© The McGraw–Hill Companies, 2007

CHAPTER 11

Analysis of Variance

Chapter Contents

Chapter Learning Objectives

When you finish this chapter you should be able to

- Use basic ANOVA terminology correctly (e.g., response variable, factors, treatments).
- Use a table or Excel to find critical values for the *F* distribution.
- Recognize from the data format which type of ANOVA is appropriate.
- Use Excel or another software package to perform ANOVA calculations.
- Explain the assumptions of ANOVA and why they are important.
- Understand and perform Tukey's test for differences in pairs of group means.
- Use the F_{max} or other tests for equal variances in *c* treatment groups.
- Interpret main effects and interaction effects in two-factor ANOVA.
- Explain the advantages of replication in two-factor ANOVA.
- Recognize when higher-order ANOVA models are needed and why Excel is insufficient.
- Explain why experimental design is important and list a few common designs.

© The McGraw–Hill Companies, 2007

11.1 OVERVIEW OF ANOVA

Chapter 12

You have already learned to compare the means of two samples. In this chapter, you will learn to compare more than two means *simultaneously* and how to trace sources of variation to potential explanatory factors by using ***analysis of variance*** (commonly referred to as ***ANOVA***). Proper *experimental design* can make efficient use of limited data to draw the strongest possible inferences. Although analysis of variance has a relatively short history, it is one of the richest and most thoroughly explored fields of statistics. Originally developed by the English statistician Ronald A. Fisher (1890–1962) in connection with agricultural research (factors affecting crop growth), it was quickly applied in biology and medicine. Because of its versatility, it is now used in engineering, psychology, marketing, and many other areas. In this chapter, we will only illustrate a few kinds of problems where ANOVA may be utilized (see Related Reading if you need to go further).

The Goal: Explaining Variation

Analysis of variance seeks to identify *sources of variation* in a numerical *dependent* variable Y (the ***response variable***). Variation in the response variable about its mean either is ***explained*** by one or more categorical *independent* variables (the ***factors***) or is ***unexplained*** (random error):

$$\begin{array}{c}\text{Variation in } Y \\ \text{(around its mean)}\end{array} = \begin{array}{c}\text{Explained Variation} \\ \text{(due to factors)}\end{array} + \begin{array}{c}\text{Unexplained Variation} \\ \text{(random error)}\end{array}$$

ANOVA is a *comparison of means*. Each possible value of a factor or combination of factors is a ***treatment.*** Sample observations within each treatment are viewed as coming from populations with possibly different means. We test whether each factor has a significant effect on Y, and sometimes we test for interaction between factors. The test uses the F distribution, which was introduced in Chapter 10. ANOVA can handle any number of factors, but the researcher often is interested only in a few. Also, data collection costs may impose practical limits on the number of factors or treatments we can choose. This chapter concentrates on ANOVA models

with one or two factors, although more complex models are briefly mentioned at the end of the chapter.

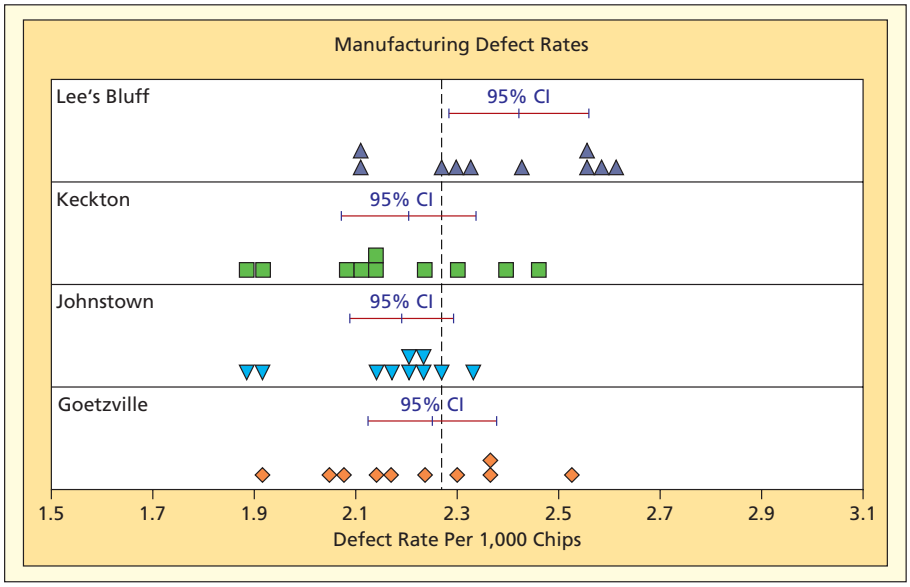
Illustration: Manufacturing Defect Rates

Figure 11.1 shows a dot plot of daily defect rates for automotive computer chips manufactured at four plant locations. Samples of 10 days' production were taken at each plant. Are the observed differences in the plants' sample mean defect rates merely due to random variation? Or are the observed differences between the plants' defect rates too great to be attributed to chance? This is the kind of question that ANOVA is designed to answer.

FIGURE 11.1

Chip defect rates at four plants. The treatment means are significantly different ($p = .02$). Note that the confidence interval for Lee's Bluff falls to the right of the dotted vertical line, which represents the overall mean.

VS Chapter 12



A simple way to state the ANOVA hypothesis is

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (mean defect rates are the same at all four plants)

H_1: Not all the means are equal (at least one mean differs from the others)

If we cannot reject H_0, then we conclude that the observations within each treatment or group actually have a common mean μ (represented by a dashed line in Figure 11.1). This one-factor ANOVA model may be visualized as in Figure 11.2.

FIGURE 11.2

ANOVA model for chip defect rate

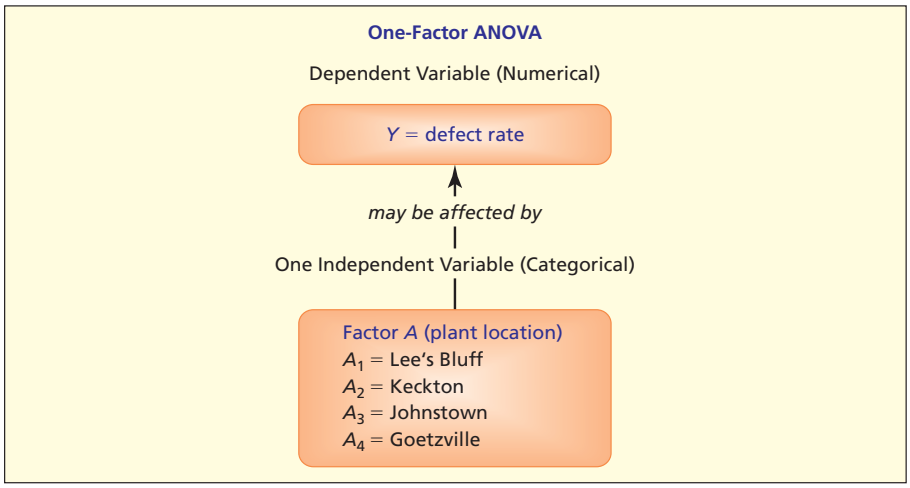


Illustration: Hospital Length of Stay

To allocate resources and fixed costs correctly, hospital management needs to test whether a patient's length of a stay (LOS) depends on the diagnostic-related group (DRG) code and the patient's age group. Consider the case of a bone fracture. LOS is a *numerical* response variable (measured in hours). The hospital organizes the data by using five diagnostic codes for type of fracture (facial, radius or ulna, hip or femur, other lower extremity, all other) and three age groups (under 18, 18 to 64, 65 and over). Although patient age is a numerical variable, it is coded into three categories based on stages of bone growth. Figure 11.3 illustrates two possible ANOVA models (one-factor or two-factor). We could also test for ***interaction*** between factors, as you will see later on.

One factor: Length of stay $= f$(Type of Fracture)

Two factors: Length of stay $= f$(Type of Fracture, Age Group)

FIGURE 11.3

ANOVA models for hospital length of stay

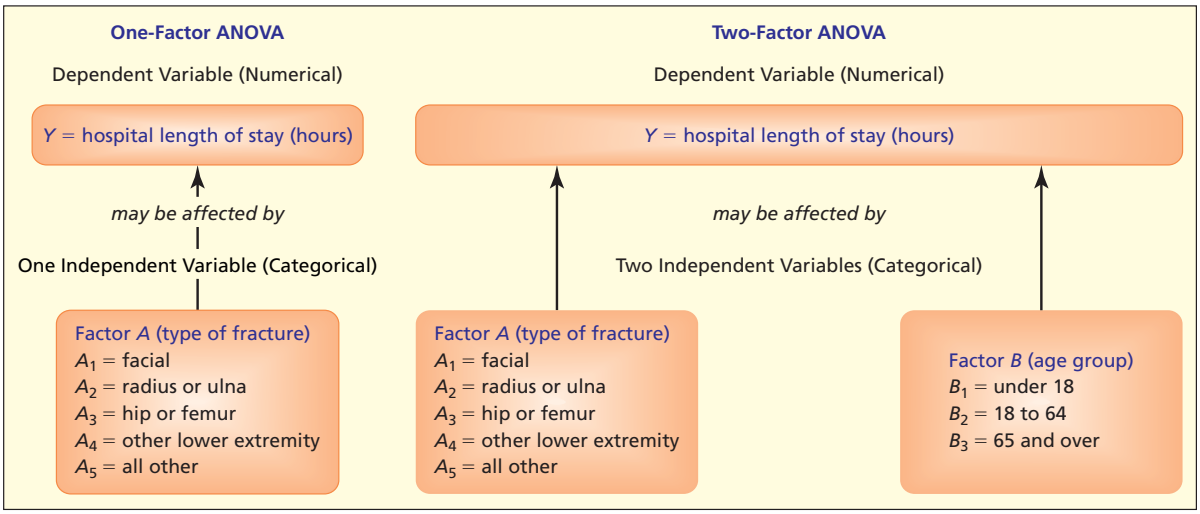


Illustration: Automobile Painting

Paint quality is a major concern of car makers. A key characteristic of paint is its viscosity, a continuous *numerical* variable. Viscosity is to be tested for dependence on application temperature (low, medium, high) and/or the supplier of the paint (Sasnak Inc., Etaoin Ltd., or Shrdlu Inc.). Although temperature is a numerical variable, it has been coded into *categories* that represent the test conditions of the experiment. Figure 11.4 illustrates two potential ANOVA models:

One factor: Viscosity $= f$(temperature)

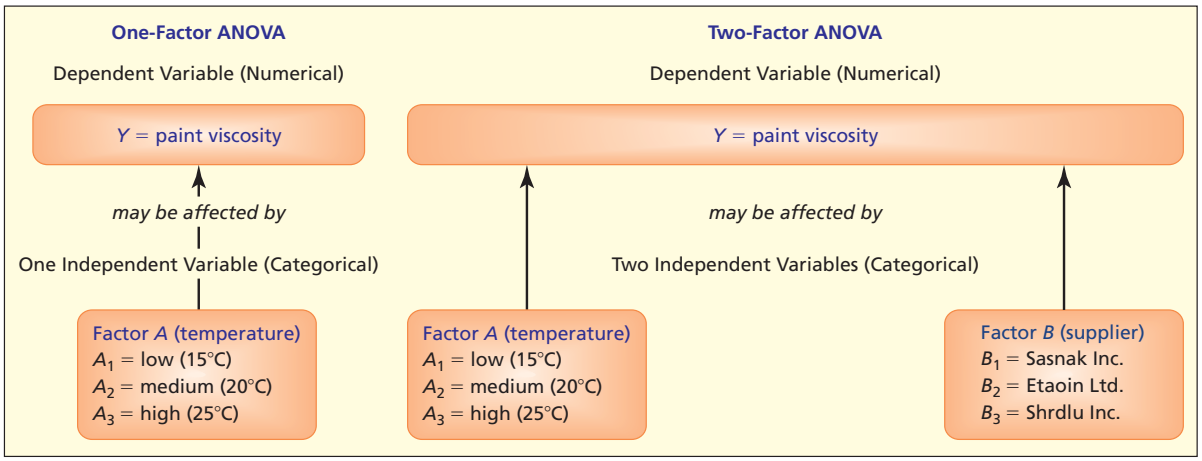
Two factors: Viscosity $= f$(temperature, supplier)

ANOVA Calculations

ANOVA calculations usually are too tedious to do by calculator, so after we choose an ANOVA model and collect the data, we rely on software (e.g., Excel, MegaStat, MINITAB, SPSS) to do the calculations. In some applications (accounting, finance, human resources, marketing) large samples can easily be taken from existing records, while in others (engineering, manufacturing, computer systems) experimental data collection is so expensive that small samples are used. Large samples increase the power of the test, but power also depends on the

FIGURE 11.4

Several ANOVA models for paint viscosity



degree of variation in Y. Lowest power would be in small samples with high variation in Y, and conversely. Specialized software is needed to calculate power for ANOVA experiments.

ANOVA Assumptions

Analysis of variance assumes that the

- Observations on Y are independent.
- Populations being sampled are normal.
- Populations being sampled have equal variances.

Fortunately, ANOVA is somewhat robust to departures from the normality and equal variance assumptions. Later in this chapter, you will see tests for equal variances and normality.

11.2 ONE-FACTOR ANOVA (COMPLETELY RANDOMIZED MODEL)

Data Format

If we are only interested in comparing the means of c groups (*treatments* or *factor levels*), we have a ***one-factor ANOVA.**** This is by far the most common ANOVA model that covers many business problems. The one-factor ANOVA is usually viewed as a comparison between several columns of data, although the data could also be presented in rows. Table 11.1 illustrates the data format for a one-factor ANOVA with c treatments, denoted $A_1, A_2, \ldots, A_c$. The group means are $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_c$.

