CHAPTER 11

Analysis of Variance

Statistics will prove anything, even the truth.

N. Moynihan (British writer)

CONTENTS

We present a general approach, called the *analysis of variance* (ANOVA), for making inferences about the mean values of a variety of random variables. In one-factor ANOVA, the mean of a variable depends on only a single factor, namely, the sample to which it belongs. In two-factor ANOVA, the random variables are thought of as being arrayed in a rectangular arrangement, and the mean of a variable depends on both its row and its column factor. We show how to test the hypothesis that the mean of a random variable does not depend on which row the random variable is in, as well as the analogous hypothesis that the mean does not depend on which column it is in.

Introductory Statistics, DOI: 10.1016/B978-0-12-374388-6.00011-9

11.1 INTRODUCTION

In recent years, many people have expressed the fear that large parts of U.S. industry are becoming increasingly unable to compete effectively in the world economy. For instance, U.S. public opinion has shifted to the belief that Japanese-made automobiles are of higher quality than their U.S.-made counterparts. Japan, and not the United States, is now considered by many to be the world leader in applying statistical techniques to improve quality.

Statistical quality control methods were developed by U.S. industrial statisticians in the 1920s and 1930s. These early methods were primarily concerned with surveillance of existing manufacturing processes. They relied to a large degree on the use of statistical sampling procedures to enable statisticians to quickly detect when something had gone wrong with the manufacturing process. In recent years, however, the emphasis in statistical quality control has shifted from overseeing a manufacturing process to designing that process. That is, led by some Japanese quality control experts, a feeling has developed that the primary contribution of statistics should be in determining effective ways of manufacturing a product.

For instance, when producing computer chips, the manufacturer needs to decide upon the raw materials to be used, the temperature at which to fuse the parts, the shape and the size of the chip, and other factors. For a given set of choices of these factors, the manufacturer wants to know the mean quality value of the resulting chip. This will enable her or him to determine the choices of the factors of production that would be most appropriate for obtaining a quality product.

In this chapter we introduce the statistical technique used for analyzing the foregoing type of problem. It is a general method for making inferences about a multitude of parameters relating to population means. Its use will enable us, for instance, to determine the mean quality level of a manufactured item for a variety of choices of factor settings. The statistical technique was invented by R. A. Fisher and is known as the *analysis of variance* (ANOVA).

Whereas the previous chapter was concerned with hypothesis tests of two population means, this chapter considers tests of multiple population means. For instance, in Sec. 11.2 we will suppose that we have data from m populations and are interested in testing the hypothesis that all the population means are equal. This scenario is said to constitute a *one-factor* analysis of variance, since the model assumes that the mean of a variable depends on only one factor, namely, the sample from which the observation is taken.

In Sec. 11.3 we consider models in which it is assumed that two factors determine the mean value of a variable. In such cases the variables to be observed can be thought of as being arranged in a rectangular array, and the mean value of a specified variable depends on both the row and the column in which it is located. For this *two-factor* analysis of variance problem we show how to estimate the mean values. In addition we show how to test the hypothesis that a specified factor does not affect the mean. For instance, we might have data of the yearly rainfall in various desert locations over a series of years. Two factors would affect the yearly amount of rainfall in a region—the location of the region and the year considered—and we might be interested in testing whether it is only the location and not the year that makes a difference in the mean yearly rainfall.

In all the models considered in this chapter, we assume that the data are normally distributed with the same (though unknown) variance σ^2. The analysis of variance approach for testing a null hypothesis H_0 concerning multiple parameters is based on deriving two estimators of the common variance σ^2. The first estimator is a valid estimator of σ^2 whether the null hypothesis is true or not, while the second one is a valid estimator only when H_0 is true. In addition, when H_0 is not true, this latter estimator will overestimate σ^2, in that the estimator will tend to exceed it. The test compares the values of these two estimators and rejects H_0 when the ratio of the second estimator to the first is sufficiently large. In other words, since the two estimators should be close to each other when H_0 is true (because they both estimate σ^2 in this case) whereas the second estimator should tend to be larger than the first when H_0 is not true, it is natural to reject H_0 when the second estimator is significantly larger than the first.

11.2 ONE-FACTOR ANALYSIS OF VARIANCE

Consider m samples, each of size n. Suppose that these samples are independent and that sample i comes from a population that is normally distributed with mean μ_i and variance $\sigma^2, i = 1, \ldots, m$. We will be interested in testing the null hypothesis

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_m$$

against

$$H_1\colon \text{ not all the means are equal}$$

That is, we will be testing the null hypothesis that all the population means are equal against the alternative that at least two population means differ.

Let $\overline{X}_i$ and S_i^2 denote the sample mean and sample variance, respectively, for the data of the ith sample, $i = 1, \ldots, m$. Our test of our null hypothesis will be carried out by comparing the values of two estimators of the common variance σ^2. Our first estimator, which will be a valid estimator of σ^2 whether or not the null hypothesis is true, is obtained by noting that each of the sample variances S_i^2 is an unbiased estimator of its population variance σ^2. Since we have m of these estimators, namely, $S_1^2, \ldots, S_m^2$, we will combine them into a single estimator by taking their average. That is, our first estimator of σ^2 is given by

$$\frac{1}{m}\sum_{i=1}^{m} S_i^2$$

Note that this estimator was obtained without assuming anything about the truth or falsity of the null hypothesis.

Our second estimator of σ^2 will be a valid estimator only when the null hypothesis is true. So let us assume that H_0 is true, and thus all the population means μ_i are equal, say, $\mu_i = \mu$ for all i. Under this condition it follows that the m sample means $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_m$ will all be normally distributed with the same mean μ and the same variance σ^2/n. In other words, when the null hypothesis is true, the data $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_m$ constitute a sample from a normal population having variance σ^2/n. As a result, the sample variance of these data will, when H_0 is true, be an estimator of σ^2/n. Designate this sample variance by $\overline{S}^2$. That is,

$$\overline{S}^2 = \frac{\sum_{i=1}^{m}(\overline{X}_i - \overline{\overline{X}})^2}{m-1}$$

where

$$\overline{\overline{X}} = \frac{1}{m}\sum_{i=1}^{m}\overline{X}_i$$

Since $\overline{S}^2$ is an unbiased estimator of σ^2/n when H_0 is true, it follows in this case that $n\overline{S}^2$ is an estimator of σ^2. That is, our second estimator of σ^2 is $n\overline{S}^2$. Hence, we have shown that

$$\sum_{i=1}^{m}\frac{S_i^2}{m} \qquad \text{always estimates } \sigma^2$$

$$n\overline{S}^2 \qquad \text{estimates } \sigma^2 \text{ when } \mathrm{H}_0 \text{ is true}$$

Since it can be shown that $n\overline{S}^2$ will tend to be larger than σ^2 when H_0 is not true, it is reasonable to let the test statistic TS be given by

$$\mathrm{TS} = \frac{n\overline{S}^2}{\sum_{i=1}^{m} S_i^2/m}$$

and to reject H_0 when TS is sufficiently large.

