, in the population,  $\beta_3 = 0$ . Interpret  $\beta_1$ .
  - The prediction equation for a certain sample is  $\hat{y} = 16 + 2s + 3r + 8(s \times r)$ . By finding the four predicted means for this equation, show that the coefficient 8 of the interaction term is the amount by which the mean for one of the four groups must increase or decrease for the interaction to disappear.
34. For the 2000 GSS, Table 12.38 shows sample means of political ideology classified by gender and by race. For  $H_0$ : no interaction, software reports  $F = 21.7$ ,  $df_1 = 1$  and  $df_2 = 2508$ , and  $P$ -value = 0.001. So, in comparing females and males on mean political ideology, we should do it separately by race.
- Compare the sample means of females and males for each race descriptively.
  - Explain how to obtain the following interpretation from the sample means: "For females there is no race effect on political ideology. For males, whites are more conservative by about half an ideology category, on the average."
  - Suppose that instead of the two-way ANOVA, you performed a one-way ANOVA with gender as the predictor and a separate one-way ANOVA with

race as the predictor. Suppose the ANOVA for gender does not show a significant effect. Explain how this could happen, even though the two-way ANOVA showed a gender effect for each race. (Hint: Will the overall sample means for females and males be more similar than they are for each race?)

**d)** Refer to (c). Summarize what you would learn about the gender effect from a two-way ANOVA that you would fail to learn from a one-way ANOVA.

Table 12.39:

Gender	Race	
	Black	White
Female	4.06 (n = 256)	4.04 (n = 1144)
Male	3.74 (n = 139)	4.25 (n = 973)

35. Refer to Table 12.17 about three influences on children.
- Using software, conduct the repeated measures analyses of Section 12.6.
  - Suppose you scored the influence categories (1, 2, 3, 4, 5). Does this affect the test statistic? Explain.
  - Suppose you used scores  $(-3, -2, 0, 2, 3)$ . What would this assume about the response categories? Repeat the analyses using these scores. Are the conclusions sensitive to the choice of scores?
36. Recently the General Social Survey asked respondents, “Compared with ten years ago, would you say that American children today are (1) much better off, (2) better off, (3) about the same, (4) worse off, or (5) much worse off.” Table 12.40 shows responses for ten of the subjects on three issues: quality of their education, safety of the neighborhoods they live in, and getting health care when they need it.
- Test the hypothesis that the population means are equal. Report the  $P$ -value, and interpret.
  - For each of the following, indicate whether it is a fixed effect, random effect, or response variable: (i) opinion, (ii) issue, (iii) subject.
37. Refer to the previous exercise. The first five respondents were female, and the last five were male. Analyze these data using both gender and issue as factors.
- Identify the between-subjects and within-subjects factors.
  - Test for interaction. Interpret.
  - Test the main effect of gender. Interpret.
  - Test the main effect of issue. Interpret.
  - Use 95% confidence intervals to compare the means for the three issues. Interpret, and summarize your findings from these analyses.

Table 12.40:

Subject	Status of Children		
	Education	Neighborhood	Health Care
1	4	4	3
2	2	4	2
3	3	3	4
4	1	2	1
5	3	4	3
6	2	5	4
7	1	4	2
8	3	3	3
9	4	5	3
10	2	4	2

38. The General Social Survey asks respondents to rate various groups using the “feeling thermometer” on a scale of 0 (most unfavorable) to 100 (most favorable). We plan to study how the mean compares for rating liberals and rating conservatives, for ratings in 2006 and ratings in 1986. Explain why a two-way ANOVA using time (1986, 2006) and group rated (Liberal, Conservative) as factors would require methods for repeated measures. Identify the within-subjects and between-subjects factors.
39. Upjohn, a pharmaceutical company, conducted a randomized clinical trial comparing an active hypnotic drug with a placebo for patients suffering from insomnia. The outcome is patient response to the question, “How quickly did you fall asleep after going to bed?” Patients suffering from insomnia were randomly assigned to receive the active drug or a placebo. The study measured patients’ responses at the start and at the conclusion of a two-week treatment period. The study analyzed whether the active drug helps subjects with insomnia problems. However, patients taking placebo may also tend to fall asleep more quickly at the conclusion of the study period, because of the *placebo effect*—thinking they are taking a beneficial drug may psychologically have a positive effect. Is the improvement with the active drug better than with the placebo? The sample means are 50.0 initially and 27.8 at follow up for the 120 subjects taking active drug, and 50.3 initially and 37.4 at follow up for the 119 subjects taking placebo. Table 12.41 shows a printout from an analysis of these data.
- For the repeated-measures ANOVA, report and interpret the  $F$  statistic and  $P$ -value for testing interaction.
  - Does it make sense to test main effects? Explain.
  - Using the sample means, interpret the interaction descriptively.
  - For within-subjects comparisons of means at the two times, a 95% confidence interval equals (18.0, 26.4) for the active drug and (8.7, 17.2) for placebo. Interpret.