TABLE 11.1
Format of One-Factor ANOVA Data

One-Factor ANOVA: Data in Columns

A_1	A_2	. . .	A_c
y_{11}	y_{12}	. . .	y_{1c}
y_{21}	y_{22}	. . .	y_{2c}
y_{31}	y_{32}	. . .	y_{3c}
. . .	. . .	. . .	. . .
etc.	etc.	. . .	etc.
n_1 obs.	n_2 obs.	. . .	n_c obs.
$\bar{y}_1$	$\bar{y}_2$	. . .	$\bar{y}_c$

One-Factor ANOVA: Data in Rows

A_1	y_{11}	y_{21}	y_{31}	. . .	etc.	n_1 obs.	$\bar{y}_1$
A_2	y_{12}	y_{22}	y_{32}	. . .	etc.	n_2 obs.	$\bar{y}_2$
. . .				. . .		. . .	. . .
A_c	y_{1c}	y_{2c}	y_{3c}	. . .	etc.	n_c obs.	$\bar{y}_c$

*If subjects (or individuals) are assigned randomly to treatments, then we call this the *completely randomized model*.

Within treatment j we have n_j observations on Y. Sample sizes within each treatment do *not* need to be equal, although there are advantages to having balanced sample sizes. The total number of observations is the sum of the sample sizes for each treatment:

$$n = n_1 + n_2 + \cdots + n_c \tag{11.1}$$

Hypotheses to Be Tested

The question of interest is whether the mean of Y varies from treatment to treatment. The hypotheses to be tested are

H_0: $\mu_1 = \mu_2 = \cdots = \mu_c$ (all the means are equal)

H_1: Not all the means are equal (at least one mean is different)

Since one-factor ANOVA is a generalization of the test for equality of two means, why not just compare all possible pairs of means by using repeated two-sample t tests (as in Chapter 10)? Consider our experiment comparing the four manufacturing plant average defect rates. To compare pairs of plant averages we would have to perform six different t tests. If each t test has a Type I error probability equal to .05, then the probability that at least one of those tests results in a Type I error is $1 - (.95)^6 = .2649$. ANOVA tests all the means *simultaneously* and therefore does not inflate our Type I error.

One-Factor ANOVA as a Linear Model

An equivalent way to express the one-factor model is to say that observations in treatment j came from a population with a common mean (μ) plus a treatment effect (A_j) plus random error (ε_{ij}):

$$y_{ij} = \mu + A_j + \varepsilon_{ij} \qquad j = 1, 2, \ldots, c \quad \text{and} \quad i = 1, 2, \ldots, n_j \tag{11.2}$$

The random error is assumed to be normally distributed with zero mean and the same variance for all treatments. If we are interested only in what happens to the response for the particular **levels** of the factor that were selected (a ***fixed-effects model***), then the hypotheses to be tested are

H_0: $A_1 = A_2 = \cdots = A_c = 0$ (all treatment effects are zero)

H_1: Not all A_j are zero (some treatment effects are nonzero)

If the null hypothesis is true ($A_j = 0$ for all j), then knowing that an observation x came from treatment j does not help explain the variation in Y and the ANOVA model collapses to

$$y_{ij} = \mu + \varepsilon_{ij} \tag{11.3}$$

Group Means

The *mean of each group* is calculated in the usual way by summing the observations in the treatment and dividing by the sample size:

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \tag{11.4}$$

The *overall sample mean* or *grand mean* $\bar{y}$ can be calculated either by summing *all* the observations and dividing by n or by taking a weighted average of the c sample means:

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{c} \sum_{i=1}^{n_j} y_{ij} = \frac{1}{n} \sum_{j=1}^{c} n_j \, \bar{y}_j \tag{11.5}$$

Partitioned Sum of Squares

To understand the logic of ANOVA, consider that for a given observation y_{ij} the following relationship must hold (on the right-hand side we just add and subtract $\bar{y}_j$):

$$(y_{ij} - \bar{y}) = (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j) \tag{11.6}$$

This says that any deviation of an observation from the grand mean $\bar{y}$ may be expressed in two parts: the deviation of the column mean $(\bar{y}_j)$ from the grand mean $(\bar{y})$, or *between* treatments, and the deviation of the observation (y_{ij}) from its own column mean $(\bar{y}_j)$, or *within* treatments. We can show that this relationship also holds for *sums* of squared deviations, yielding the ***partitioned sum of squares:***

$$\textbf{(11.7)} \qquad \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2 = \sum_{j=1}^{c} n_j(\bar{y}_j-\bar{y})^2 + \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$$

This important relationship may be expressed simply as

$$\textbf{(11.8)} \qquad SST = SSA + SSE \qquad \text{(partitioned sum of squares)}$$

Partitioned Sum of Squares

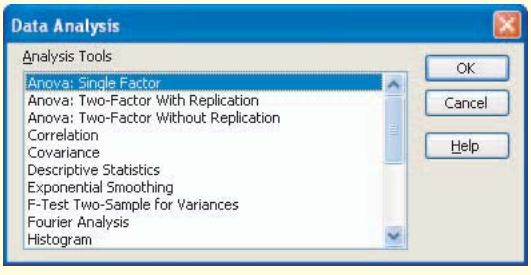
Sum of Squares Total (*SST*)	=	Sum of Squares Between Treatments (*SSA*) ↑ Explained by Factor *A*	+	Sum of Squares Within Treatments (*SSE*) ↑ Unexplained Random Error

If the treatment means do not differ greatly from the grand mean, *SSA* will be small and *SSE* will be large (and conversely). The sums *SSA* and *SSE* may be used to test the hypothesis that the treatment means differ from the grand mean. However, we first divide each sum of squares by its *degrees of freedom* (to adjust for group sizes). The *test statistic* is the ratio of the resulting ***mean squares.*** These calculations can be arranged in the tabular format shown in Table 11.2.

TABLE 11.2
One-Factor ANOVA Table

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Statistic*
Treatment (between groups)	$SSA = \sum_{j=1}^{c} n_j(\bar{y}_j-\bar{y})^2$	$c-1$	$MSA = \frac{SSA}{c-1}$	$F = \frac{MSA}{MSE}$
Error (within groups)	$SSE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$	$n-c$	$MSE = \frac{SSE}{n-c}$	
Total	$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2$	$n-1$		

FIGURE 11.5
Excel's ANOVA menu



The ANOVA calculations are mathematically simple but involve tedious sums. These calculations are almost always done on a computer.* For example, Excel's one-factor ANOVA menu using Tools > Data Analysis is shown in Figure 11.5. MegaStat uses a similar menu.

*Detailed examples of ANOVA calculations can be found in the case studies in *LearningStats* Unit 11.

Test Statistic

At the beginning of this chapter we described the variation in Y as consisting of explained variation and unexplained variation. To test whether the independent variable explains a significant proportion of the variation in Y, we need to compare the explained (due to treatments) and unexplained (due to error) variation. Recall that the F distribution describes the *ratio of two variances*. Therefore it makes sense that the ANOVA test statistic is the *F test statistic.* The F statistic is the ratio of the variance due to treatments to the variance due to error. *MSA* is the mean square due to treatments and *MSE* is the mean square within treatments. Equation 11.9 shows the F statistic and its degrees of freedom.

$$F = \frac{MSA}{MSE} = \frac{\left(\frac{SSA}{c-1}\right) \leftarrow \text{d.f.}_1 = c - 1 \text{ (numerator)}}{\left(\frac{SSE}{n-c}\right) \leftarrow \text{d.f.}_2 = n - c \text{ (denominator)}} \tag{11.9}$$

If there is little difference among treatments, we would expect *MSA* to be near zero because the treatment means $\bar{y}_j$ would be near the overall mean $\bar{y}$. Thus, when F is near zero we would not expect to reject the hypothesis of equal group means. The larger the F statistic, the more we are inclined to reject the hypothesis of equal means. But how large must F be to convince us that the means differ? Just as with a z test or a t test, we need a *decision rule*.

Decision Rule

The F distribution is a right-skewed distribution that starts at zero (F cannot be negative since variances are sums of squares) and has no upper limit (since the variances could be of any magnitude). For ANOVA, the F test is a right-tailed test. For a given level of significance α, we can use Appendix F to obtain the right-tail critical value of F. Alternatively, we can use Excel's function =FINV(α,df$_1$,df$_2$). The decision rule is illustrated in Figure 11.6. This critical value is denoted $F_{\text{df}_1,\text{df}_2}$ or $F_{c-1,n-c}$.

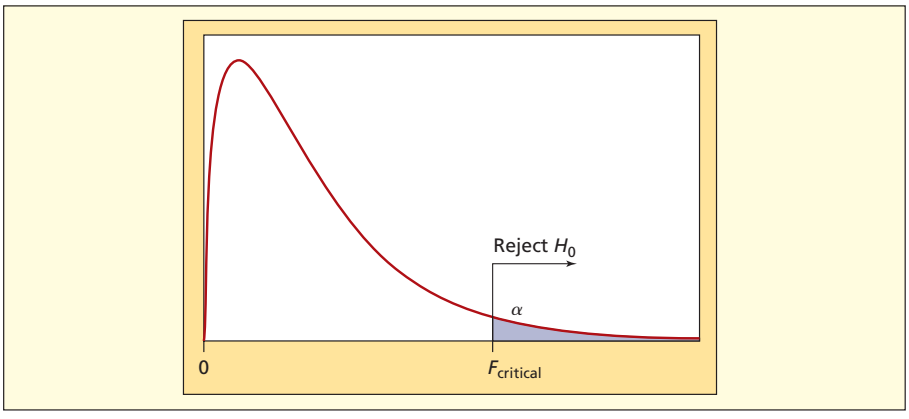


FIGURE 11.6
Decision rule for an F test

EXAMPLE

Carton Packing

A cosmetics manufacturer's regional distribution center has four workstations that are responsible for packing cartons for shipment to small retailers. Each workstation is staffed by two workers. The task involves assembling each order, placing it in a shipping carton, inserting packing material, taping the carton, and placing a computer-generated shipping label on each carton. Generally, each station can pack 200 cartons a day, and often more. However, there is variability, due to differences in orders, labels, and cartons. Table 11.3 shows the

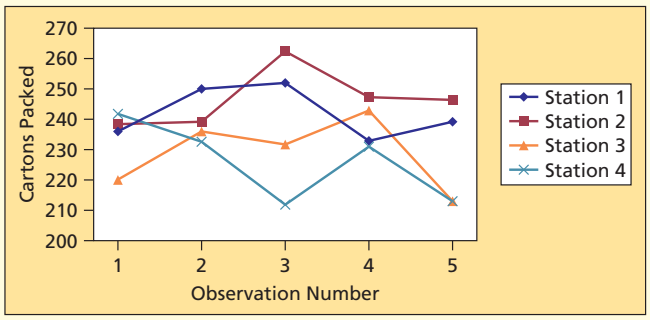
number of cartons packed per day during a recent week. Is the variation among stations within the range attributable to chance, or do these samples indicate actual differences in the means?

TABLE 11.3 Number of Cartons Packed Cartons

	Station 1	*Station 2*	*Station 3*	*Station 4*
	236	238	220	241
	250	239	236	233
	252	262	232	212
	233	247	243	231
	239	246	213	213
Sum	1,210	1,232	1,144	1,130
Mean	242.0	246.4	228.8	226.0
St. Dev.	8.515	9.607	12.153	12.884
n	5	5	5	5

As a preliminary step, we plot the data (Figure 11.7) to check for any time pattern and just to visualize the data. We see some potential differences in means, but no obvious time pattern (otherwise we would have to consider observation order as a second factor). We proceed with the hypothesis test.

FIGURE 11.7
Plot of the data



Step 1: State the Hypotheses

The hypotheses to be tested are

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (the means are the same)

H_1: Not all the means are equal (at least one mean is different)

Step 2: State the Decision Rule

There are $c = 4$ groups and $n = 20$ observations, so degrees of freedom for the F test are

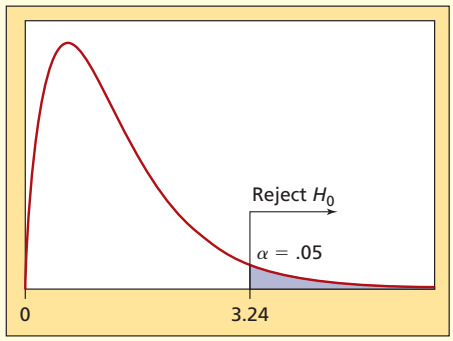
Numerator: $\text{d.f.}_1 = c - 1 = 4 - 1 = 3$ (between treatments, factor)

Denominator: $\text{d.f.}_2 = n - c = 20 - 4 = 16$ (within treatments, error)

We will use $\alpha = .05$ for the test. The 5 percent right-tail critical value from Appendix F is $F_{3,16} = 3.24$. Instead of Appendix F we could use Excel's function =FINV(0.05,3,16) which yields $F_{.05} = 3.238872$. This decision rule is illustrated in Figure 11.8.

FIGURE 11.8

F test using $\alpha = .05$ with $F_{3,16}$



Appendix F: Critical Values for $F_{.05}$

This table shows the critical value that defines the 5% right-tail area for the stated degrees of freedom.

Denominator Degrees of Freedom (ν_2)	Numerator Degrees of Freedom (ν_1) 1	2	3	4	5	6	7	8	9	10	12
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28

Step 3: Perform the Calculations

Using Excel for the calculations, we obtain the results shown in Figure 11.9. You can specify the desired level of significance (Excel's default is $\alpha = .05$). Note that Excel labels *SSA* "between groups" and *SSE* "within groups." This is an intuitive and attractive way to describe the variation.