To determine how large TS needs to be to justify rejecting H_0, we use the fact that when H_0 is true, TS will have what is known as an *F distribution* with $m-1$ numerator and $m(n-1)$ denominator degrees of freedom. Let $F_{m-1,m(n-1),\alpha}$

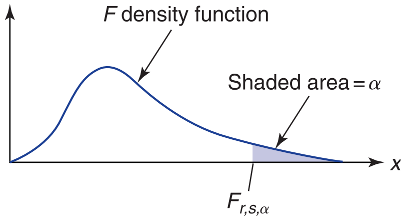

FIGURE 11.1

Random variable F with degrees of freedom r, s: $P\{F \geq F_{r,s,a}\} = \alpha$.

Table 11.1 Values of $F_{r,s,0.05}$

s = Degrees of freedom for denominator	r = Degrees of freedom for numerator			
	1	2	3	4
4	7.71	6.94	6.59	6.39
5	6.61	5.79	5.41	5.19
⋮	⋮	⋮	⋮	⋮
10	4.96	4.10	3.71	3.48

denote the α critical value of this distribution. That is, the probability that an F random variable having numerator and denominator degrees of freedom $m - 1$ and $m(n - 1)$, respectively, will exceed $F_{m-1,m(n-1),\alpha}$ is equal to α (see Fig. 11.1). The significance-level-α test of H_0 is as follows:

$$\begin{array}{ll} \text{Reject } \text{H}_0 & \text{if } \dfrac{n\overline{S}^2}{\sum_{i=1}^{m} \frac{S_i^2}{m}} \geq F_{m-1,m(n-1),\alpha} \\[2ex] \text{Do not reject } \text{H}_0 & \text{otherwise} \end{array}$$

Values of $F_{r,s,0.05}$ for various values of r and s are presented in App. Table D.4. Part of this table is presented now in Table 11.1. For instance, from Table 11.1, we see that there is a 5 percent chance that an F random variable having 3 numerator and 10 denominator degrees of freedom will exceed 3.71.

A Remark on the Degrees of Freedom

The numerator degrees of freedom of the F random variable are determined by the numerator estimator $n\overline{S}^2$. Since $\overline{S}^2$ is the sample variance from a sample of size m, it follows that it has $m - 1$ degrees of freedom. Similarly, the denominator estimator is based on the statistic $\sum_{i=1}^{m} S_i^2$. Since each of the sample variances

S_i^2 is based on a sample of size n, it follows that they each have $n - 1$ degrees of freedom. Summing the m sample variances then results in a statistic with $m(n - 1)$ degrees of freedom.

■ Example 11.1

An investigator for a consumer cooperative organized a study of the mileages obtainable from three different brands of gasoline. Using 15 identical motors set to run at the same speed, the investigator randomly assigned each brand of gasoline to 5 of the motors. Each of the motors was then run on 10 gallons of gasoline, with the total mileages obtained as follows.

Gas 1	Gas 2	Gas 3
220	244	252
251	235	272
226	232	250
246	242	238
260	225	256

Test the hypothesis that the average mileage obtained is the same for all three types of gasoline. Use the 5 percent level of significance.

Solution

Since there are three samples, each of size 5, we see that $m = 3$ and $n = 5$. The sample means are

$$\overline{X}_1 = \frac{1203}{5} = 240.6$$
$$\overline{X}_2 = \frac{1178}{5} = 235.6$$
$$\overline{X}_3 = \frac{1268}{5} = 253.6$$

The average of the three sample means is

$$\overline{\overline{X}} = \frac{240.6 + 235.6 + 253.6}{3} = 243.2667$$

Therefore, the sample variance of the data $\overline{X}_i, i = 1, 2, 3$, is

$$\overline{S}^2 = \frac{(240.6 - 243.2667)^2 + (235.6 - 243.2667)^2 + (253.6 - 243.2667)^2}{2}$$
$$= 86.3333$$

The numerator estimate is thus

$$5\overline{S}^2 = 431.667$$

Computing the sample variances from the three samples yields $S_1^2 = 287.8$, $S_2^2 = 59.3$, and $S_3^2 = 150.8$, so the denominator estimate is

$$\sum_{i=1}^{3} \frac{S_i^2}{3} = 165.967$$

Therefore, the value of the test statistic is

$$\text{TS} = \frac{431.667}{165.967} = 2.60$$

Since $m - 1 = 2$ and $m(n - 1) = 12$, we must compare the value of the TS with the value of $F_{2,12,0.05}$. Now, from App. Table D.4, we see that $F_{2,12,0.05} = 3.89$. Since the value of the test statistic does not exceed 3.89, it follows that at the 5 percent level of significance we cannot reject the null hypothesis that the gasolines give equal mileage.

Another way of doing the computations for the hypothesis test that all the population means are equal is by computing the p value. If the value of the test statistic TS is v, then the p value will be given by

$$p \text{ value } = P\{F_{m-1,m(n-1)} \geq v\}$$

where $F_{m-1,m(n-1)}$ is an F random variable with $m - 1$ numerator and $m(n - 1)$ denominator degrees of freedom. ■

Program 11-1 will compute the value of the test statistic TS and the resulting p value.

■ Example 11.2

Let us do the computations of Example 11.1 by using Program 11-1. After the data have been entered, we get the following output.

```
The denominator estimate is 165.967
The numerator estimate is 431.667
The value of the f-statistic is 2.6009
The p-value is 0.11525
```

■

Table 11.2 summarizes the results of this section.

Remark *When $m = 2$, the preceding is a test of the null hypothesis that two independent samples, having a common population variance, have the same mean. The reader might*

Table 11.2 One-Factor ANOVA Table

Variables $\overline{X}_i$ and S_i^2, $i = 1, \ldots, m$, are the sample means and sample variances, respectively, of independent samples of size n from normal populations having means μ_i and a common variance σ^2.

Source of estimator	Estimator of σ^2	Value of test statistic
Between samples	$n\overline{S}^2 = \dfrac{n\sum_{i=1}^{m}(\overline{X}_i - \overline{\overline{X}})^2}{m-1}$	$\text{TS} = \dfrac{n\overline{S}^2}{\sum_{i=1}^{m}\frac{S_i^2}{m}}$
Within samples	$\displaystyle\sum_{i=1}^{m}\frac{S_i^2}{m}$	

Significance-level-α test of H_0: all μ_1 values are equal:

Reject H_0	if TS $\geq F_{m-1,m(n-1),\alpha}$
Do not reject H_0	otherwise

If $\text{TS} = v$, then p value $= P\{F_{m-1,m(n-1)\geq v}\}$

where $F_{m-1,m(n-1)}$ is an F random variable with $m-1$ numerator and $m(n-1)$ denominator degrees of freedom.

wonder how this compares with the one presented in Chap. 10. It turns out that the tests are exactly the same. That is, assuming the same data are used, they always give rise to exactly the same p value.