Table 12.41:

---

Tests of Within-Subject Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig
TIME	36815.95	1	36815.95	133.41	0.0001
TIME*TREATMENT	2543.19	1	2543.19	9.22	0.0027
Error(TIME)	65403.36	237	275.96		

  

Tests of Between-Subjects Effects.					
Source	Type III Sum of Squares	df	Mean Square	F	Sig
TREATMENT	2973.865	1	2973.86	4.08	0.0444
Error	172591.093	127	728.23		

---

40. Table 12.42 shows results of using SAS for analyzing the anorexia data of Table 12.21.
- Explain how to use the information in this table to test  $H_0$ : no interaction between treatment and time.
  - Explain how to determine the  $df$  values for treatment, time, and their interaction.
41. Using software, conduct the repeated measures ANOVA of the anorexia data in Table 12.21, available at the text website.

### Concepts and Applications

42. Refer to the student survey data set (Exercise 1.11), with response variable the number of weekly hours engaged in sports and other physical exercise. Using software, conduct an analysis of variance and follow-up estimation, and prepare a report summarizing your analyses and interpretations using
- Gender as a predictor.
  - Gender and whether a vegetarian as predictors.
43. Refer to the data file your class created in Exercise 1.12. For variables chosen by your instructor, use ANOVA methods. Interpret and summarize your findings.
44. Use software with the “house selling price” data file at the text website.
- Conduct an ANOVA to test equality of the mean selling prices for homes with one, two, and three bathrooms. Interpret.
  - Conduct a  $F$  test as part of a regression analysis that is equivalent to the

Table 12.42:

---

Repeated Measures Analysis of Variance  
Tests of Hypotheses for Between Subjects Effects

Source	DF	Anova SS	Mean Square	F Value	Pr > F
TREATMNT	2	644.23	322.12	6.20	0.0033
Error	69	3584.03	51.94		

Univariate Tests of Hypotheses for Within Subject Effects

Source: TIME

DF	Anova SS	Mean Square	F Value	Pr > F
1	275.0069444	275.0069444	9.70	0.0027

Source: TIME\*TREATMNT

DF	Anova SS	Mean Square	F Value	Pr > F
2	307.3218334	153.6609167	5.42	0.0065

Source: Error(TIME)

DF	Anova SS	Mean Square	F Value	Pr > F
69	1955.3712221	28.3387134		

---

ANOVA in (a).

- c) Explain the difference between conducting a test of independence of selling price and number of bathrooms using the  $F$  test in (a) or (b) and using a regression  $t$  test for the coefficient of the number of bathrooms in a regression model. Give an example of three population means for which the regression test would be less appropriate than the ANOVA test.
45. Go to the GSS website [sda.berkeley.edu/GSS](http://sda.berkeley.edu/GSS), and click on *Comparison of Means* and conduct an ANOVA for the most recent year available to compare the mean of political ideology (POLVIEWS) by the categories (lower, working, middle, upper) for social class (CLASS). Report results and interpret.
46. Repeat the previous exercise, using the nine regions of the country (REGION) as the explanatory variable.
47. For female respondents to the GSS, you would like to determine whether the mean of the number of male sex partners since the 18th birthday (NUMMEN) varies significantly by level of marital status (MARITAL). Show how to do this using the data for the most recent year posted at the website [sda.berkeley.edu/GSS](http://sda.berkeley.edu/GSS).
48. A study<sup>2</sup> of American armed forces who had served in Iraq or Afghanistan reported that of those who had been in a firefight, the  $P$ -value was 0.001 for comparing the number of firefights reported by soldiers deployed in Afghanistan, soldiers deployed in Iraq, and Marines deployed in Iraq. Explain how this relates to analysis of variance, identifying the response variable, the explanatory variable, and hypotheses that could be tested to yield this  $P$ -value. Then summarize how you could interpret the result to someone who has not studied statistics and does not know what a  $P$ -value is.
49. A recent GSS asked, “How often do you attend religious services?” Table 12.43 shows results of ANOVA for comparing three levels of this variable (at least 2-3 times a month for the “high” group, several times a year to about once a month for the “medium” group, and at most once a year for the “low” group) on the mean number of good friends. Use the reported results to make inferences (a significance test and estimation). Summarize your analyses and interpretations.
50. An experiment used four randomly selected groups of five individuals each. The overall sample mean was 60.
- a) What did the data look like if the one-way ANOVA for comparing the means had test statistic  $F = 0$ ?
- b) What did the data look like if  $F = \infty$ ?
51. A study about smoking and personality (by A. Terracciano and P. Costa, *Addiction*, vol. 99, pp. 472-481, 2004) used a sample of 1638 adults in the Baltimore Longitudinal Study on Aging. The subjects formed three groups according to