FIGURE 11.9

Excel's one-factor ANOVA results

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Station 1	5	1210	242	72.5
Station 2	5	1232	246.4	92.3
Station 3	5	1144	228.8	147.7
Station 4	5	1130	226	166

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1479.2	3	493.0667	4.121769	0.024124	3.238872
Within Groups	1914	16	119.625			
Total	3393.2	19				

Step 4: Make the Decision

Since the test statistic $F = 4.12$ exceeds the critical value $F_{.05} = 3.24$, we can reject the hypothesis of equal means. Since Excel gives the p-value, you don't actually need Excel's critical value. The p-value ($p = .024124$) is less than the level of significance ($\alpha = .05$) which confirms that we should reject the hypothesis of equal treatment means. For comparison, Figure 11.10 shows MegaStat's ANOVA table for the same data. The results are the same, although MegaStat rounds things off, highlights significant p-values, and gives standard deviations instead of variances for each treatment.

FIGURE 11.10

MegaStat's one-factor ANOVA results

One factor ANOVA

Mean	*n*	*Std. Dev*	
242.0	5	8.51	Station 1
246.4	5	9.61	Station 2
228.8	5	12.15	Station 3
226.0	5	12.88	Station 4
235.8	20	13.36	Total

ANOVA table

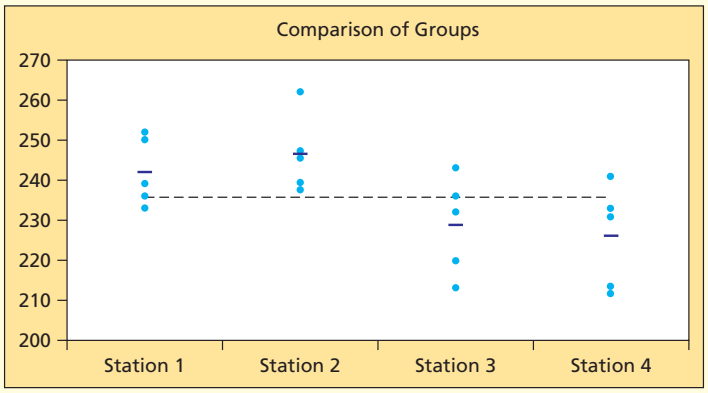
Source	*SS*	*df*	*MS*	*F*	*p-value*
Treatment	1,479.20	3	493.067	4.12	.0241
Error	1,914.00	16	119.625		
Total	3,393.20	19			

MegaStat provides additional insights by showing a dot plot of observations by group, shown in Figure 11.11. The display includes group means (shown as short horizontal tick marks) and the overall mean (shown as a dashed line). The dot plot suggests that stations 3 and 4 have means below the overall mean, while stations 1 and 2 are above the overall mean.

FIGURE 11.11

Dot plot of four samples

Cartons



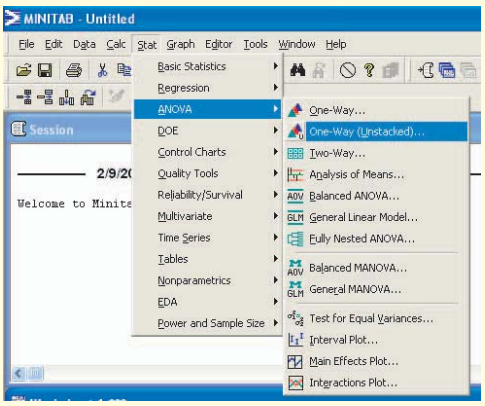
Using MINITAB

MINITAB's output, shown in Figure 11.12, is the same as Excel's except that MINITAB rounds off the results and displays a confidence interval for each group mean, an attractive feature.* In our carton example, the confidence intervals overlap, except possibly stations 2 and 4.

*MINITAB and most other statistical packages prefer the data in *stacked* format. Each variable has its own column (e.g., column one contains all the *Y* values, while column two contains group labels like "Station 1"). MINITAB will convert *unstacked* data to *stacked* data for one-factor ANOVA, but not for other ANOVA models. See *LearningStats* Unit 11 for examples of *stacked* versus *unstacked* data.

FIGURE 11.12

MINITAB's one-factor ANOVA



One-way ANOVA: Station 1, Station 2, Station 3, Station 4

```
Source  DF    SS   MS     F      P
Factor   3  1479  493  4.12  0.024
Error   16  1914  120
Total   19  3393

S = 10.94   R-Sq = 43.59%   R-Sq(adj) = 33.02%

                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level      N    Mean  StDev   +---------+---------+---------+---------
Station 1  5  242.00   8.51                (--------*-------)
Station 2  5  246.40   9.61                   (-------*--------)
Station 3  5  228.80  12.15      (--------*-------)
Station 4  5  226.00  12.88    (-------*--------)
                               +---------+---------+---------+---------
                              216       228       240       252

Pooled StDev = 10.94
```

SECTION EXERCISES

Instructions: For each data set: (a) State the hypotheses. (b) Use Excel's Tools > Data Analysis (or MegaStat or MINITAB) to perform the one-factor ANOVA, using $\alpha = .05$. (c) State your conclusion about the population means. Was the decision close? (d) Interpret the p-value carefully. (e) Include a plot of the data for each group if you are using MegaStat, and confidence intervals for the group means if you are using MINITAB. What do the plots show?

11.1 Scrap rates per thousand (parts whose defects cannot be reworked) are compared for 5 randomly selected days at three plants. Does the data prove a significant difference in mean scrap rates? **ScrapRate**

Scrap Rate (Per Thousand Units)

Plant A	*Plant B*	*Plant C*
11.4	11.1	10.2
12.5	14.1	9.5
10.1	16.8	9.0
13.8	13.2	13.3
13.7	14.6	5.9

11.2 One particular morning, the length of time spent in the examination rooms is recorded for each patient seen by each physician at an orthopedic clinic. Does the data prove a significant difference in mean times? **Physicians**

Time in Examination Rooms (minutes)

Physician 1	*Physician 2*	*Physician 3*	*Physician 4*
34	33	17	28
25	35	30	33
27	31	30	31
31	31	26	27
26	42	32	32
34	33	28	33
21		26	40
		29	

11.3 Semester GPAs are compared for seven randomly chosen students in each class level at Oxnard University. Does the data prove a significant difference in mean GPAs? GPA1

GPA for Randomly Selected Students in Four Business Majors			
Accounting	*Finance*	*Human Resources*	*Marketing*
2.48	3.16	2.93	3.54
2.19	3.01	2.89	3.71
2.62	3.07	3.48	2.94
3.15	2.88	3.33	3.46
3.56	3.33	3.53	3.50
2.53	2.87	2.95	3.25
3.31	2.85	3.58	3.20

11.4 Sales of *People* magazine are compared over a 5-week period at four Borders outlets in Chicago. Does the data prove a significant difference in mean weekly sales? Magazines

Weekly Sales			
Store 1	*Store 2*	*Store 3*	*Store 4*
102	97	89	100
106	77	91	116
105	82	75	87
115	80	106	102
112	101	94	100

11.3 MULTIPLE COMPARISONS

Tukey's Test

In Figure 11.12, we naturally want to compare the confidence intervals to see whether they overlap. In so doing, we are trying to answer the question, Which means differ significantly? However, to maintain the desired overall probability of Type I error, we need to create a *simultaneous confidence interval* for the difference of means and then see which pairs exclude zero. For c groups, there are $c(c-1)/2$ distinct pairs of means to be compared.

Several ***multiple comparison*** tests are available. Their logic is similar. We will discuss only one, called ***Tukey's studentized range test*** (sometimes called the *HSD* or "honestly significant difference" test). It has good power and is widely used. We will refer to it as *Tukey's test,* named for statistician John Wilder Tukey (1915–2000). This test is available in most statistical packages (but not in Excel's Tools > Data Analysis). It is a two-tailed test for equality of paired means from c groups compared simultaneously and is a natural follow-up when the results of the one-factor ANOVA test show a significant difference in at least one mean. The hypotheses are

$$H_0: \mu_j = \mu_k$$
$$H_1: \mu_j \neq \mu_k$$

The decision rule is

(11.10)
$$\text{Reject } H_0 \text{ if } \frac{|\bar{y}_j - \bar{y}_k|}{\sqrt{MSE\left[\frac{1}{n_j} + \frac{1}{n_k}\right]}} > T_\alpha$$

where $T_\alpha = 0.707 q_{c,n-c}$ and $q_{c,n-c}$ is a critical value of the *studentized range* for the desired level of significance. Table 11.4 shows 5 percent critical values of $q_{c,n-c}$. If the desired degrees of freedom cannot be found, we could interpolate or better yet rely on a computer package like MegaStat to provide the exact critical value. We take *MSE* directly from the ANOVA calculations (see Table 11.2).

TABLE 11.4
Upper 5 Percent Points of Studentized Range

Source: E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians,* 3rd ed. (Oxford University Press, 1970), p. 192. Copyright © 1970 Oxford University Press. Used with permission.

	Numerator d.f.								
Denominator d.f.	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.36	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47

We will illustrate the Tukey test for the carton-packing data. We assume that a one-factor ANOVA has already been performed and the results showed that at least one mean was significantly different. We will use the *MSE* from the ANOVA. For the carton-packing data there are 4 groups and 20 observations, so $c = 4$ and $n - c = 20 - 4 = 16$. From Table 11.4 we must interpolate between $q_{4,15} = 4.08$ and $q_{4,20} = 3.96$ to get $q_{4,16} = 4.056$ so the approximate critical value is $T_\alpha = (0.707)(4.056) = 2.87$ and the decision rule for any pair of means is

$$\text{Reject } H_0 \text{ if } \frac{|\bar{y}_j - \bar{y}_k|}{\sqrt{MSE\left[\frac{1}{n_j} + \frac{1}{n_k}\right]}} > 2.87$$

There may be a different decision rule for every pair of cities unless the sample sizes n_j and n_k are identical (in our example, the group sizes are the same). For example, to compare groups 2 and 4 the test statistic is

$$\frac{|\bar{y}_2 - \bar{y}_4|}{\sqrt{MSE\left[\frac{1}{n_2} + \frac{1}{n_4}\right]}} = \frac{|246.4 - 226.0|}{\sqrt{119.625\left[\frac{1}{5} + \frac{1}{5}\right]}} = 2.95$$

Since 2.95 exceeds 2.87, we reject the hypothesis of equal means for station 2 and station 4. We conclude that there is a significant difference between the mean output of stations 2 and 4. A similar test must be performed for every possible pair of means.

Using MegaStat

MegaStat includes all six possible comparisons of means, as shown in Figure 11.13. Only stations 2 and 4 differ at $\alpha = .05$. However, if we use the independent sample t test (as in Chapter 10) shown in MegaStat's lower table, we obtain two p-values smaller than $\alpha = .05$ (stations 1, 4 and stations 2, 3) and one that is below $\alpha = .01$ (stations 2, 4). This demonstrates that a *simultaneous* Tukey t test is not the same as comparing individual pairs of means. As noted in section 11.2, using multiple independent t tests results in a greater probability of making a Type I error. An attractive feature of MegaStat's Tukey test is that it highlights significant results using color-coding for $\alpha = .05$ and $\alpha = .01$. Note that MegaStat's Tukey critical value T_α is slightly more accurate than our interpolated $T_\alpha = 2.87$ from Table 11.4.

FIGURE 11.13

MegaStat's Tukey tests and independent sample *t* tests Cartons

Tukey simultaneous comparison *t*-values (d.f. = 16)

		Station 4 226.0	Station 3 228.8	Station 1 242.0	Station 2 246.4
Station 4	226.0				
Station 3	228.8	0.40			
Station 1	242.0	2.31	1.91		
Station 2	246.4	2.95	2.54	0.64	

critical values for experimentwise error rate:

0.05	2.86
0.01	3.67

p-values for pairwise *t*-tests

		Station 4 226.0	Station 3 228.8	Station 1 242.0	Station 2 246.4
Station 4	226.0				
Station 3	228.8	.6910			
Station 1	242.0	.0344	.0745		
Station 2	246.4	.0094	.0217	.5337	

SECTION EXERCISES

Instructions: Use MegaStat, MINITAB, or another software package to perform Tukey's test for significant pairwise differences. Perform the test using both the 5 percent and 1 percent levels of significance.

11.5 Refer to Exercise 11.1. Which pairs of mean scrap rates differ significantly (3 plants)? ScrapRate

11.6 Refer to Exercise 11.2. Which pairs of mean examination times differ significantly (4 physicians)? Physicians

11.7 Refer to Exercise 11.3. Which pairs of mean GPAs differ significantly (4 majors)? GPA1

11.8 Refer to Exercise 11.4. Which pairs of mean weekly sales differ significantly (4 stores)? Magazines

11.4 TESTS FOR HOMOGENEITY OF VARIANCES (OPTIONAL)

ANOVA Assumptions

Analysis of variance assumes that observations on the response variable are from normally distributed populations that have the same variance. We have noted that few populations meet these requirements perfectly and unless the sample is quite large, a test for normality is impractical. However, we can easily test the assumption of ***homogeneous*** (equal) ***variances.*** Although the one-factor ANOVA test is only slightly affected by inequality of variance when group sizes are equal or nearly so, it is still a good idea to test this assumption. In general, surprisingly large differences in variances must exist to conclude that the population variances are unequal.