PROBLEMS

1. Consider the data from three samples, each of size 4. (That is, $m = 3$, $n = 4$.)

Sample 1	5	9	12	6
Sample 2	13	12	20	11
Sample 3	8	12	16	8

   (a) Determine the three sample means $\overline{X}_i$, $i = 1, 2, 3$.
   (b) Find $\overline{\overline{X}}$, the average of the three sample means.
   (c) Show that $\overline{\overline{X}}$ is equal to the average of the 12 data values.

2. Use the data in Prob. 1 to test the hypothesis that the three population means are equal. Use the 5 percent level of significance.
3. A nutritionist randomly divided 15 bicyclists into three groups of five each. Members of the first group were given vitamin supplements to take with each of their meals over the next 3 weeks. The second group was instructed to eat a particular type of high-fiber whole-grain cereal

for the next 3 weeks. Members of the third group were instructed to eat as they normally do. After the 3-week period elapsed, the nutritionist had each bicyclist ride 6 miles. The following times were recorded:

Vitamin group	15.6	16.4	17.2	15.5	16.3
Fiber cereal group	17.1	16.3	15.8	16.4	16.0
Control group	15.9	17.2	16.4	15.4	16.8

Are these data consistent with the hypothesis that neither the vitamin nor the fiber cereal affects the speed of a bicyclist? Use the 5 percent level of significance.

4. To determine whether the percentage of calories in a person's diet that is due to fat is the same across the country, random samples of 20 volunteers each were chosen in the three different regions. Each volunteer's percentage of total calories due to fat was determined, with the following summarized data resulting.

Region i	$\overline{X}_i$	S_i^2
$i = 1$	32.4	102
$i = 2$	36.4	112
$i = 3$	37.1	138

Test the null hypothesis that the percentage of calories due to fat does not vary for individuals living in the three regions. Use the 5 percent level of significance.

5. Six servings each of three different brands of processed meat were tested for fat content. The following data (in fat percentage per gram of weight) resulted.

Brand			**Fat content**			
1	32	34	31	35	33	30
2	40	36	33	29	35	32
3	37	30	28	33	37	39

Do the data enable us to reject, at the 5 percent level of significance, the hypothesis that the average fat content is the same for all three brands?

6. An important factor in the sales of a new golf ball is how far it will travel when hit. Four different types of balls were hit by an automatic driving machine 25 times each, and their distances (in yards)

were recorded. The following data, referring to the sample means and sample variances obtained with each type of ball, resulted.

Ball i	$\overline{X}_i$	S_i^2
1	212	26
2	220	23
3	198	25
4	214	24

Using the 5 percent level of significance, test the null hypothesis that the mean distance traveled is the same for each type of ball.

7. Three standard chemical procedures are used to determine the magnesium content in a certain chemical compound. Each procedure was used 4 times on a given compound with the following data resulting.

Method 1	76.43	78.61	80.40	78.22
Method 2	80.40	82.24	72.70	76.04
Method 3	82.16	84.14	80.20	81.33

Test the hypothesis that the mean readings are the same for all three methods. Use the 5 percent level of significance.

8. An emergency room physician wanted to learn whether there were any differences in the time it takes for three different inhaled steroids to clear a mild asthmatic attack. Over a period of weeks, she randomly administered these steroids to asthma sufferers and noted the number of minutes it took for the patient's lungs to become clear. Afterward, she discovered that 12 patients had been treated with each type of steroid, with the following sample means and sample variances resulting.

Steroid	$\overline{X}_i$	S_i^2
A	32	145
B	40	138
C	30	150

Test the hypothesis that the mean time to clear a mild asthmatic attack is the same for all three steroids. Use the 5 percent level of significance.

9. The following data refer to the numbers of deaths per 10,000 adults in a large eastern city in different seasons of the years from 1982 to 1986.

Year	Winter	Spring	Summer	Fall
1982	33.6	31.4	29.8	32.1
1983	32.5	30.1	28.5	29.9
1984	35.3	33.2	29.5	28.7
1985	34.4	28.6	33.9	30.1
1986	37.3	34.1	28.5	29.4

Test the hypothesis that death rates do not depend on the season. Use the 5 percent level of significance.

10. A nutrition expert claims that the amount of running a person does relates to that person's blood cholesterol level. Six runners from each of three running categories were randomly chosen to have their blood cholesterol levels checked. The sample means and sample variances were as follows:

Weekly miles run	$\overline{X}_i$	S_i^2
Less than 15	188	190
Between 15 and 30	181	211
More than 30	174	202

Do these data prove the nutritionist's claim? Use the 5 percent level of significance.

11. A college administrator claims that there is no difference in first-year grade-point averages for students entering the college from any of three different local high schools. The following data give the first-year grade-point averages of 15 randomly chosen students—5 from each of the three high schools. Are they strong enough, at the 5 percent level, to disprove the claim of the administrator?

School A	School B	School C
3.2	2.8	2.5
2.7	3.0	2.8
3.0	3.3	2.4
3.3	2.5	2.2
2.6	3.1	3.0

12. A psychologist conducted an experiment concerning maze test scores of a strain of laboratory mice trained under different laboratory conditions. A group of 24 mice was randomly divided into three groups of 8 each. Members of the first group were given a type of cognitive training, those in the second group were given a type of behavioral

training, and those in the third group were not trained at all. The maze test scores (judged by someone who did not know which training particular mice received) were summarized as follows:

Group	$\overline{X}_i$	S_i^2
1	74.2	111.4
2	78.5	102.1
3	80.0	124.0

Is this sufficient evidence to conclude that the different types of training have an effect on maze test scores? Use the 5 percent level of significance.

11.3 TWO-FACTOR ANALYSIS OF VARIANCE: INTRODUCTION AND PARAMETER ESTIMATION

Whereas the model of Sec. 11.2 enabled us to study the effect of a single factor on a data set, we can also study the effects of several factors. In this section we suppose that each data value is affected by two factors.

■ Example 11.3

Four different standardized reading achievement tests were administered to each of five students. Their scores were as follows:

	Student				
Examination	**1**	**2**	**3**	**4**	**5**
1	75	73	60	70	86
2	78	71	64	72	90
3	80	69	62	70	85
4	73	67	63	80	92

Each value in this set of 20 data points is affected by two factors: the examination and the student whose score on that examination is being recorded. The examination factor has four possible values, or *levels*, and the student factor has five possible levels. ■

In general, let us suppose that there are m possible levels of the first factor and n possible levels of the second. Let X_{ij} denote the value of the data obtained when the first factor is at level i and the second factor is at level j. We often portray the data set in the following array of rows and columns:

$$\begin{array}{ccccccc}
X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\
X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\
\multicolumn{6}{c}{\dotfill} \\
X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\
\multicolumn{6}{c}{\dotfill} \\
X_{m1} & X_{m2} & \cdots & X_{mj} & \cdots & X_{mn}
\end{array}$$

Because of this we refer to the first factor as the *row* factor and the second factor as the *column* factor. Also, the data value X_{ij} is the value in row i and column j.