---

<sup>2</sup>C. Hoge et al., *New England J. Medic.*, vol. 351, 13-21, 2004

Table 12.43:

		How Often Attend Religious Services		
		High	Medium	Low
Mean number good friends		12.1	6.4	6.2
Sample size		337	159	330

  

Source	Sum of Squares	df	Mean Square	F	Sig
Group	6748.2	2	3374.1	14.2	0.000
Error	196136.9	823	238.3		
Total	202885.1	825			

smoking status (never, former, current). Each subject completed a personality questionnaire that provided scores on various personality scales designed to have overall means of about 50 and standard deviations of about 10. Table 12.44 shows some results for three traits, giving the means with standard deviations in parentheses.

- For the  $F$  test for the extraversion scale, using the 0.05 significance level, what conclusion would you make?
- Refer to (a). Does this mean that the population means are necessarily equal? Explain.
- The study measured 35 personality scales and reported an  $F$  test comparing the three smoking groups for each scale. The researchers mentioned doing a Bonferroni correction for the 35  $F$  tests. If the nominal overall probability of Type I error was 0.05 for the 35 tests, how small did the  $P$ -value have to be for a particular test to be significant?

Table 12.44:

	Never smokers (n = 828)	Former smokers (n = 694)	Current smokers (n = 116)	F
Neuroticism	46.7 (9.6)	48.5 (9.2)	51.9 (9.9)	17.77
Extraversion	50.4 (10.3)	50.2 (10.0)	50.9 (9.4)	0.24
Conscientiousness	51.8 (10.1)	48.9 (9.7)	45.6 (10.3)	29.42

52. A study<sup>3</sup> compared verbal memory of men and women for abstract words and for concrete words. It found a gender main effect in favor of women. It also reported, “There was no sex  $\times$  word-type interaction ( $F = .408$ ,  $P = .525$ ), indicating that women were equally advantaged on the two kinds of words.” How would you explain what this sentence means to someone who has never studied statistics?
53. A study<sup>4</sup> compared mental impairment for a group of men who had just stopped smoking to a group of men who continued to smoke, at three times (before cessation, after 6 months, after one year). It reported that the change in the mental health score in the cessation group was significantly larger than in the smoking group. Put this in the context of ANOVA, identifying the groups, the response variable, the hypotheses to be tested, and the type of ANOVA to be employed. Describe what “no interaction” means in this study and whether it seemed plausible.
54. Explain carefully the difference between a probability of Type I error of 0.05 for a single comparison of two means and a multiple comparison error rate of 0.05 for comparing all pairs of means.
55. In multiple comparisons following a one-way ANOVA with equal sample sizes, the margin of error with a 95% confidence interval for comparing each pair of means equals 10. Give three sample means illustrating that it is possible that group A is not significantly different from group B and group B is not significantly different from group C, yet group A is significantly different from group C.
56. Table 12.45 is a contingency table summarizing responses on political ideology in the 2004 General Social Survey by race and gender (Recall 1 = extremely liberal, 4 = moderate, 7 = extremely conservative). The table shows results of 1-way and 2-way ANOVAs comparing the four groups.
- a) Write a paragraph explaining how to interpret the results of the 1-way ANOVA.
- b) Write a paragraph explaining how to interpret the results of the 2-way ANOVA. Write a separate paragraph explaining what you learn from the 2-way ANOVA that you cannot learn from the 1-way ANOVA.
57. Table 7.25 in Exercise 49 in Chapter 7 summarized the mean number of dates in the past three months by gender and by level of physical attractiveness. Identify the response variable and the factors, and indicate whether these data appear to show interaction. Explain.
58. Construct a numerical example of means for a two-way classification under the following conditions:

---

<sup>3</sup>D. Kimura and P. Clarke, *Psych. Reports*, 91, 1137-1142, 2002

<sup>4</sup>Y. Mino et al., *Psychiatry and Clin. Neurosciences*, vol. 54, 169-172, 2000

Table 12.45:

Race	Gender	<i>n</i>	Political Ideology	
			Mean	Standard Deviation
White	Female	553	4.15	1.40
	Male	501	4.42	1.42
Black	Female	100	3.95	1.47
	Male	54	3.93	1.30

Table 12.46:

## 1-way ANOVA

Source	Sum of Squares	df	Mean Square	F	P
Groups	34.46	3	11.49	5.78	0.001
Error	2393.83	1204	1.99		
Total	2428.29	1207			

## 2-way ANOVA

Source	Sum of Squares	df	Mean Square	F	P
Race	12.74	1	12.74	6.41	.011
Sex	16.42	1	16.42	8.26	.004
Interaction	2.66	1	2.66	1.34	.248
Error	2393.83	1204	1.99		
Total	2428.29	1207			

- a) Main effects are present only for the row variable.
  - b) Main effects are present for each variable, but there is no interaction.
  - c) Interaction effects are present.
  - d) No effects of any type are present.
59. The null hypothesis of equality of means for a factor is rejected in a two-way ANOVA. Does this imply that the hypothesis will be rejected in a one-way ANOVA  $F$  test, if the data are collapsed over the levels of the second variable? Explain.
60. For a two-way classification of means by factors A and B, at each level of B the means are equal for the levels of A. Does this imply that the overall means are equal at the various levels of A, ignoring B? Explain the implications, in terms of how results may differ between two-way ANOVA and one-way ANOVA.
61. A random sample of 26 female graduate students at the University of Florida were surveyed about their attitudes toward abortion. Each received a score on abortion attitude according to how many from a list of eight possible reasons for abortion she would accept as a legitimate reason for a woman to seek abortion. Thus, the higher the score, the more favorable the attitude toward abortion as an option in a variety of circumstances. The students were classified as “fundamentalist” or “nonfundamentalist” in their religious beliefs. They were also classified according to their church attendance frequency, “frequent” (more than once a month) or “infrequent.” Table 12.47 displays the abortion attitude scores, classified by religion and church attendance. Using software, analyze the data. Discuss your findings in a short report, indicating the models fitted, hypotheses tested, parameters estimated, and interpretations that follow from your analyses.

Table 12.47:

		Religion	
		Fundamentalist	Nonfundamentalist
Church Attendance	Frequent	0, 3, 4, 0, 3 2, 0, 1, 1	2, 5, 1, 2 3, 3
	Infrequent	4, 3, 4	6, 8, 6, 4 6, 3, 7, 4

62. Table 12.48, based on GSS data, is a contingency table summarizing responses of 29 subjects about government spending on the environment, assistance to big cities, and law enforcement. The response scale was 1 = too little, 2 = about right, 3 = too much. Analyze these data, using repeated measures ANOVA to compare the mean responses for the three types of spending.

Table 12.48:

	Cities								
	1			2			3		
	1	2	3	1	2	3	1	2	3
Law Enforcement Environment	1	2	3	1	2	3	1	2	3
1	3	1	0	6	3	0	5	3	1
2	1	0	0	2	1	0	1	1	0
3	1	0	0	0	0	0	0	0	0

63. Refer to the student survey data set in Exercise 1.11. Use repeated measures analyses to model the weekly number of hours of recreation in terms of type of activity (levels  $S$  = sports and physical exercise,  $T$  = TV watching) and gender. Interpret results of all analyses, and summarize.
64. The Profile of Mood States (POMS) is a depression scale based on responses of a subject to five items – how (sad, gloomy, lonely, unworthy, discouraged) they have felt in the past week, with responses for each item ranging from 0 = not at all to 4 = extremely. The book *Interpreting Basic Statistics* by Z. Holcomb (5th ed., 2007, Pyrczak Publishing, p. 62) describes a study in which formerly sedentary results were given a 10-week cardiovascular exercise program. Their mean POMS score changed from 11.3 to 5.9. By contrast, for a control group who did not exercise, the mean changed from 10.9 to 11.1. Explain how the analysis for this study could use ANOVA methods, highlighting (a) whether it is one-way or two-way, (b) which samples are independent and which are dependent, (c) whether the sample data show evidence of interaction.
65. For subjects aged under 50, there is little difference in mean annual medical expenses for smokers and non-smokers. For subjects aged over 50, there is a large difference. True or false: There is interaction between smoking status and age in their effects on annual medical expenses.
- Select the correct response(s) in Exercises 66–69. (More than one response may be correct.)
66. Analysis of variance and regression are similar in the sense that
- They both assume a quantitative response variable.
  - They both have  $F$  tests for testing that the response variable is statistically independent of the explanatory variable(s).
  - For inferential purposes, they both assume that the response variable  $Y$  is normally distributed with the same standard deviation at all combinations of levels of the explanatory variable(s).
  - They both provide ways of partitioning the variation in  $Y$  into ‘explained’ and ‘unexplained’ components.