Hartley's F_{max} Test

If we had only two groups, we could use the *F* test you learned in Chapter 10 to compare the variances. But for *c* groups, a more general test is required. One such test is ***Hartley's F_{max} test,***

named for statistician H. O. Hartley (1912–1980). The hypotheses are

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_c^2$

H_1: Not all the σ_j^2 are equal

The test statistic is the ratio of the largest sample variance to the smallest sample variance:

$$F_{\max} = \frac{s^2_{\max}}{s^2_{\min}} \tag{11.11}$$

Critical values of $F_{\max}$ may be found in Table 11.5 using degrees of freedom given by

Numerator: $\text{d.f.}_1 = c$

Denominator: $\text{d.f.}_2 = \dfrac{n}{c} - 1$

where n is the total number of observations. This test assumes equal group sizes, so d.f._2 would be an integer. For group sizes that are not drastically unequal, this procedure will still be approximately correct, using the next lower integer if d.f._2 is not an integer. Note that this is *not* the same table as the F table you have used previously.

TABLE 11.5
Critical 5 Percent Values of Hartley's $F_{\max} = s^2_{\max}/s^2_{\min}$

Source: E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, 3rd. ed. (Oxford University Press, 1970), p. 202. Copyright © 1970 Oxford University Press. Used with permission.

	Numerator d.f.								
Denominator d.f.	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	39.0	87.5	142	202	266	333	403	475	550
3	15.4	27.8	39.2	50.7	62.0	72.9	83.5	93.9	104
4	9.60	15.5	20.6	25.2	29.5	33.6	37.5	41.1	44.6
5	7.15	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5
6	5.82	8.38	10.4	12.1	13.7	15.0	16.3	17.5	18.6
7	4.99	6.94	8.44	9.7	10.8	11.8	12.7	13.5	14.3
8	4.43	6.00	7.18	8.12	9.03	9.78	10.5	11.1	11.7
9	4.03	5.34	6.31	7.11	7.80	8.41	8.95	9.45	9.91
10	3.72	4.85	5.67	6.34	6.92	7.42	7.87	8.28	8.66
12	3.28	4.16	4.79	5.30	5.72	6.09	6.42	6.72	7.00
15	2.86	3.54	4.01	4.37	4.68	4.95	5.19	5.40	5.59
20	2.46	2.95	3.29	3.54	3.76	3.94	4.10	4.24	4.37
30	2.07	2.40	2.61	2.78	2.91	3.02	3.12	3.21	3.29
60	1.67	1.85	1.96	2.04	2.11	2.17	2.22	2.26	2.30
∞	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

EXAMPLE

Carton Packing: Tukey Test Cartons

Using the carton-packing data in Table 11.3, there are 4 groups and 20 total observations, so we have

Numerator: $\text{d.f.}_1 = c = 4$

Denominator: $\text{d.f.}_2 = n/c - 1 = 20/4 - 1 = 5 - 1 = 4$

From Table 11.5 we choose the critical value $F_{\max} = 20.6$ using $\text{d.f.}_1 = 4$ and $\text{d.f.}_2 = 4$. The sample statistics (from Excel) for our workstations are

Work Station	n	Mean	Variance
Station 1	15	242.0	72.5
Station 2	17	246.4	92.3
Station 3	15	228.8	147.7
Station 4	12	226.0	166.0

The test statistic is

$$F_{\max} = \frac{s^2_{\max}}{s^2_{\min}} = \frac{166.0}{72.5} = 2.29$$

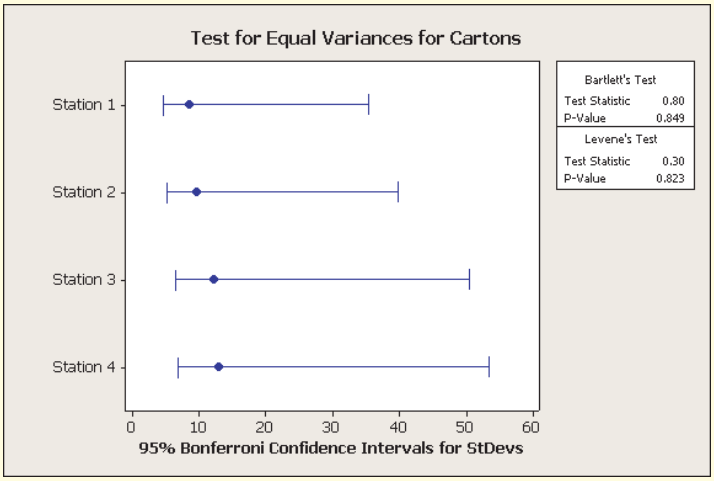
In this case, we cannot reject the hypothesis of equal variances. Indeed, Table 11.5 makes it clear that unless the sample size is very large, the variance ratio would have to be quite large to reject the hypothesis of equal population variances. If the F_{max} test is significant, we prefer an alternative* to one-factor ANOVA, which does not require this assumption.

Levene's Test

The F_{max} test relies on the assumption of normality in the populations from which the sample observations are drawn. A more robust alternative is ***Levene's test,*** which does not assume a normal distribution. This test requires a computer package. It is not necessary to discuss the computational procedure except to say that Levene's test is based on the distances of the observations from their sample *medians* rather than their sample *means*. As long as you know how to interpret a *p*-value, Levene's test is easy to use. Figure 11.14 shows MINITAB's output for the test of homogeneity of variance for the carton-packing data using Levene's test, with the added attraction of confidence intervals for each population standard deviation. Since the confidence intervals overlap and the *p*-value (.823) is large, we cannot reject the hypothesis of equal population variances. This confirms that the one-factor ANOVA procedure was appropriate for the carton-packing data.

FIGURE 11.14

MINITAB's equal-variance test **Cartons**



SECTION EXERCISES

Instructions: For each data set, use Hartley's F_{max} test to test the hypothesis of equal variances, using the 5 percent table of critical values from this section and the largest and smallest sample variances from your previous ANOVA. Alternatively, if you have access to MINITAB or another software package, perform Levene's test for equal group variances, discuss the *p*-value, and interpret the graphical display of confidence intervals for standard deviations.

11.9 Refer to Exercise 11.1. Are the population variances the same for scrap rates (3 plants)? **ScrapRate**

11.10 Refer to Exercise 11.2. Are the population variances the same for examination times (4 physicians)? **Physicians**

11.11 Refer to Exercise 11.3. Are the population variances the same for the GPAs (4 majors)? **GPA1**

11.12 Refer to Exercise 11.4. Are the population variances the same for weekly sales (4 stores)? **Magazines**

*For one-factor ANOVA, we could use the nonparametric *Kruskal-Wallis* test described in Chapter 15.

Mini Case 11.1

Hospital Emergency Arrivals

To plan its staffing schedule, a large urban hospital examined the number of arrivals per day over a 3-month period, as shown in Table 11.6. Each day has 13 observations except Tuesday, which has 14. Data are shown in rows rather than in columns to make a more compact table.

TABLE 11.6 Number of Emergency Arrivals by Day of the Week Emergency

Mon	188	175	208	176	179	184	191	194	174	191	198	213	217	
Tue	174	167	165	164	169	164	150	175	178	164	202	175	191	180
Wed	177	169	180	173	182	181	168	165	174	175	174	177	182	
Thu	170	164	190	169	164	170	153	150	156	173	177	183	208	
Fri	177	167	172	185	185	170	170	193	212	171	175	177	209	
Sat	162	184	173	175	144	170	163	157	181	185	199	203	198	
Sun	182	176	183	228	148	178	175	174	188	179	220	207	193	

We perform a one-factor ANOVA to test the model $Arrivals = f(Weekday)$. The single factor (*Weekday*) has 7 treatments. The Excel results, shown in Figure 11.15, indicate that *Weekday* does have a significant effect on *Arrivals,* since the test statistic $F = 3.257$ exceeds the 5 percent critical value $F_{6,85} = 2.207$. The p-value (.006) indicates that a test statistic this large would arise by chance only about 6 times in 1,000 samples if the hypothesis of equal daily means were true.

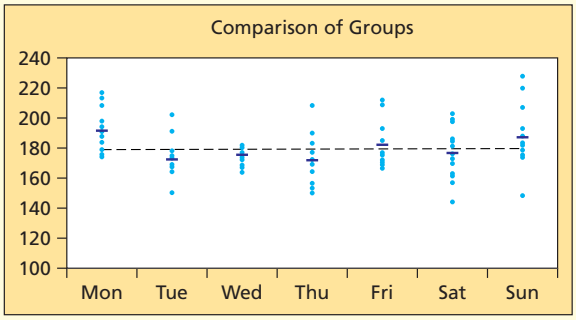
FIGURE 11.15 One-factor ANOVA for emergency arrivals and sample plot

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Mon	13	2488	191.385	206.423
Tue	14	2418	172.714	164.220
Wed	13	2277	175.154	29.808
Thu	13	2227	171.308	250.564
Fri	13	2363	181.769	216.692
Sat	13	2294	176.462	308.769
Sun	13	2431	187.000	445.667

ANOVA: Single Factor

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4514.283	6	752.3806	3.257899	0.006238	2.20723
Within Groups	19629.93	85	230.9404			
Total	24144.22	91				



The Tukey multiple comparison test (Figure 11.16) shows that the only pairs of *significantly* different means at $\alpha = .05$ are (*Mon, Tue*) and (*Mon, Thu*). In testing for equal variances, we get

FIGURE 11.16 MegaStat's Tukey test for $\mu_j - \mu_k$

Tukey simultaneous comparison *t*-values (d.f. = 84)

		Thu 171.3	Tue 172.2	Wed 175.2	Sat 176.5	Fri 181.8	Sun 187.0	Mon 191.4
Thu	171.3							
Tue	172.2	0.14						
Wed	175.2	0.64	0.50					
Sat	176.5	0.86	0.72	0.22				
Fri	181.8	1.75	1.61	1.10	0.89			
Sun	187.0	2.62	2.48	1.98	1.76	0.87		
Mon	191.4	3.35	3.21	2.71	2.49	1.61	0.73	

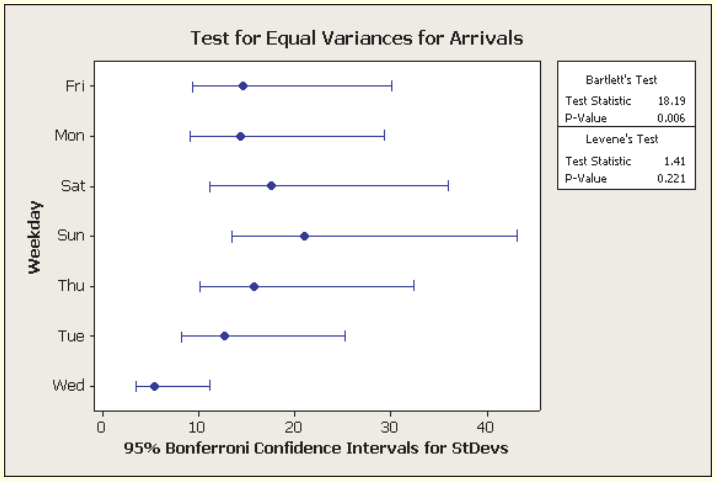
critical values for experimentwise error rate:

0.05	3.03
0.01	3.59

conflicting conclusions, depending on which test we use. Hartley's test gives $F_{max} = (445.667)/(29.808) = 14.95$, which exceeds the critical value $F_{7,12} = 6.09$ (note that *Wed* has a *very* small variance). But Levene's test for homogeneity of variances (Figure 11.17) has a *p*-value of .221, which at $\alpha = .05$ does not allow us to reject the equal-variance assumption that underlies the ANOVA test. When it is available, we prefer Levene's test because it does not depend on the assumption of normality.

FIGURE 11.17

MINITAB test for equal variances



11.5 TWO-FACTOR ANOVA WITHOUT REPLICATION (RANDOMIZED BLOCK MODEL)

Data Format

Suppose that two factors *A* and *B* may affect *Y*. One way to visualize this is to imagine a data matrix with *r* rows and *c* columns. Each row is a level of factor *A*, while each column is a level of factor *B*. Initially, we will consider the case where all levels of both factors occur, and each cell contains only one observation. In this ***two-factor ANOVA without replication*** (or *nonrepeated measures design*) each factor combination is observed exactly once. The mean of *Y* can be computed either across the rows or down the columns, as shown in Table 11.7. The grand mean $\bar{y}$ is the sum of all data values divided by the sample size *rc*.

TABLE 11.7

Format of Two-Factor ANOVA Data Set Without Replication

	Levels of Factor *B*				
Levels of Factor *A*	B_1	B_2	. . .	B_c	***Row Mean***
A_1	y_{11}	y_{12}	. . .	y_{1c}	$\bar{y}_{1.}$
A_2	y_{21}	y_{22}	. . .	y_{2c}	$\bar{y}_{2.}$
. . .	. . .	. . .	. . .	. . .	. . .
A_r	y_{r1}	y_{r2}	. . .	y_{rc}	$\bar{y}_{r.}$
Col Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	. . .	$\bar{y}_{.c}$	$\bar{y}$

For example, *Y* might be computer chip defects per thousand for three different deposition techniques (A_1, A_2, A_3) on four different types of silicon substrate (B_1, B_2, B_3, B_4) yielding a table with $3 \times 4 = 12$ cells. Each factor combination is a *treatment*. With only one observation per treatment, no interaction between the two factors is included.*

*There are not enough degrees of freedom to estimate an interaction unless the experiment is replicated.

Two-Factor ANOVA Model

Expressed in linear form, the two-factor ANOVA model is

$$y_{jk} = \mu + A_j + B_k + \varepsilon_{jk} \tag{11.12}$$

where

y_{jk} = observed data value in row j and column k
μ = common mean for all treatments
A_j = effect of row factor A ($j = 1, 2, \ldots, r$)
B_k = effect of column factor B ($k = 1, 2, \ldots, c$)
ε_{jk} = random error

The random error is assumed to be normally distributed with zero mean and the same variance for all treatments.