As in Sec. 11.2, we suppose that all the data values $X_{ij}, i = 1, \ldots, m, j = 1, \ldots, n$, are independent normal random variables with common variance σ^2. However, whereas in Sec. 11.2 we supposed that only a single factor affected the mean value of a data point—namely, the sample to which it belonged—in this section we will suppose that the mean value of the data point depends on both its row and its column. However, before specifying this model, we first recall the model of Sec. 11.2. If we let X_{ij} represent the value of the jth member of sample i, then this model supposes that

$$E[X_{ij}] = \mu_i$$

If we now let μ denote the average value of the μ_i, that is,

$$\mu = \frac{\sum_{i=1}^{m} \mu_i}{m}$$

then we can write the preceding as

$$E[X_{ij}] = \mu + \alpha_i$$

where $\alpha_i = \mu_i. - \mu$. With this definition of α_i equal to the deviation of μ_i from the average of the means μ, it is easy to see that

$$\sum_{i=1}^{m} \alpha_i = 0$$

In the case of two factors, we write our model in terms of row and column deviations. Specifically, we suppose that the expected value of variable X_{ij} can be expressed as follows:

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

The value μ is referred to as the *grand mean*, α_i is the *deviation from the grand mean due to row i*, and β_j is the *deviation from the grand mean due to column j*. In addition, these quantities satisfy the following equalities:

$$\sum_{i=1}^{m} \alpha_i = \sum_{j-1}^{n} \beta_j = 0$$

Let us start by determining estimators for parameters μ, α_i, and $\beta_j, i = 1, \ldots, m$, $j = 1, \ldots, n$. To do so, we will find it convenient to introduce the following "dot" notation. Let

$$X_{i.} = \frac{\sum_{j=1}^{n} X_{ij}}{n} = \text{average of all values in row } i$$

$$X_{.j} = \frac{\sum_{i=1}^{m} X_{ij}}{m} = \text{average of all values in column } j$$

$$X_{..} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}}{nm} = \text{average of all } nm \text{ data values}$$

It is not difficult to show that

$$E[X_{i.}] = \mu + \alpha_i$$

$$E[X_{.j}] = \mu + \beta_j$$

$$E[X_{..}] = \mu$$

Since the preceding is equivalent to

$$E[X_{..}] = \mu$$

$$E[X_{i.} - X_{..}] = \alpha_i$$

$$E[X_{.j} - X_{..}] = \beta_j$$

we see that unbiased estimators of μ, α_i and β_j—call them $\hat{\mu}, \hat{\alpha}_i$, and $\hat{\beta}_j$—are given by

$$\hat{\mu} = X_{..}$$

$$\hat{\alpha}_i = X_{i.} - X_{..}$$

$$\hat{\beta}_j = X_{.j} - X_{..}$$

Example 11.4

The following data from Example 11.3 give the scores obtained when four different reading tests were given to each of five students. Use it to estimate the parameters of the model.

	Student						
Examination	1	2	3	4	5	Row totals	$X_{i\cdot}$
1	75	73	60	70	86	364	72.8
2	78	71	64	72	90	375	75
3	80	69	62	70	85	366	73.2
4	73	67	63	80	92	375	75
Column totals	306	280	249	292	353	1480	← grand total
$X_{\cdot j}$	76.5	70	62.25	73	88.25	$X_{\cdot\cdot} = \frac{1480}{20} = 74$	

The estimators are

$$\hat{\mu} = 74$$

$$\begin{array}{ll}
\hat{\alpha}_1 = 72.8 - 74 = -1.2 & \hat{\beta}_1 = 76.5 - 74 = 2.5 \\
\hat{\alpha}_2 = 75 - 74 = 1 & \hat{\beta}_2 = 70 - 74 = -4 \\
\hat{\alpha}_3 = 73.2 - 74 = -0.8 & \hat{\beta}_3 = 62.25 - 74 = -11.75 \\
\hat{\alpha}_4 = 75 - 74 = 1 & \hat{\beta}_4 = 73 - 74 = -1 \\
 & \hat{\beta}_5 = 88.25 - 74 = 14.25
\end{array}$$

Therefore, for instance, if one of the students is randomly chosen and then given a randomly chosen examination, then our estimate of the mean score that will be obtained is $\hat{\mu} = 74$. If we were told that examination i was taken, then this would increase our estimate of the mean score by the amount $\hat{\alpha}_i$; if we were told that the student chosen was number j, then this would increase our estimate of the mean score by the amount $\hat{\beta}_j$. Thus, for instance, we would estimate that the score obtained on examination 1 by student 2 is the value of a random variable whose mean is $\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_2 = 74 - 1.2 - 4 = 68.8$. ■

Remark *In the preceding we defined* $X_{\cdot\cdot}$ *by using the double-summation notation. That is, we used notation of the form*

$$\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}$$

This expression is meant to be the sum of the terms X_{ij} *for all* nm *possible values of the pair* i, j.

Equivalently, suppose that the data values X_{ij} are arranged in rows and columns as given at the beginning of this section. Let T_i denote the sum of the values in row i. That is,

$$T_i = \sum_{j=1}^{n} X_{ij}$$

Then the double summation notation is defined by

$$\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij} = \sum_{i=1}^{m} T_i$$

In words, the double summation is equal to the sum of all the row sums; that is, it is just the sum of all the nm data values X_{ij}. (It is easy to see that it is also equal to the sum of the n column sums.)

PROBLEMS

1. In a study of air pollution, samples of air were taken at three different locations at five different times. The following data refer to the amount of particulate matter present in the air (in units of milligrams per cubic meter).

	Location		
Time	**1**	**2**	**3**
1. January 2006	78	84	87
2. July 2006	75	69	82
3. January 2007	66	60	70
4. July 2007	71	64	61
5. January 2008	58	55	52

Assuming the model

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

estimate the unknown parameters.

2. Using the data of Prob. 1, verify that

$$X_{..} = \frac{\sum_{i=1}^{m} X_{i.}}{m} = \frac{\sum_{j=1}^{n} X_{j.}}{n}$$

Express in words what this equation states.

3. The following data refer to the numbers of boxes packed by each of three men during three different shifts.

	Man		
Shift	**1**	**2**	**3**
1. 9–11 a.m.	32	27	29
2. 1–3 p.m.	31	26	22
3. 3–5 p.m.	33	30	25

Assuming the model of this section, estimate the unknown parameters.

4. Use the results of Example 11.4 to estimate $E[X_{ij}] = \mu + \alpha_i + \beta_j$ for all the possible values of i and $j, i = 1, 2, 3, 4,\ j = 1, 2, 3, 4, 5$. Compare the estimated values $E[X_{ij}]$ with the observed values of X_{ij} as given in that example.