67. One-way ANOVA provides relatively more evidence that  $H_0: \mu_1 = \dots = \mu_g$  is false
- a) The smaller the ‘between-groups’ variation and the larger the ‘within-groups’ variation.
  - b) The smaller the ‘between-groups’ variation and the smaller the ‘within-groups’ variation.
  - c) The larger the ‘between-groups’ variation and the smaller the ‘within-groups’ variation.
  - d) The larger the ‘between-groups’ variation and the larger the ‘within-groups’ variation.
68. For four means, a multiple comparison method provides 95% confidence intervals for the differences between the six pairs. Then
- a) For each confidence interval, there is a 0.95 chance that it contains the population difference.
  - b) The probability that all six confidence intervals are correct is 0.70.
  - c) The probability that all six confidence intervals are correct is 0.95.
  - d) The probability that all six confidence intervals are correct is  $(0.95)^6$ .
  - e) The probability is 0.05 that at least one confidence interval does not contain the true difference.
  - f) The confidence intervals are wider than separate 95% confidence intervals for each difference.
69. Interaction terms are needed in a two-way ANOVA model when
- a) Each pair of variables is associated.
  - b) Both explanatory variables have significant effects in the model without interaction terms.
  - c) The difference in means between two categories of one explanatory variable varies greatly among the categories of the other explanatory variable.
  - d) The mean square for interaction is huge compared to the mean square error.
70. \* You know the sample mean, standard deviation, and sample size for each of three groups. Can you conduct an ANOVA  $F$  test comparing the population means, or would you need more information?
71. \* You form a 95% confidence interval in five different situations.
- a) Assuming that the results of the intervals are statistically independent, find the probability that *all* five intervals contain the parameters they are designed to estimate. Find the probability that at least one interval is in error. (*Hint*: Use the binomial distribution.)
  - b) Explain why you should use confidence level 0.9898 for each interval so that the probability that all five intervals contain the parameters equals exactly 0.95. (*Hint*: Find  $(0.9898)^5$ .) Compare this to the confidence coefficient for each interval used in the Bonferroni method.
72. \* Show how to construct a regression model for the analysis of mean income over a three-way classification of gender (male, female), race (white, black), and

type of job (assembly-line worker, janitorial). Interpret the parameters in the model. Assume that there is no interaction.

73. \* This exercise motivates the formula for the between-groups variance estimate in one-way ANOVA. Suppose the sample sizes all equal  $n$  and the population means all equal  $\mu$ . The sampling distribution of each  $\bar{y}_i$  then has mean  $\mu$  and variance  $\sigma^2/n$ . The sample mean of the  $\bar{y}_i$  values is  $\bar{y}$ .
- a) Treating  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_g$  as  $g$  observations having sample mean  $\bar{y}$ , explain why  $\sum(\bar{y}_i - \bar{y})^2/(g - 1)$  estimates the variance  $\sigma^2/n$  of the sampling distribution of the  $\bar{y}_i$ -values.
- b) Using (a), explain why  $\sum n(\bar{y}_i - \bar{y})^2/(g - 1)$  estimates  $\sigma^2$ . For the unequal sample size case, replacing  $n$  by  $n_i$  yields the between-groups estimate.

### 12.9.1 Bibliography

- Howell, D. C. (2006) *Statistical Methods for Psychology*, 6th ed. Wadsworth.
- Kirk, R. E. (1995) *Experimental Design: Procedures for the Behavioral Sciences*, 3rd ed. Brooks/Cole.
- Kutner, M.H., Nachtsheim, C. J., and Neter, J. (2004). *Applied Linear Regression Models*, 5th ed. McGraw-Hill/Irwin.
- Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day.
- Winer, B. J., Brown, D. R., and Michels, K. M. (1991). *Statistical Principles in Experimental Design*, 3rd ed. McGraw-Hill.