Hypotheses to Be Tested

If we are interested only in what happens to the response for the particular levels of the factors that were selected (a ***fixed-effects* model**) then the hypotheses to be tested are

Factor *A*

H_0: $A_1 = A_2 = \cdots = A_r = 0$ (row factor has no effect)
H_1: Not all the A_j are equal to zero (row factor has an effect)

Factor *B*

H_0: $B_1 = B_2 = \cdots = B_c = 0$ (column factor has no effect)
H_1: Not all the B_k are equal to zero (column factor has an effect)

If we are unable to reject either null hypothesis, all variation in Y is just a random disturbance around the mean μ:

$$y_{jk} = \mu + \varepsilon_{jk} \tag{11.13}$$

Randomized Block Model

A special terminology is used when only one factor is of research interest and the other factor is merely used to control for potential confounding influences. In this case, the two-factor ANOVA model with one observation per cell is sometimes called the ***randomized block model.*** In the randomized block model, it is customary to call the column effects *treatments* (as in one-factor ANOVA to signify that they are the effect of interest) while the row effects are called *blocks*.* For example, a North Dakota agribusiness might want to study the effect of four kinds of fertilizer (F_1, F_2, F_3, F_4) in promoting wheat growth (Y) on three soil types (S_1, S_2, S_3). To control for the effects of soil type, we could define three blocks (rows) each containing one soil type, as shown in Table 11.8. Subjects within each block (soil type) would be randomly assigned to the treatments (fertilizer).

TABLE 11.8
Format of Randomized Block Experiment: Two Factors

Block (Soil Type)	Treatment (Fertilizer)			
	F_1	F_2	F_3	F_4
S_1				
S_2				
S_3				

*In principle, either rows or columns could be the blocking factor, but it is customary to put the blocking factor in rows.

A randomized block model looks like a two-factor ANOVA and is computed exactly like a two-factor ANOVA. However, its interpretation by the researcher may resemble a one-factor ANOVA since only the column effects (treatments) are of interest. The blocks exist only to reduce variance. The effect of the blocks will show up in the hypothesis test, but is of no interest to the researcher as a separate factor. In short, the difference between a randomized block model and a standard two-way ANOVA model lies in the mind of the researcher. Since calculations for a randomized block design are identical to the two-factor ANOVA with one observation per cell, we will not call the row factor a "block" and the column factor a "treatment." Instead, we just call them *factor A* and *factor B*. Interpretation of the factors is not a mathematical issue. If only the column effect is of interest, you may call the column effect the "treatment."

Format of Calculation of Nonreplicated Two-Factor ANOVA

Calculations for the unreplicated two-factor ANOVA may be arranged as in Table 11.9. Degrees of freedom sum to $n - 1$. For a data set with r rows and c columns, notice that $n = rc$. The total sum of squares shown in Table 11.9 has three components:

(11.14) $$SST = SSA + SSB + SSE$$

where

SST = total sum of squared deviations about the mean

SSA = between rows sum of squares (effect of factor A)

SSB = between columns sum of squares (effect of factor B)

SSE = error sum of squares (residual variation)

SSE is a measure of unexplained variation. If SSE is relatively high, we would fail to reject the null hypothesis that the factor effects do not differ significantly from zero. Conversely, if SSE is relatively small, it is a sign that at least one factor is a relevant predictor of Y, and we would expect either SSA or SSB (or both) to be relatively large. Before doing the F test, each sum of squares must be divided by its degrees of freedom to obtain the *mean square.* Calculations are almost always done by a computer. For details of two-factor calculation methods see *LearningStats* Unit 11. There are case studies for each ANOVA.

TABLE 11.9 Format of Two-Factor ANOVA with One Observation per Cell

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Ratio*
Factor A (row effect)	$SSA = c\sum_{j=1}^{r}(\bar{y}_{j.} - \bar{y})^2$	$r - 1$	$MSA = \frac{SSA}{r-1}$	$F_A = \frac{MSA}{MSE}$
Factor B (column effect)	$SSB = r\sum_{k=1}^{c}(\bar{y}_{.k} - \bar{y})^2$	$c - 1$	$MSB = \frac{SSB}{c-1}$	$F_B = \frac{MSB}{MSE}$
Error	$SSE = \sum_{j=1}^{r}\sum_{k=1}^{c}(y_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y})^2$	$(r - 1)(c - 1)$	$MSE = \frac{SSE}{(c-1)(r-1)}$	
Total	$SST = \sum_{j=1}^{c}\sum_{k=1}^{r}(y_{jk} - \bar{y})^2$	$rc - 1$		

EXAMPLE

Vehicle Acceleration

Drivers expect a car to have good acceleration. A driver is coasting on the highway, with his foot off the accelerator. He steps on the gas to speed up. What is the peak acceleration to a final speed of 80 mph? Tests were carried out on one vehicle at 4 different initial speeds (10, 25, 40, 55 mph) and three different levels of rotation of accelerator pedal (5, 8, 10 degrees). The acceleration results are shown in Table 11.10. Does this sample show that the two experimental factors (pedal rotation, initial speed) are significant predictors of acceleration? Bear in mind that a different sample could yield different results; this is an *unreplicated* experiment.

TABLE 11.10 Maximum Acceleration Under Test Conditions Acceleration

	Initial Speed			
Pedal Rotation	***10 mph***	***25 mph***	***40 mph***	***55 mph***
5 degrees	0.35	0.19	0.14	0.10
8 degrees	0.37	0.28	0.19	0.19
10 degrees	0.42	0.30	0.29	0.23

Note: Maximum acceleration is measured as a fraction of acceleration due to gravity (32 ft./sec.^{2}).

Step 1: State the Hypotheses

It is helpful to assign short, descriptive variable names to each factor. The general form of the model is

$$Acceleration = f(PedalRotation, InitialSpeed)$$

Stated as a linear model:

$$y_{jk} = \mu + A_j + B_k + \varepsilon_{jk}$$

The hypotheses are

Factor *A (PedalRotation)*

H_0: $A_1 = A_2 = A_3 = 0$ (pedal rotation has no effect)

H_1: Not all the A_j are equal to zero

Factor *B (InitialSpeed)*

H_0: $B_1 = B_2 = B_3 = B_4 = 0$ (initial speed has no effect)

H_1: Not all the B_k are equal to zero

Step 2: State the Decision Rule

Each F test may require a different right-tail critical value because the numerator degrees of freedom depend on the number of factor levels, while denominator degrees of freedom (error *SSE*) are the same for all three tests:

Factor A: $\text{d.f.}_1 = r - 1 = 3 - 1 = 2$ ($r = 3$ pedal rotations)

Factor B: $\text{d.f.}_1 = c - 1 = 4 - 1 = 3$ ($c = 4$ initial speeds)

Error: $\text{d.f.}_2 = (r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$

From Appendix F, the 5 percent critical values in a right-tailed test (all ANOVA tests are right-tailed tests) are

$F_{2,6} = 5.14$ for factor A

$F_{3,6} = 4.76$ for factor B

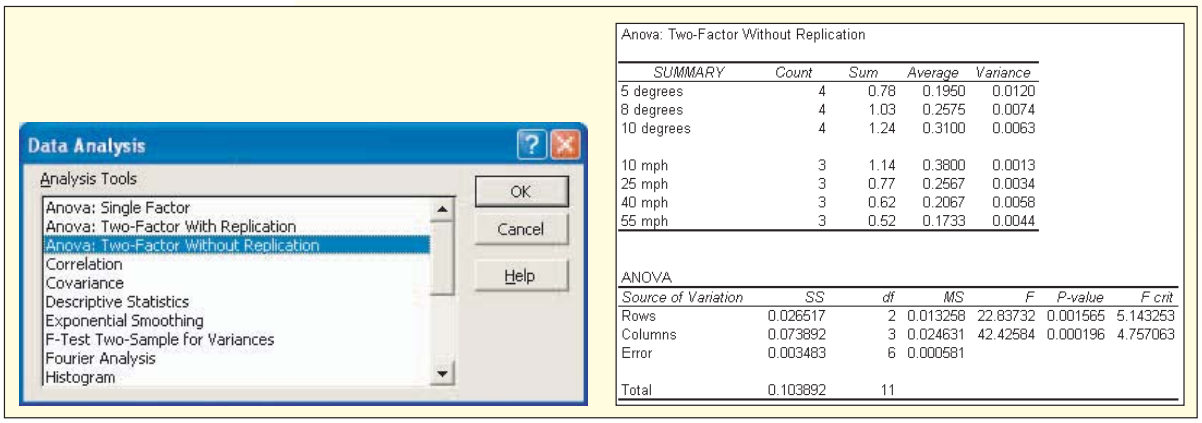
We will reject the null hypothesis (no factor effect) if the F test statistic exceeds the critical value.

Step 3: Perform the Calculations

Calculations are done by using Excel's Tools > Data Analysis. The menu and results are shown in Figure 11.18. There is a table of means and variances, followed by the ANOVA table.

FIGURE 11.18

Excel's ANOVA: two-factor without replication **Acceleration**



Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
5 degrees	4	0.78	0.1950	0.0120
8 degrees	4	1.03	0.2575	0.0074
10 degrees	4	1.24	0.3100	0.0063
10 mph	3	1.14	0.3800	0.0013
25 mph	3	0.77	0.2567	0.0034
40 mph	3	0.62	0.2067	0.0058
55 mph	3	0.52	0.1733	0.0044

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	0.026517	2	0.013258	22.83732	0.001565	5.143253
Columns	0.073892	3	0.024631	42.42584	0.000196	4.757063
Error	0.003483	6	0.000581			
Total	0.103892	11				

Step 4: Make the Decision

Since $F_A = 22.84$ (rows) exceeds $F_{2,6} = 5.14$, we see that factor A (pedal rotation) has a significant effect on acceleration. The p-value for pedal rotation is very small ($p = .001565$), which says that the F statistic is not due to chance. Similarly, $F_B = 42.43$ exceeds $F_{3,6} = 4.76$, so we see that factor B (initial speed) also has a significant effect on acceleration. Its tiny p-value (.000196) is unlikely to be a chance result. In short, we conclude that

- Acceleration is significantly affected by pedal rotation ($p = .001565$).
- Acceleration is significantly affected by initial speed ($p = .000196$).

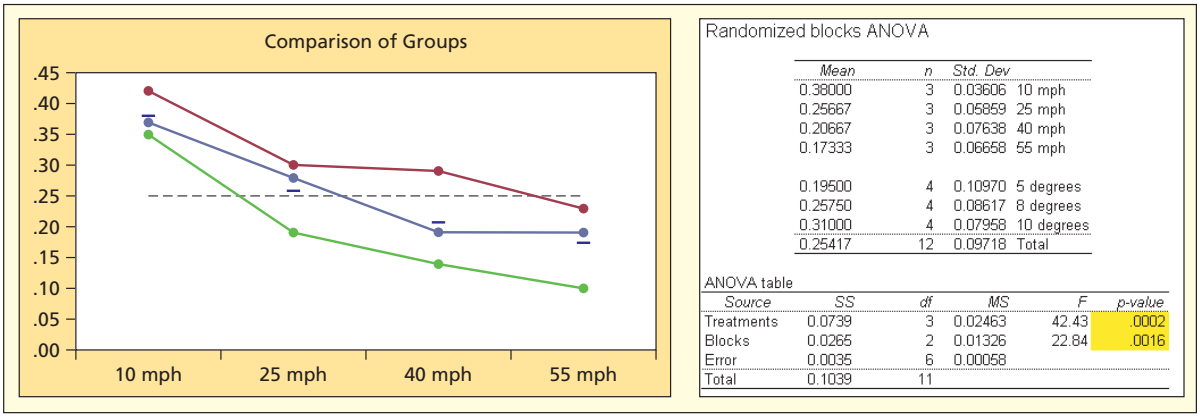
The p-values suggest that initial speed is a more significant predictor than pedal rotation, although both are highly significant. These results conform to your own experience. Maximum acceleration ("pushing you back in your seat") from a low speed or standing stop is greater than when you are driving down the freeway, and of course the harder you press the accelerator pedal, the faster you will accelerate. In fact, you might think of the pedal rotation as a blocking factor since its relationship to acceleration is tautological and of little research interest. Nonetheless, omitting pedal rotation and using a one-factor model would not be a correct model specification. Further, the engineers who did this experiment were actually interested in both effects.

Using MegaStat

Figure 11.19 shows MegaStat's dot plot and ANOVA table. The dot plot shows the column factor (presumed to be the factor of research interest) on the horizontal axis, while the row factor (presumed to be a blocking factor) is only used to define the line graphs. MegaStat rounds its ANOVA results more than Excel and highlights significant p-values. MegaStat does not provide critical F values, which are basically redundant since you have the p-values.

FIGURE 11.19

MegaStat's two-factor ANOVA (randomized block model) **Acceleration**



Randomized blocks ANOVA

Mean	n	Std. Dev	
0.38000	3	0.03606	10 mph
0.25667	3	0.05859	25 mph
0.20667	3	0.07638	40 mph
0.17333	3	0.06658	55 mph
0.19500	4	0.10970	5 degrees
0.25750	4	0.08617	8 degrees
0.31000	4	0.07958	10 degrees
0.25417	12	0.09718	Total

ANOVA table

Source	SS	df	MS	F	p-value
Treatments	0.0739	3	0.02463	42.43	.0002
Blocks	0.0265	2	0.01326	22.84	.0016
Error	0.0035	6	0.00058		
Total	0.1039	11			

Multiple Comparisons

Figure 11.20 shows MegaStat's Tukey simultaneous comparisons of the treatment pairs using a pooled variance. There are also *p*-values for corresponding independent two-sample *t* tests. However, MegaStat presents Tukey comparisons *only for the column factor* (the row factor is presumed merely to be a blocking factor). For this data, the Tukey *t* tests and independent sample *t* tests agree on all comparisons except for 25 mph versus 40 mph. (The *t* test shows a significant difference between 25 mph and 40 mph at the .05 level of significance.) Both tests show no significant difference in acceleration between 40 mph and 55 mph.