5. The following table gives the birth rates per 1000 population for four different countries in four different years.

	2003	**2002**	**2001**	**1990**
Australia	12.6	12.71	12.86	15.4
Austria	9.4	9.58	9.74	11.6
Belgium	10.4	10.58	10.74	12.6
Czech Republic	9.0	9.08	9.11	13.4

Assuming the model of this section, estimate the

(a) Grand mean of the birth rates

(b) Deviation from the grand mean of Australian birth rates

(c) Deviation from the grand mean of the 1990 birth rates

6. The following table provides the unemployment rates for three levels of educational attainment in four different years.

Level of education	**1980**	**1984**	**1988**	**2000**
Did not graduate from high school	8.4	12.1	9.6	8.8
High school graduate	5.1	7.2	5.4	6.1
College graduate	1.9	2.7	1.7	2.2

Source: U.S. Bureau of Labor Statistics, *Labor Force Statistics.*

Assuming the model of this section, estimate

(a) The grand mean μ

(b) The row deviations, $\alpha_i, i = 1, 2, 3, 4$

(c) The column deviations, $\beta_j, j = 1, 2, 3, 4$

7. The following table provides the unemployment rates for five different industries in three different years.

Industry	2000	2001	2002
Transportation	3.4	4.3	4.9
Mining	4.4	4.2	6.3
Construction	6.2	7.1	9.2
Manufacturing	3.5	5.2	6.7
Information	3.2	4.9	6.9

Source: U.S. Bureau of Labor Statistics, *Employment and Earnings*, monthly.

Assuming the model of this section, estimate the unknown parameters.

8. Suppose that $x_{ij} = i + 4j$. (So, for instance, $x_{11} = 1 + 4 = 5$, and $x_{23} = 2 + 4 \cdot 3 = 14$.) Write out in a rectangular array of rows and columns all the 12 values of x_{ij} where i is 1 or 2 or 3 and j is 1 or 2 or 3 or 4. Put the value of x_{ij} in the location joining row i and column j.

9. In Prob. 8, determine

(a) $\sum_{j=1}^{4} x_{1j}$ (b) $\sum_{j=1}^{4} x_{2j}$

(c) $\sum_{j=1}^{4} x_{3j}$ (d) $\sum_{i=1}^{3}\sum_{j=1}^{4} x_{ij}$

11.4 TWO-FACTOR ANALYSIS OF VARIANCE: TESTING HYPOTHESES

Consider the two-factor model in which one has data values X_{ij}, $i = 1, \dots, m$ and $j = 1, \dots, n$. These data are assumed to be independent normal random variables with a common variance σ^2 and with mean values satisfying

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

where

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 0$$

In this section we will test the hypothesis

$$H_0\colon \text{ all } \alpha_i = 0$$

against

$$\text{H}_1\text{: not all } \alpha_i \text{ are } 0$$

This null hypothesis states that there is no row effect, in that the value of a datum is not affected by its row factor level.

We will also test the analogous hypothesis for columns, namely,

$$\text{H}_0\text{: all } \beta_j \text{ are } 0$$

against

$$\text{H}_1\text{: not all } \beta_j \text{ are } 0$$

To obtain tests for the foregoing null hypotheses, we will apply the analysis of variance approach in which two different estimators are derived for the variance σ^2. The first will always be a valid estimator, whereas the second will be a valid estimator only when the null hypothesis is true. In addition, the second estimator will tend to overestimate σ^2 when the null hypothesis is not true.

To obtain our first estimator of σ^2, we recall that the sum of the squares of N standard normal random variables is a chi-squared random variable with N degrees of freedom. Since the nm standardized variables

$$\frac{X_{ij} - E[X_{ij}]}{\sigma}$$

$i = 1, \ldots, m, j = 1, \ldots, n$ are all standard normal, it follows that

$$\frac{\sum_{i=1}^m \sum_{j=1}^n (X_{ij} - E[X_{ij}])^2}{\sigma^2} = \frac{\sum_{i=1}^m \sum_{j=1}^n (X_{ij} - \mu - \alpha_i - \beta_j)^2}{\sigma^2}$$

is chi squared with nm degrees of freedom. If in the preceding expression we now replace the unknown parameters $\mu, \alpha_1, \alpha_2, \ldots, \alpha_m, \beta_1, \beta_2, \ldots, \beta_n$ by their estimators $\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_m, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$, then it turns out that the resulting expression will remain chi squared but will lose 1 degree of freedom for each parameter that is estimated. To determine how many parameters are to be estimated, we must be careful to remember that $\sum_{i=1}^m \alpha_i = \sum_{j=1}^n \beta_j = 0$. Since the sum of all the α_i is 0, it follows that once we have estimated $m - 1$ of the α_i then we have also estimated the final one. Hence, only $m - 1$ parameters are to be estimated in order to determine all the estimators $\hat{\alpha}_i$. For the same reason only $n - 1$ of the β_j need to be estimated to determine estimators for all n of them. Since μ is also to be estimated, we see that the number of parameters to be estimated is

$$1 + (m - 1) + (n - 1) = m + n - 1$$

As a result, it follows that

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\hat{\mu}-\hat{\alpha}_i-\hat{\beta}_j)^2}{\sigma^2}$$

is a chi-squared random variable with $nm-(n+m-1)=(n-1)(m-1)$ degrees of freedom.

Since

$$\begin{aligned}\hat{\mu} &= X_{..}\\ \hat{\alpha}_i &= X_{i.}-X_{..}\\ \hat{\beta}_j &= X_{.j}-X_{..}\end{aligned}$$

we see that

$$\begin{aligned}\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j &= X_{..}+X_{i.}-X_{..}+X_{.j}-X_{..}\\ &= Xi.+X_{.j}-X_{..}\end{aligned}$$

Thus, the statistic

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-X_{i.}-X_{.j}+X_{..})^2}{\sigma^2} \tag{11.1}$$

is chi squared with $(n-1)(m-1)$ degrees of freedom.

The sum of squares SS_e defined by

$$SS_e=\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-X_{i.}-X_{.j}+X_{..})^2$$

is called the *error sum of squares*.

If we think of the difference between a random variable and its estimated mean as being an "error," then SS_e is equal to the sum of the squares of the errors. Since SS_e/σ^2 is just the expression in Eq. (11.1), we see that SS_e/σ^2 is chi squared with $(n-1)(m-1)$ degrees of freedom. As the expected value of a chi-squared random variable is equal to its number of degrees of freedom, we have

$$E\left[\frac{SS_e}{\sigma^2}\right]=(n-1)(m-1)$$

or

$$E\left[\frac{SS_e}{(n-1)(m-1)}\right] = \sigma^2$$

That is, letting $N = (n-1)(m-1)$, we have shown the following.

$\dfrac{SS_e}{N}$ is an unbiased estimator of σ^2.

Suppose now that we want to test the null hypothesis that there is no row effect; that is, we want to test

$$H_0 : \text{all the } \alpha_i \text{ are } 0$$

against

$$H_1 : \text{not all the } \alpha_i \text{ are } 0$$

To obtain a second estimator of σ^2, consider the row averages $X_{i\cdot}, i = 1, \ldots, m$. Note that when H_0 is true, each α_i is equal to 0, and so

$$E[X_{i\cdot}] = \mu + \alpha_i = \mu$$

Since each $X_{i\cdot}$ is the average of n random variables, each having variance σ^2, it follows that

$$\text{Var}(X_{i\cdot}) = \frac{\sigma^2}{n}$$

Thus, we see that when H_0 is true,

$$\frac{\sum_{i=1}^m (X_{i\cdot} - E[X_{i\cdot}])^2}{\text{Var}(X_{i\cdot})} = \frac{n\sum_{i=1}^m (X_{i\cdot} - \mu)^2}{\sigma^2}$$

will be chi squared with m degrees of freedom. If we now substitute $X_{\cdot\cdot}$ (the estimator of μ) for μ in the preceding, the resulting expression will remain chi squared but with one less degree of freedom. That is, it will have $m-1$ degrees of freedom. We thus have the following.

When H_0 *is true*, then

$$\frac{n\sum_{i=1}^m (X_{i\cdot} - X_{\cdot\cdot})^2}{\sigma^2}$$

is chi squared with $m-1$ degrees of freedom.

The statistic SS_r defined by

$$\mathrm{SS}_r = n\sum_{i=1}^{m}(X_{i\cdot} - X_{\cdot\cdot})^2$$

is called the *row sum of squares*.

We have already seen that when H_0 is true, SS_r/σ^2 is chi squared with $m - 1$ degrees of freedom. As a result, when H_0 is true,

$$E\left[\frac{\mathrm{SS}_r}{\sigma^2}\right] = m - 1$$

or, equivalently,

$$E\left[\frac{\mathrm{SS}_r}{m-1}\right] = \sigma^2$$

In addition, it can be shown that $\mathrm{SS}_r/(m-1)$ will tend to be larger than σ^2 when H_0 is not true. Thus, once again we have obtained two estimators of σ^2. The first estimator, SS_e/N, where $N = (n-1)(m-1)$, is a valid estimator whether or not the null hypothesis is true. The second estimator, $\mathrm{SS}_r/(m-1)$, is a valid estimator of σ^2 only when H_0 is true and tends to be larger than σ^2 when H_0 is not true.

The test of the null hypothesis H_0 that there is no row effect involves comparing the two estimators just given, and it calls for rejection when the second is significantly larger than the first. Specifically, we use the test statistic

$$\mathrm{TS} = \frac{\mathrm{SS}_r/(m-1)}{\mathrm{SS}_e/N}$$

and the significance-level-α test is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if TS} \geq F_{m-1,N,\alpha} \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

Alternatively the test can be performed by calculating the p value. If the value of the test statistic is v, then the p value is given by

$$p \text{ value} = P\{F_{m-1,N} \geq v\}$$

where $F_{m-1,N}$ is an F random variable with $m - 1$ numerator and N denominator degrees of freedom.

A similar test can be derived to test the null hypothesis that there is no column effect, that is, that all the β_j are equal to 0. The results of both tests are summarized in Table 11.3.

Table 11.3 Two-Factor ANOVA

	Sum of squares	Degrees of freedom
Row	$\mathrm{SS}_r = n\sum_{i=1}^{m}(X_{i.} - X_{..})^2$	$m - 1$
Column	$\mathrm{SS}_c = m\sum_{j=1}^{n}(X_{.ij} - X_{..})^2$	$n - 1$
Error	$\mathrm{SS}_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i}. - X_{.j} + X_{..})^2$	$(n - 1)(m - 1)$
	Let $N = (n - 1)(m - 1)$	

Null hypothesis	Test statistic	Significance-level-α test	p Value if $\mathrm{TS} = v$
No row effect (all $\alpha_i = 0$)	$\dfrac{\mathrm{SS}_r/(m-1)}{\mathrm{SS}_e/N}$	Reject if $\mathrm{TS} \geq F_{m-1,N,\alpha}$	$P\{F_{m-1,N} \geq v\}$
No column effect (all $\beta_j = 0$)	$\dfrac{\mathrm{SS}_e/(n-1)}{\mathrm{SS}_e/N}$	Reject if $\mathrm{TS} \geq F_{n-1,N,\alpha}$	$P\{F_{n-1,N} \geq v\}$

Program 11-2 will do the computations and give the p value.

■ Example 11.5

The following are the numbers of defective items produced by four workers using, in turn, three different machines.

	Worker			
Machine	**1**	**2**	**3**	**4**
1	41	42	40	35
2	35	42	43	36
3	42	39	44	47

Test whether there are significant differences between the machines and the workers.

Solution

Since there are three rows and four columns, we see that $m = 3$ and $n = 4$ Computing the row and column averages gives the following results:

$$X_{1.} = \frac{41 + 42 + 40 + 35}{4} = 39.5 \qquad X_{.1} = \frac{41 + 35 + 42}{3} = 39.33$$

$$X_{2.} = \frac{35+42+43+36}{4} = 39 \qquad X_{.2} = \frac{42+42+39}{3} = 41$$

$$X_{3.} = \frac{42+39+44+47}{4} = 43 \qquad X_{.3} = \frac{40+43+44}{3} = 42.3$$

$$X_{.4} = \frac{35+36+47}{3} = 39.33$$

Also

$$X_{..} = \frac{39.5+39+43}{3} = 40.5$$

Thus,

$$\begin{aligned} SS_r &= n\sum_{i=1}^{m}(X_{i.} - X_{..})^2 \\ &= 4[1^2 + (1.5)^2 + (2.5)^2] \\ &= 38 \end{aligned}$$

and

$$\begin{aligned} SS_c &= m\sum_{j=1}^{n}(X_{.j} - X_{..})^2 \\ &= 3[(1.17)^2 + (0.5)^2 + (1.83)^2 + (1.17)^2] \\ &= 19.010 \end{aligned}$$

The calculation of SS_e is more involved because we must add the sum of the squares of the terms $X_{ij} - X_{i.} - X_{.j} + X_{..}$ as i ranges from 1 to 3 and j from 1 to 4. The first term in this sum, when $i = 1$ and $j = 1$, is

$$(41 - 39.5 - 39.33 + 40.5)^2$$

Adding all 12 terms gives

$$SS_e = 94.05$$

Since $m - 1 = 2$ and $N = 2 \cdot 3 = 6$, the test statistic for the hypothesis that there is no row effect is

$$\text{TS(row)} = \frac{38/2}{94.05/6} = 1.21$$

From App. Table D.4 we see that $F_{2,6,0.05} = 5.14$, and so the hypothesis that the mean number of defective items is unaffected by which machine is used is not rejected at the 5 percent level of significance.