FIGURE 11.20

MegaStat's Tukey and independent sample *t* tests **Acceleration**

Tukey simultaneous comparison *t*-values (d.f. = 6)

	55 mph 0.17333	40 mph 0.20667	25 mph 0.25667	10 mph 0.38000
55 mph 0.17333				
40 mph 0.20667	1.69			
25 mph 0.25667	4.24	2.54		
10 mph 0.38000	10.50	8.81	6.27	

critical values for experimentwise error rate:

0.05	3.46
0.01	4.97

p-values for pairwise *t*-tests

	55 mph 0.17333	40 mph 0.20667	25 mph 0.25667	10 mph 0.38000
55 mph 0.17333				
40 mph 0.20667	.1411			
25 mph 0.25667	.0055	.0440		
10 mph 0.38000	4.87E-05	.0001	.0008	

Limitations of Two-Factor ANOVA Without Replication

When replication is impossible or extremely expensive, two-factor ANOVA without replication must suffice. For example, crash-testing of automobiles to estimate collision damage is very costly. However, whenever possible, there is a strong incentive to replicate the experiment to add power to the tests. Would different results have been obtained if the car had been tested

not once but several times at each speed? Or if several different cars had been tested? For testing acceleration, there would seem to be no major cost impediment to replication except the time and effort required to take the measurements. Of course, it could be argued that if the measurements of acceleration were careful and precise the first time, replication would be a waste of time. And yet, some random variation is found in any experiment. These are matters to ponder. But two-factor ANOVA *with replication* does offer advantages, as you will see.

SECTION EXERCISES

Instructions: For each data set: (a) State the hypotheses. If you are viewing this data set as a randomized block, which is the blocking factor, and why? (b) Use Excel's Tools > Data Analysis (or MegaStat or MINITAB) to perform the two-factor ANOVA without replication, using $\alpha = .05$. (c) State your conclusions about the treatment means. (d) Interpret the p-values carefully. (e) Include a plot of the data for each group if you are using MegaStat, or individual value plots if you are using MINITAB. What do the plots show?

11.13 Concerned about Friday absenteeism, management examined absenteeism rates for the last three Fridays in four assembly plants. Does this sample prove that there is a significant difference in treatment means? **Absences**

	Plant 1	*Plant 2*	*Plant 3*	*Plant 4*
March 4	19	18	27	22
March 11	22	20	32	27
March 18	20	16	28	26

11.14 Engineers are testing company fleet vehicle fuel economy (miles per gallon) performance by using different types of fuel. One vehicle of each size is tested. Does this sample prove that there is a significant difference in treatment means? **MPG2**

	87 Octane	*89 Octane*	*91 Octane*	*Ethanol 5%*	*Ethanol 10%*
Compact	27.2	30.0	30.3	26.8	25.8
Mid-Size	23.0	25.6	28.6	26.6	23.3
Full-Size	21.4	22.5	22.2	18.9	20.8
SUV	18.7	24.1	22.1	18.7	17.4

11.15 Five statistics professors are using the same textbook with the same syllabus and common exams. At the end of the semester, the department committee on instruction looked at average exam scores. Does this sample prove a significant difference in treatment means? **ExamScores**

	Prof. Argand	*Prof. Blague*	*Prof. Clagmire*	*Prof. Dross*	*Prof. Ennuyeux*
Exam 1	80.9	72.3	84.9	81.2	70.9
Exam 2	75.5	74.6	78.7	76.5	70.3
Exam 3	79.0	76.0	79.6	75.0	73.7
Final	69.9	78.0	77.8	74.1	73.9

11.16 A beer distributor is comparing quarterly sales of Coors Light (number of six-packs sold) at three convenience stores. Does this sample prove a significant difference in treatment means? **BeerSales**

	Store 1	Store 2	Store 3
Qtr 1	1,521	1,298	1,708
Qtr 2	1,396	1,492	1,382
Qtr 3	1,178	1,052	1,132
Qtr 4	1,730	1,659	1,851

Mini Case 11.2

Automobile Interior Noise Level

Most consumers prefer quieter cars. Table 11.11 shows interior noise level for five vehicles selected from tests performed by a popular magazine. Noise level (in decibels) was measured at idle, at 60 miles per hour, and under hard acceleration from 0 to 60 mph. For reference, 60 dB is a normal conversation, 75 dB is a typical vacuum cleaner, 85 dB is city traffic, 90 dB is a typical hair dryer, and 110 dB is a chain saw. Two questions may be asked: (1) Does noise level vary significantly among the vehicles? (2) Does noise level vary significantly with speed? If you wish to think of this as a randomized block experiment, the column variable (*vehicle type*) is the research question, while the row variable (*speed*) is the blocking factor.

TABLE 11.11 Interior Noise Levels in Five Randomly Selected Vehicles
NoiseLevel

Speed	Chrysler 300M	BMW 528i Sport Wagon	Ford Explorer Sport Trac	Chevy Malibu LS	Subaru Outback H6-3.0
Idle	41	45	44	45	46
60 mph	65	67	66	66	76
0–60 mph	76	72	76	77	64

Source: *Popular Science* 254–258 (selected issues).

Note: Data are a random sample to be used for educational purposes only and should not be viewed as a guide to vehicle performance.

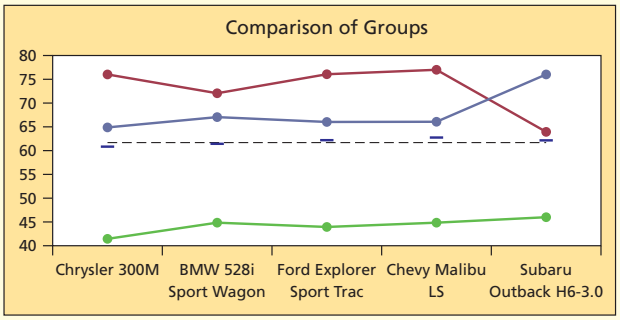
The general form of the model is $NoiseLevel = f(CarSpeed, CarType)$. Degrees of freedom for *CarSpeed* (rows) will be $r - 1 = 3 - 1 = 2$, while degrees of freedom for *CarType* (columns) will be $c - 1 = 5 - 1 = 4$. Denominator degrees of freedom will be the same for both factors since *SSE* has degrees of freedom $(r - 1)(c - 1) = (3 - 1)(5 - 1) = 8$. Excel's ANOVA results and MegaStat's dot plot are shown in Figure 11.21.

FIGURE 11.21 Results of two-factor ANOVA without replication for car noise

SUMMARY	Count	Sum	Average	Variance
Idle	5	221	44.200	3.700
60 mph	5	340	68.000	20.500
0-60 mph	5	365	73.000	29.000
Chrysler 300M	3	182	60.667	320.333
BMW 528i Sport Wagon	3	184	61.333	206.333
Ford Explorer Sport Trac	3	186	62.000	268.000
Chevy Malibu LS	3	188	62.667	264.333
Subaru Outback H6-3.0	3	186	62.000	228.000

ANOVA: Two-Factor Without Replication

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	2368.133	2	1184.067	46.01295	0.00004	4.45897
Columns	6.933333	4	1.733333	0.06736	0.99006	3.83785
Error	205.8667	8	25.73333			
Total	2580.933	14				



Since $F = 46.01$ exceeds $F_{2,8} = 4.46$, we see that *CarSpeed* (row factor) does have a highly significant effect on noise level. Its very small p-value ($p = .00004$) is unlikely to be a chance result. But *CarType* (column factor) has no significant effect on noise level since $F = 0.07$ does not exceed $F_{4,8} = 3.84$. The p-value for *CarType* ($p = .99006$) says that its F statistic could easily have arisen by chance. In short, we conclude that

- Interior noise *is* significantly affected by car speed ($p = .00004$).
- Interior noise *is not* significantly affected by car type ($p = .9901$).

We do not bother with Tukey multiple comparisons of means since we know that car type has no significant effect on noise level (the research hypothesis) and the effect of initial speed is of less research interest (a blocking factor).

11.6 TWO-FACTOR ANOVA WITH REPLICATION (FULL FACTORIAL MODEL)

What Does Replication Accomplish?

In a two-factor model, suppose that each factor combination is observed m times. With an equal number of observations in each cell (*balanced data*) we have a two-factor ANOVA model *with* ***replication.*** Replication allows us to test not only the factors' ***main effects*** but also an ***interaction effect.*** This model is often called the ***full factorial*** model. In linear model format it may be written

(11.15)
$$y_{ijk} = \mu + A_j + B_k + AB_{jk} + \varepsilon_{ijk}$$

where

y_{ijk} = observation i for row j and column $k (i = 1, 2, \ldots, m)$
μ = common mean for all treatments
A_j = effect attributed to factor A in row $j (j = 1, 2, \ldots, r)$
B_k = effect attributed to factor B in column $k (k = 1, 2, \ldots, c)$
AB_{jk} = effect attributed to interaction between factors A and B
ε_{ijk} = random error (normally distributed, zero mean, same variance for all treatments)

Interaction effects can be important. For example, an agribusiness researcher might postulate that corn yield is related to seed type (A), soil type (B), interaction between seed type and soil type (AB), or all three. In the absence of any factor effects, all variation about the mean μ is purely random.

Format of Hypotheses

For a fixed-effects ANOVA model, the hypotheses that could be tested in the two-factor ANOVA model with replicated observations are

Factor *A*: Row Effect

H_0: $A_1 = A_2 = \cdots = A_r = 0$ (factor A has no effect)
H_1: Not all the A_j are equal to zero (factor A has an effect)

Factor *B*: Column Effect

H_0: $B_1 = B_2 = \cdots = B_c = 0$ (factor B has no effect)
H_1: Not all the B_k are equal to zero (factor B has an effect)

Interaction Effect

H_0: All the AB_{jk} are equal to zero (there is no interaction effect)
H_1: Not all AB_{jk} are equal to zero (there is an interaction effect)

If none of the proposed factors has anything to do with Y, then the model collapses to

$$y_{ijk} = \mu + \varepsilon_{ijk} \tag{11.16}$$

Format of Data

Table 11.12 shows the format of a data set with two factors and a balanced (equal) number of observations per treatment (each row/column intersection is a treatment). To avoid needless subscripts, the m observations in each treatment are represented simply as yyy. Except for the replication within cells, the format is the same as the unreplicated two-factor ANOVA.

TABLE 11.12
Data Format of Replicated Two-Factor ANOVA

	Levels of Factor *B*				
Levels of Factor *A*	B_1	B_2	. . .	B_c	***Row Mean***
A_1	*yyy* *yyy* . . . *yyy*	*yyy* *yyy* . . . *yyy*	. . . .	*yyy* *yyy* . . . *yyy*	$\bar{y}_{1.}$
A_2	*yyy* *yyy* . . . *yyy*	*yyy* *yyy* . . . *yyy*	. . . .	*yyy* *yyy* . . . *yyy*	$\bar{y}_{2.}$
. . .	. . .	. . .	. . .	. . .	. . .
A_r	*yyy* *yyy* . . . *yyy*	*yyy* *yyy* . . . *yyy*	. . . .	*yyy* *yyy* . . . *yyy*	$\bar{y}_{r.}$
Col Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	. . .	$\bar{y}_{.c}$	$\bar{y}$

Sources of Variation

There are now three F tests that could be performed: one for each main effect (factors A and B) and a third F test for interaction. The total sum of squares is partitioned into four components:

$$SST = SSA + SSB + SSI + SSE \tag{11.17}$$

where

SST = total sum of squared deviations about the mean
SSA = between rows sum of squares (effect of factor A)
SSB = between columns sum of squares (effect of factor B)
SSI = interaction sum of squares (effect of AB)
SSE = error sum of squares (residual variation)

For an experiment with r rows, c columns, and m replications per treatment, the sums of squares and ANOVA calculations may be presented in a table, shown in Table 11.13.

If SSE is relatively high, we expect that we would fail to reject H_0 for the various hypotheses. Conversely, if SSE is relatively small, it is likely that at least one of the factors (row effect, column effect, or interaction) is a relevant predictor of Y. Before doing the F test, each sum of squares must be divided by its degrees of freedom to obtain its *mean square*. Degrees of freedom sum to $n - 1$ (note that $n = rcm$).

TABLE 11.13 Two-Factor ANOVA with Replication

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Factor A (row effect)	$SSA = cm\sum_{j=1}^{r}(\bar{y}_{j.} - \bar{y})^2$	$r - 1$	$MSA = \frac{SSA}{r-1}$	$F_A = \frac{MSA}{MSE}$
Factor B (column effect)	$SSB = rm\sum_{k=1}^{c}(\bar{y}_{.k} - \bar{y})^2$	$c - 1$	$MSB = \frac{SSB}{c-1}$	$F_B = \frac{MSB}{MSE}$
Interaction ($A \times B$)	$SSI = m\sum_{j=1}^{r}\sum_{k=1}^{c}(\bar{y}_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y})^2$	$(r - 1)(c - 1)$	$MSI = \frac{SSI}{(r-1)(c-1)}$	$F_I = \frac{MSI}{MSE}$
Error	$SSE = \sum_{i=1}^{m}\sum_{j=1}^{r}\sum_{k=1}^{c}(y_{ijk} - \bar{y}_{jk})^2$	$rc(m - 1)$	$MSE = \frac{SSE}{rc(m-1)}$	
Total	$SST = \sum_{i=1}^{m}\sum_{j=1}^{r}\sum_{k=1}^{c}(y_{ijk} - \bar{y})^2$	$rcm - 1$		

EXAMPLE

Delivery Time

A health maintenance organization orders weekly medical supplies for its four clinics from five different suppliers. Delivery times (in days) for 4 recent weeks are shown in Table 11.14.