The test statistic for the hypothesis that there is no column effect is

$$\text{TS(col.)} = \frac{19.010/3}{94.05/6} = 0.40$$

From App. Table D.4 we see that $F_{3,6,0.05} = 4.76$, and so the hypothesis that the mean number of defective items is unaffected by which worker is used is also not rejected at the 5 percent level of significance. ■

We could also have solved the above by running a program such as Program 11-2. Running Program 11-2 yields the following output:

```
The value of the F-statistic for testing that there is no row effect is 1.212766
The p-value for testing that there is no row effect is 0.3571476
The value of the F-statistic for testing that there is no column effect is 0.4042554
The p-value for testing that there is no column effect is 0.7555629
```

Since both p values are greater than 0.05, we cannot reject at the 5 percent significance level the hypothesis that the machine used does not affect the mean number of defective items produced; nor can we reject the hypothesis that the worker employed does not affect the mean number of defective items produced.

PROBLEMS

1. An experiment was performed to determine the effect of three different fuels and three different types of launchers on the range of a certain missile. The following data, in the number of miles traveled by the missile, resulted.

	Fuel 1	Fuel 2	Fuel 3
Launcher 1	70.4	71.7	78.5
Launcher 2	80.2	82.8	76.4
Launcher 3	90.4	85.7	84.8

Find out whether these data imply, at the 5 percent level of significance, that there are differences in the mean mileages obtained by using

(a) Different launchers

(b) Different fuels

2. An important consideration in deciding which database management system to employ is the mean time required to learn how to use the system. A test was designed involving three systems and four users.

Each user took the following amount of time (in hours) in training with each system:

	User			
	1	**2**	**3**	**4**
System 1	20	23	18	17
System 2	20	21	17	16
System 3	28	26	23	22

(a) Using the 5 percent level of significance, test the hypothesis that the mean training time is the same for all the systems.

(b) Using the 5 percent level of significance, test the hypothesis that the mean training time is the same for all the users.

3. Five different varieties of oats were planted in each of four separated fields. The following yields resulted.

	Field			
Oat variety	**1**	**2**	**3**	**4**
1	296	357	340	348
2	402	390	420	335
3	345	342	358	308
4	360	322	336	270
5	324	339	357	308

Find out whether the data are consistent with the hypothesis that the mean yield does not depend on

(a) The field

(b) The oat variety

Use the 5 percent level of significance.

4. In Example 11.3, test the hypothesis that the mean score of a student does not depend on which test is taken.

5. In Prob. 1 of Sec. 11.3, test the hypothesis that the mean air pollution level

(a) Is unchanging in time

(b) Does not depend on the location

Use the 5 percent level of significance.

6. In Prob. 3 of Sec. 11.3, test the hypothesis that the mean number of boxes packed does not depend on

(a) The worker doing the packing

(b) The shift

Use the 5 percent level of significance.

7. The following data give the percentages of random samples of United Kingdom citizens who were smokers, in a variety of years.

	Age (years)					
Year	**16–19**	**20–24**	**25–34**	**35–49**	**50–59**	**60+**
1978	34	44	45	45	45	30
1988	28	37	36	36	33	23
1998	31	40	35	30	27	16
2000	29	35	35	29	27	16
2002	25	38	34	28	26	15
2007	20	31	27	22	21	12

(a) Test the hypothesis that the actual percentages of smokers do not depend on the year considered.
(b) Test the hypothesis that there is no effect due to age group.

8. In Prob. 5 of Sec. 11.3, test the hypothesis that
(a) The mean birth rates do not depend on the particular country being considered.
(b) The mean birth rates do not depend on the particular year being considered.

9. In Prob. 7 of Sec. 11.3, test the hypothesis that
(a) The mean unemployment rates do not depend on the particular industry being considered.
(b) The mean unemployment rates do not depend on the particular year being considered.

11.5 FINAL COMMENTS

This chapter presented a brief introduction to a powerful statistical technique known as the *analysis of variance* (ANOVA). This technique enables statisticians to draw inferences about population means when these mean values are affected by many different factors. For instance, whereas we have considered only one- and two-factor ANOVA problems, any number of factors could affect the value of an outcome. In addition, there could be interactions between some of these factors. For instance, in two-factor ANOVA, it might be the case that the combination of a particular row and a particular column greatly affects a mean value. For example, while individually each of two carcinogens may be relatively harmless, perhaps in conjunction they are devastating. The general theory of ANOVA shows how to deal with these and a variety of other situations.

ANOVA was developed by R. A. Fisher, who applied it to a large number of agricultural problems during his tenure as chief scientist at the Rothamstead Experimental Laboratories. ANOVA has since been widely applied in a variety of fields. For instance, in education one might want to study how a student's learning of algebra is affected by such factors as the instructor, the syllabus of the algebra course, the time spent on each class, the number of classes, the number of students in each class, and the textbook used. ANOVA has also been widely applied in studies in psychology, social science, manufacturing, biology, and many other fields. Indeed, ANOVA is probably the most widely used of all the statistical techniques.

KEY TERMS

One-factor analysis of variance: A model concerning a collection of normal random variables. It supposes that the variances of these random variables are equal and that their mean values depend on only a single factor, namely, the sample to which the random variable belongs.

***F* statistic:** A test statistic that is, when the null hypothesis is true, a ratio of two estimators of a common variance.

Two-factor analysis of variance: A model in which a set of normal random variables having a common variance is arranged in an array of rows and columns. The mean value of any of them depends on two factors, namely, the row and the column in which the variable lies.

SUMMARY

One-Factor Analysis of Variance Consider m independent samples, each of size n. Let $\mu_1, \mu_2, \ldots, \mu_m$ be the respective means of these m samples, and consider a test of

$$\mathrm{H}_0\text{: all the means are equal}$$

against

$$\mathrm{H}_1\text{: not all the means are equal}$$

Let $\overline{X}_i$ *and* S_i^2 denote the sample mean and sample variance, respectively, from sample $i, i = 1, \ldots, m$. Also, let $\overline{S}^2$ be the sample variance of the data set $\overline{X}_1, \ldots, \overline{X}_m$.

To test H_0 *against* H_1, use the test statistic

$$\mathrm{TS} = \frac{n\overline{S}^2}{\sum_{i=1}^m S_i^2/m}$$

The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if TS } \geq F_{m-1,m(n-1),\alpha} \\ \text{Not reject } H_0 & \text{Otherwise} \end{array}$$

If the value of TS is v, then

$$p \text{ value} = P\{F_{m-1,m(n-1)\geq v}\}$$

Program 11-1 can be used both to compute the value of TS and to obtain the resulting p value.