TABLE 11.14 Delivery Times (in days) Deliveries

	Supplier 1	Supplier 2	Supplier 3	Supplier 4	Supplier 5
Clinic A	8	14	10	8	17
	8	9	15	7	12
	10	14	10	13	9
	13	11	7	10	10
Clinic B	13	9	12	6	15
	14	9	10	10	12
	12	7	10	12	12
	13	8	11	8	10
Clinic C	11	8	12	10	14
	10	9	10	11	13
	12	11	13	7	10
	14	12	10	10	12
Clinic D	7	8	7	8	14
	10	13	5	5	13
	10	9	6	11	8
	13	12	5	4	11

Using short variable names, the two-factor ANOVA model has the general form

$$DeliveryTime = f(Clinic, Supplier, Clinic \times Supplier)$$

The effects are assumed additive. The linear model is

$$y_{ijk} = \mu + A_j + B_k + AB_{jk} + \varepsilon_{ijk}$$

Step 1: State the Hypotheses

The hypotheses are

Factor *A*: Row Effect (*Clinic*)

H_0: $A_1 = A_2 = \cdots = A_r = 0$ (clinic means are the same)

H_1: Not all the A_j are equal to zero (clinic means differ)

Factor *B*: Column Effect (*Supplier*)

H_0: $B_1 = B_2 = \cdots = B_c = 0$ (supplier means are the same)

H_1: Not all the B_k are equal to zero (supplier means differ)

Interaction Effect (*Clinic*×*Supplier*)

H_0: All the AB_{jk} are equal to zero (there is no interaction effect)

H_1: Not all AB_{jk} are equal to zero (there is an interaction effect)

Step 2: State the Decision Rule

Each F test may require a different right-tail critical value because the numerator degrees of freedom depend on the number of factor levels, while denominator degrees of freedom (error *SSE*) are the same for all three tests:

Factor A: $\text{d.f.}_1 = r - 1 = 4 - 1 = 3$ ($r = 4$ clinics)

Factor B: $\text{d.f.}_1 = c - 1 = 5 - 1 = 4$ ($c = 5$ suppliers)

Interaction (AB): $\text{d.f.}_1 = (r - 1)(c - 1) = (4 - 1)(5 - 1) = 12$

Error $\text{d.f.}_2 = rc(m - 1) = 4 \times 5 \times (4 - 1) = 60$

Excel provides the right-tail F critical values for $\alpha = .05$, which we can verify using Appendix F:

$F_{3,60} = 2.76$ for Factor A

$F_{4,60} = 2.53$ for Factor B

$F_{12,60} = 1.92$ for Factor AB

We reject the null hypothesis if an F test statistic exceeds its critical value.

Step 3: Perform the Calculations

Excel provides tables of row and column sums and means (not shown here because they are lengthy). The ANOVA table in Figure 11.22 summarizes the partitioning of variation into its component sums of squares, degrees of freedom, mean squares, F test statistics, p-values, and critical F-values for $\alpha = .05$.

FIGURE 11.22

Excel's two-factor ANOVA with replication **Deliveries**

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	51.350	3	17.1167	3.434783	0.022424	2.758078
Columns	104.425	4	26.1063	5.238712	0.001097	2.525215
Interaction	102.775	12	8.5646	1.718645	0.085176	1.917396
Within	299.000	60	4.9833			
Total	557.550	79				

Step 4: Make the Decision

For the row variable (*Clinic*) the test statistic $F = 3.435$ and its p-value ($p = .0224$) lead us to conclude that the mean delivery times among clinics are not the same at $\alpha = .05$. For the

column variable (*Supplier*) the test statistic $F = 5.239$ and its p-value ($p = .0011$) lead us to conclude that the mean delivery times from suppliers are not the same at $\alpha = .05$. For the interaction effect, the test statistic $F = 1.719$ and its p-value ($p = .0852$) lack significance at $\alpha = .05$. The p-values permit a more flexible interpretation since α need not be specified in advance. In summary:

Variable	*p-Value*	*Interpretation*
Clinic	.0224	Clinic means differ (significant at $\alpha = .05$)
Supplier	.0011	Supplier means differ (significant at $\alpha = .01$)
Clinic×Supplier	.0852	Weak interaction effect (significant at $\alpha = .10$)

Using MegaStat

MegaStat's two-factor ANOVA results, shown in Figure 11.23, are similar to Excel's except that the table of treatment means is more compact, the results are rounded, and significant p-values are highlighted (bright yellow for $\alpha = .01$, light green for $\alpha = .05$).

FIGURE 11.23

MegaStat's two-factor ANOVA **Deliveries**

Two factor ANOVA

Means:

Factor 1	Factor 2: Supplier 1	Supplier 2	Supplier 3	Supplier 4	Supplier 5	
Clinic A	9.8	12.0	10.5	9.5	12.0	10.8
Clinic B	13.0	8.3	10.8	9.0	12.3	10.7
Clinic C	11.8	10.0	11.3	9.5	12.3	11.0
Clinic D	10.0	10.5	5.8	7.0	11.5	9.0
	11.1	10.2	9.6	8.8	12.0	10.3

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Factor 1	51.35	3	17.117	3.43	.0224
Factor 2	104.43	4	26.106	5.24	.0011
Interaction	102.78	12	8.565	1.72	.0852
Error	299.00	60	4.983		
Total	557.56	79			

Interaction Effect

The statistical test for interaction is just like any other F test. But you might still wonder, What *is* an interaction, anyway? You may be familiar with the idea of drug interaction. If you consume a few ounces of vodka, it has an effect on you. If you take an allergy pill, it has an effect on you. But if you combine the two, the effect may be different (and possibly dramatic) compared with either drug by itself. That is why many medications carry a warning like "Avoid alcohol while using this medication."

To visualize an interaction, we plot the treatment means for one factor against the levels of the other factor. Within each factor level, we connect the means. In the absence of an interaction, the lines will be roughly parallel or will tend to move in the same direction at the same time. If there is a strong interaction, the lines will have differing slopes and will tend to cross one another.

Figure 11.24 illustrates several possible situations, using a hypothetical two-factor ANOVA model in which factor A has three levels and factor B has two levels. For the delivery time example, a significant *interaction effect* would mean that suppliers have different mean delivery times for different clinics. However, Figure 11.25 shows that, while the interaction plot lines do cross, there is no consistent pattern, and the lines tend to be parallel more than crossing. The visual indications of interaction are, therefore, weak for the delivery time data. This conclusion is consistent with the interaction p-value ($p = .085$) for the F test of $A \times B$.

© The McGraw–Hill Companies, 2007

FIGURE 11.24

Possible interaction patterns

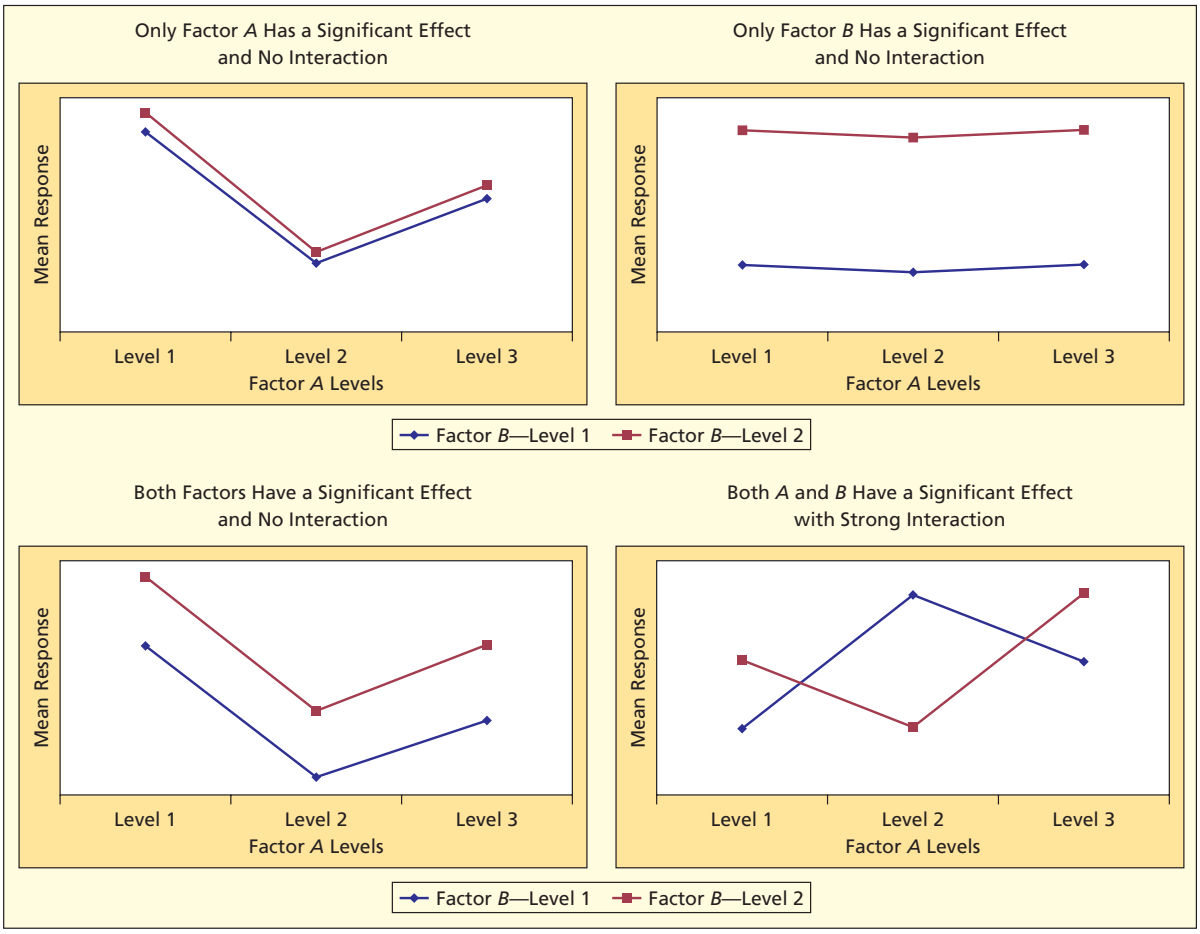
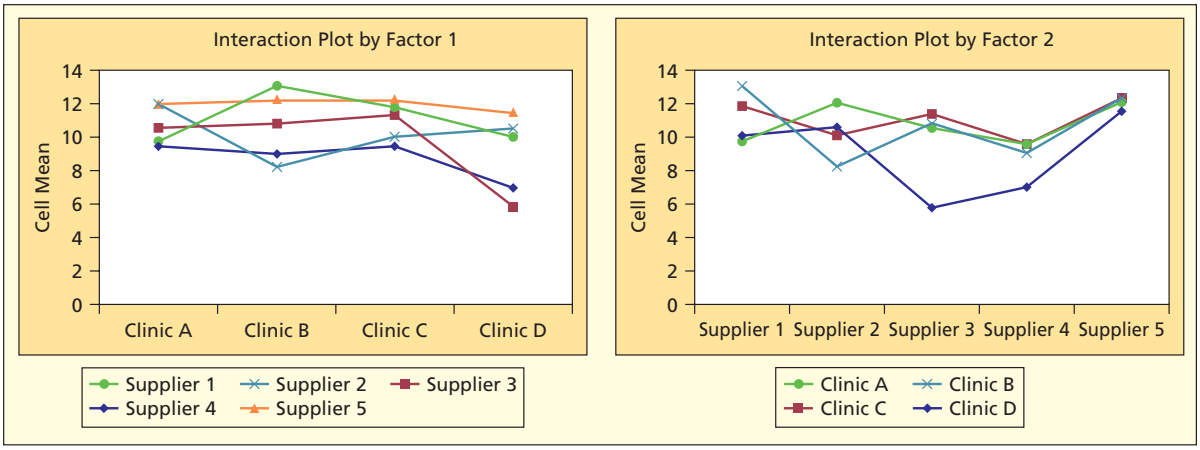


FIGURE 11.25

Interaction plots from MegaStat **Deliveries**



Tukey Tests of Pairs of Means

MegaStat's Tukey comparisons, shown in Figure 11.26, reveal significant differences at $\alpha = .05$ between plants *C*, *D* and between suppliers (1, 4) and (3, 5). At $\alpha = .01$ there is also a significant difference in means between one pair of suppliers (4, 5).

FIGURE 11.26

MegaStat table of Tukey comparisons **Deliveries**

Tukey simultaneous comparison *t*-values (d.f. = 60)

		Clinic D	Clinic B	Clinic A	Clinic C
		9.0	10.7	10.8	11.0
Clinic D	9.0				
Clinic B	10.7	2.41			
Clinic A	10.8	2.55	0.14		
Clinic C	11.0	2.83	0.42	0.28	

critical values for experimentwise error rate:

0.05	2.64
0.01	3.25

Tukey simultaneous comparison *t*-values (d.f. = 60)

		Supplier 4	Supplier 3	Supplier 2	Supplier 1	Supplier 5
		8.8	9.6	10.2	11.1	12.0
Supplier 4	8.8					
Supplier 3	9.6	1.03				
Supplier 2	10.2	1.82	0.79			
Supplier 1	11.1	3.01	1.98	1.19		
Supplier 5	12.0	4.12	3.09	2.30	1.11	

critical values for experimentwise error rate:

0.05	2.81
0.01	3.41

Significance versus Importance

MegaStat's table of means (Figure 11.23) allows us to explore these differences further and to assess the question of *importance* as well as *significance*. The largest differences in means between clinics or suppliers are about 2 days. Such a small difference might be unimportant most of the time. However, if their inventory is low, a 2-day difference could be important.