Note: Variable $F_{r,s}$ represents an F random variable having r numerator and s denominator degrees of freedom. Also, $F_{r,s,\alpha}$ is defined to be such that

$$P\{F_{r,s} \geq F_{r,s,\alpha}\} = \alpha$$

Two-Factor Analysis of Variance: *The Model*. Suppose that each data value is affected by two factors, and suppose that there are m possible values, or levels, of the first factor and n of the second factor. Let X_{ij} denote the datum obtained when the first factor is at level i and the second factor is at level j. The data set can be arranged in the following array of rows and columns:

$$\begin{array}{cccccc} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots\dots} \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots\dots} \\ X_{m1} & X_{m2} & \cdots & X_{mj} & \cdots & X_{mn} \end{array}$$

The two-factor ANOVA model supposes that the X_{ij} are normal random variables having means given by

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

and a common variance

$$\text{Var}(X_{ij}) = \sigma^2$$

The foregoing parameters satisfy

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 0$$

Estimating the Parameters. Let

$$X_{i\cdot} = \frac{\sum_{j=1}^{n} X_{ij}}{n}$$
$$X_{\cdot j} = \frac{\sum_{i=1}^{m} X_{ij}}{m}$$
$$X_{\cdot\cdot} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}}{nm}$$

The estimators of the parameters are as follows:

$$\hat{\mu} = X_{\cdot\cdot}$$
$$\hat{\alpha}_i = X_{i\cdot} - X_{\cdot\cdot}$$
$$\hat{\beta}_j = X_{\cdot j} - X_{\cdot\cdot}$$

Testing Hypotheses. Let

$$SS_e = \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - X_{i\cdot} - X_{\cdot j} + X_{\cdot\cdot})^2$$
$$SS_r = n \sum_{i=1}^{m} (X_{i\cdot} - X_{\cdot\cdot})^2$$
$$SS_c = m \sum_{j=1}^{n} (X_{\cdot j} - X_{\cdot\cdot})^2$$

SS_e, SS_r, and SS_c are called, respectively, the error sum of squares, the row sum of squares, and the column sum of squares. Also let $N = (n-1)(m-1)$.

To test H_0: *all* $\alpha_i = 0$ *against* H_1: *not all* $\alpha_i = 0$, use test statistic

$$TS = \frac{SS_r/(m-1)}{SS_e/N}$$

The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } TS \geq F_{m-1,N,\alpha} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

If $TS = v$, then the p value is given by

$$p \text{ value} = P\{F_{m-1,N} \geq v\}$$

To test H_0: *all* $\beta_j = 0$ *versus* H_1: *not all* $\beta_j = 0$, use test statistic

$$TS = \frac{SS_c/(n-1)}{SS_e/N}$$

The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if TS } \geq F_{n-1,N,\alpha} \\ \text{Not reject } H_0 & \text{Otherwise} \end{array}$$

If TS $= v$, then the p value is given by

$$p \text{ value} = P\{F_{n-1,N} \geq v\}$$

Program 11-2 can be used for the foregoing hypothesis tests. It will compute the values of the two test statistics and give the resulting p values.

REVIEW PROBLEMS

1. A corporation has three apparently identical manufacturing plants. Wanting to see if these plants are equally effective, management randomly chose 30 days. On 10 of these days it determined the daily output at plant 1. On another 10 days, it determined the daily output at plant 2, and on the final 10 days management determined the daily output at plant 3. The following summary data give the sample means and sample variances of the daily numbers of items produced at the three plants over those days.

Plant i	$\overline{X}_i$	S_i^2
$i = 1$	325	450
$i = 2$	413	520
$i = 3$	366	444

Test the hypothesis that the mean number of items produced daily is the same for all three plants. Use the 5 percent level of significance.

2. Sixty nonreading preschool students were randomly divided into four groups of 15 each. Each group was given a different type of course in learning how to read. Afterward, the students were tested with the following results.

Group	$\overline{X}_i$	S_i^2
1	65	224
2	62	241
3	68	233
4	61	245

Test the null hypothesis that the reading courses are equally effective. Use the 5 percent level of significance.

3. Preliminary studies indicate a possible connection between one's natural hair color and threshold for pain. A sample of 12 women were classified as to having light, medium, or dark hair. Each was then given a pain sensitivity test, with the following scores resulting.

Light	Medium	Dark
63	60	45
72	48	33
52	44	57
60	53	40

Are the given data sufficient to establish that hair color affects the results of a pain sensitivity test? Use the 5 percent level of significance.

4. Three different washing machines were employed to test four different detergents. The following data give a coded score of the effectiveness of each washing.

	Machine		
Detergent	**1**	**2**	**3**
1	53	50	59
2	54	54	60
3	56	58	62
4	50	45	57

(a) Estimate the improvement in mean value with detergent 1 over detergent (i) 2, (ii) 3, and (iii) 4.

(b) Estimate the improvement in mean value when machine 3 is used as opposed to machine (i) 1 and (ii) 2.

(c) Test the hypothesis that the detergent used does not affect the score.

(d) Test the hypothesis that the machine used does not affect the score.

In both (c) and (d), use the 5 percent level of significance.

5. Suppose in Prob. 4 that the 12 applications of the detergents were all on different randomly chosen machines. Test the hypothesis, at the 5 percent significance level, that the detergents are equally effective.

6. In Example 11.3 test the hypothesis that the mean test score depends only on the test taken and not on which student is taking the test.

7. A manufacturer of women's beauty products is considering four new variations of a hair dye. An important consideration in a hair dye is

its lasting power, defined as the number of days until treated hair becomes indistinguishable from untreated hair. To learn about the lasting power of its new variations, the company hired three long-haired women. Each woman's hair was divided into four sections, and each section was treated by one of the dyes. The following data concerning the lasting power resulted.

	Dye			
Woman	**1**	**2**	**3**	**4**
1	15	20	27	21
2	30	33	25	27
3	37	44	41	46

(a) Test, at the 5 percent level of significance, the hypothesis that the four variations have the same mean lasting power.

(b) Estimate the mean lasting power obtained when woman 2 uses dye 2.

(c) Test, at the 5 percent level of significance, the hypothesis that the mean lasting power does not depend on which woman is being tested.

8. Use the following data to test the hypotheses of

(a) No row effect

(b) No column effect

17	23	35	39	5
42	28	19	40	14
36	23	31	44	13
27	40	25	50	17

9. Problem 9 of Sec. 11.2 implicitly assumes that the number of deaths is not affected by the year under consideration. However, consider a two-factor ANOVA model for this problem.

(a) Test the hypothesis that there is no effect due to the year.

(b) Test the hypothesis that there is no seasonal effect.

10. The following data relate to the ages at death of a certain species of rats that were fed one of three types of diet. The rats chosen were of a type having a short life span, and they were randomly divided into three groups. The data are the sample means and sample variances

of the ages of death (measured in months) of the three groups. Each group is of size 8.

	Very low-calorie	Moderate-calorie	High-calorie
Sample mean	22.4	16.8	13.7
Sample variance	24.0	23.2	17.1

Test the hypothesis, at the 5 percent level of significance, that the mean lifetime of a rat is not affected by its diet. What about at the 1 percent level?