SECTION EXERCISES

Instructions: For each data set: (a) State the hypotheses. (b) Use Excel's Tools > Data Analysis (or MegaStat or MINITAB) to perform the two-factor ANOVA with replication, using $\alpha = .05$. (c) State your conclusions about the main effects and interaction effects. (d) Interpret the *p*-values carefully. (e) Create interaction plots and interpret them.

11.17 A small independent stock broker has created four sector portfolios for her clients. Each portfolio always has five stocks that may change from year to year. The volatility (coefficient of variation) of each stock is recorded for each year. Are the main effects significant? Is there an interaction? **Volatility**

	Stock Portfolio Type			
Year	*Health*	*Energy*	*Retail*	*Leisure*
2004	14.5	23.0	19.4	17.6
	18.4	19.9	20.7	18.1
	13.7	24.5	18.5	16.1
	15.9	24.2	15.5	23.2
	16.2	19.4	17.7	17.6
2005	21.6	22.1	21.4	25.5
	25.6	31.6	26.5	24.1
	21.4	22.4	21.5	25.9
	26.6	31.3	22.8	25.5
	19.0	32.5	27.4	26.3
2006	12.6	12.8	22.0	12.9
	13.5	14.4	17.1	11.1
	13.5	13.1	24.8	4.9
	13.0	8.1	13.4	13.3
	13.6	14.7	22.2	12.7

11.18 Oxnard Petro, Ltd., has three interdisciplinary project development teams that function on an ongoing basis. Team members rotate from time to time. Every 4 months (three times a year) each department head rates the performance of each project team (using a 0 to 100 scale, where 100 is the best rating). Are the main effects significant? Is there an interaction? **Ratings**

Year	Marketing	Engineering	Finance
2004	90	69	96
	84	72	86
	80	78	86
2005	72	73	89
	83	77	87
	82	81	93
2006	92	84	91
	87	75	85
	87	80	78

11.19 A market research firm is testing consumer reaction to a new shampoo on four age groups in four regions. There are five consumers in each test panel. Each consumer completes a 10-question product satisfaction instrument with a 5-point scale (5 is the highest rating) and the average score is recorded. Are the main effects significant? Is there an interaction? **Satisfaction**

	Northeast	Southeast	Midwest	West
Youth (under 18)	3.9	3.9	3.6	3.9
	4.0	4.2	3.9	4.4
	3.7	4.4	3.9	4.0
	4.1	4.1	3.7	4.1
	4.3	4.0	3.3	3.9
College (18–25)	4.0	3.8	3.6	3.8
	4.0	3.7	4.1	3.8
	3.7	3.7	3.8	3.6
	3.8	3.6	3.9	3.6
	3.8	3.7	4.0	4.1
Adult (26–64)	3.2	3.5	3.5	3.8
	3.8	3.3	3.8	3.6
	3.7	3.4	3.8	3.4
	3.4	3.5	4.0	3.7
	3.4	3.4	3.7	3.1
Senior (65+)	3.4	3.6	3.3	3.4
	2.9	3.4	3.3	3.2
	3.6	3.6	3.1	3.5
	3.7	3.6	3.1	3.3
	3.5	3.4	3.1	3.4

11.20 Oxnard Petro, Ltd., has three suppliers of catalysts. Orders are placed with each supplier every 15 working days, or about once every 3 weeks. The delivery time (days) is recorded for each order over 1 year. Are the main effects significant? Is there an interaction? **Deliveries2**

	Supplier 1	Supplier 2	Supplier 3
Qtr 1	12 15 11 11	10 13 11 9	16 13 14 14
Qtr 2	13 11 13 12	10 10 13 11	14 11 12 12
Qtr 3	12 8 8 13	11 9 8 6	13 8 13 6
Qtr 4	8 10 13 11	8 10 10 10	11 11 10 11

Mini Case 11.3

Turbine Engine Thrust

Engineers testing turbofan aircraft engines wanted to know if oil pressure and turbine temperature are related to engine thrust (pounds). They chose four levels for each factor and observed each combination five times, using the two-factor replicated ANOVA model $Thrust = f(OilPres, TurbTemp, OilPres \times TurbTemp)$. The test data are shown in Table 11.15.

TABLE 11.15 Turbofan Engine Thrust Test Results Turbines

	Turbine Temperature			
Oil Pressure	T1	T2	T3	T4
P1	1,945.0 1,933.0 1,942.4 1,948.0 1,930.0	1,942.3 1,931.7 1,946.0 1,959.0 1,939.9	1,934.2 1,930.0 1,944.0 1,941.0 1,942.0	1,916.7 1,943.0 1,948.8 1,928.0 1,946.0
P2	1,939.4 1,952.8 1,940.0 1,948.0 1,925.0	1,922.0 1,936.8 1,928.0 1,930.7 1,939.0	1,950.6 1,947.9 1,950.0 1,922.0 1,918.0	1,929.6 1,930.0 1,934.0 1,923.0 1,914.0
P3	1,932.0 1,955.0 1,949.7 1,933.0 1,936.5	1,939.0 1,932.0 1,933.1 1,952.0 1,943.0	1,952.0 1,963.0 1,923.0 1,965.0 1,944.0	1,960.4 1,946.0 1,931.0 1,949.0 1,906.0
P4	1,960.2 1,909.3 1,950.0 1,920.0 1,964.9	1,937.0 1,941.0 1,928.2 1,938.9 1,919.0	1,940.0 1,984.0 1,971.0 1,930.0 1,944.0	1,924.0 1,906.0 1,925.8 1,923.0 1,916.7

Source: Research project by three engineering students enrolled in an MBA program. Data are disguised.

The ANOVA results in Figure 11.27 indicate that only turbine temperature is significantly related to thrust. The table of means suggests that, because mean thrust varies only over a tiny range, the effect may not be very important. The lack of interaction is revealed by the nearly parallel ***interaction plots.*** Levene's test for equal variances (not shown) shows a p-value of $p = .42$ indicating that variances may be assumed equal, as is desirable for an ANOVA test.

FIGURE 11.27

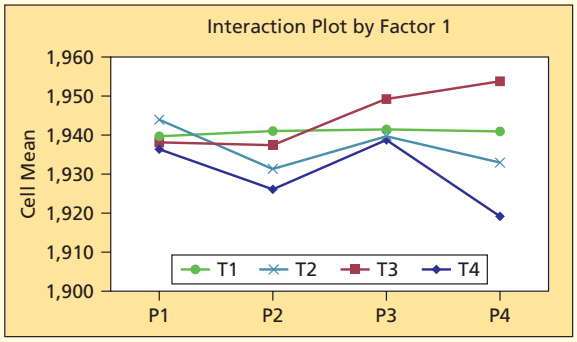
MegaStat two-factor ANOVA results

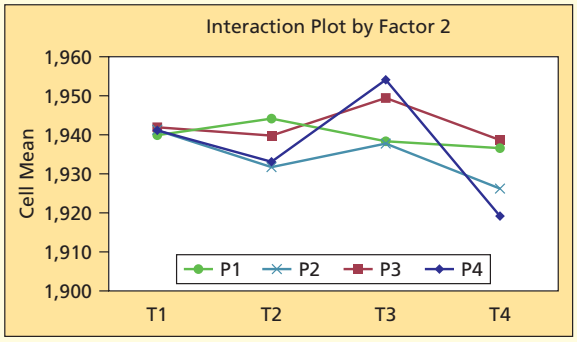
Means: Temperature

Pressure	T1	T2	T3	T4	
P1	1,939.68	1,943.78	1,938.24	1,936.50	1,939.55
P2	1,941.04	1,931.30	1,937.70	1,926.12	1,934.04
P3	1,941.24	1,939.82	1,949.40	1,938.48	1,942.24
P4	1,940.88	1,932.82	1,953.80	1,919.10	1,936.65
	1,940.71	1,936.93	1,944.79	1,930.05	1,938.12

ANOVA table: Two-Factor with Replication

Source	*SS*	*df*	*MS*	*F*	*p-value*
Pressure	755.708	3	251.9028	1.32	.2756
Temperature	2,353.426	3	784.4755	4.11	.0099
Interaction	1,989.095	9	221.0106	1.16	.3367
Error	12,212.052	64	190.8133		
Total	17,310.282	79			





11.7 GENERAL LINEAR MODEL (OPTIONAL)

Higher-Order ANOVA Models

Why limit ourselves to two factors? Although a three-factor data set cannot be shown in a two-dimensional table, the idea of a three-factor ANOVA is not difficult to grasp. Consider the hospital LOS and paint viscosity problems introduced at the beginning of this chapter. Figure 11.28 adds a third factor (gender) to the hospital model and Figure 11.29 adds a third factor (solvent ratio) to the paint viscosity model.

FIGURE 11.28

Three-factor ANOVA model for hospital length of stay

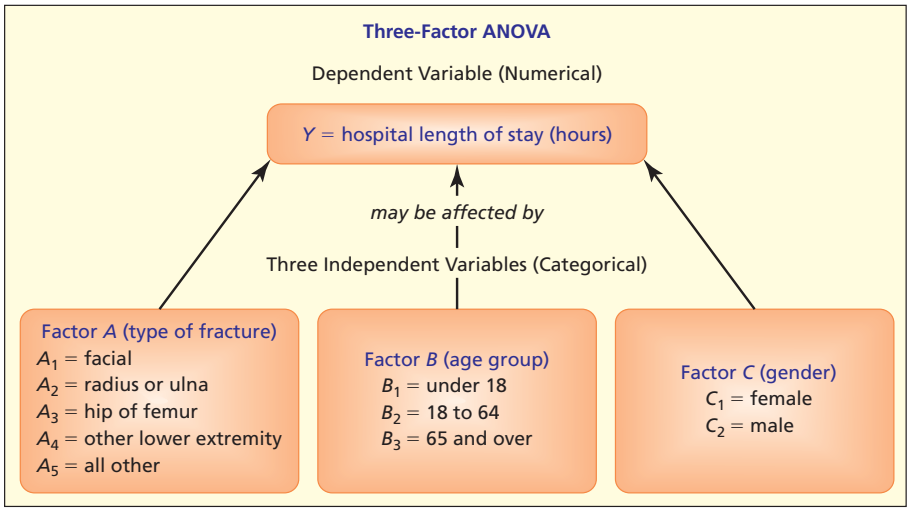
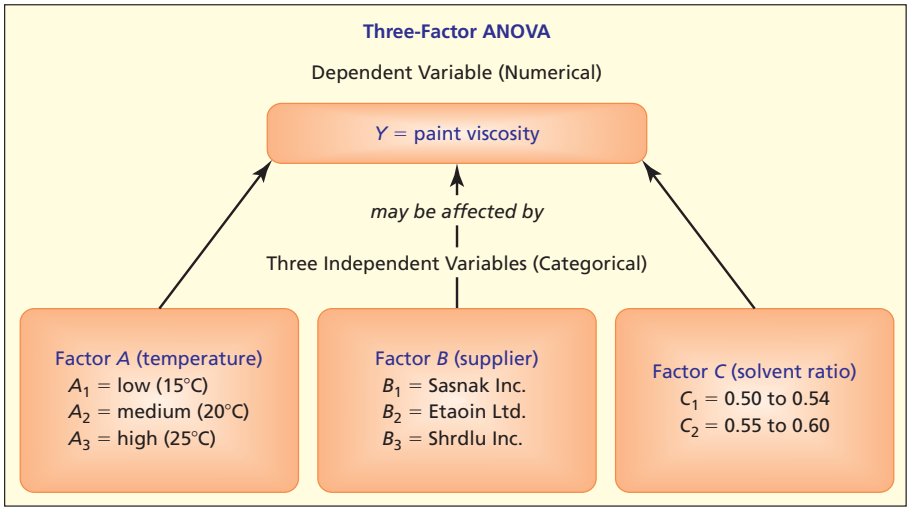


FIGURE 11.29

Three-factor ANOVA model for paint viscosity



A three-factor ANOVA allows more two-factor interactions ($A \times B$, $A \times C$, $B \times C$) and even a three-factor interaction ($A \times B \times C$). However, since the computations are already done by computer, the analysis would be no harder than a two-factor ANOVA. The "catch" is that higher-order ANOVA models are beyond Excel's capabilities, so you will need fancier software. Fortunately, any general-purpose statistical package (e.g., MINITAB, SPSS, SAS) can handle ANOVA with *any* number of factors with *any* number of levels (subject to computer software limitations).

What Is GLM?

The ***general linear model*** (GLM) is a versatile tool for estimating large and complex ANOVA models. Besides allowing more than two factors, GLM permits unbalanced data (unequal sample size within treatments) and any desired subset of interactions among factors (including three-way interactions or higher) as long as you have enough observations (i.e., enough degrees of freedom) to compute the effects. GLM can also provide predictions and identify unusual observations. GLM does not require equal variances, although care must be taken to avoid sparse or empty cells in the data matrix. Data are expected to be in stacked format (one column for Y and one column for each factor A, B, C, etc.). The output of GLM is easily understood by anyone who is familiar with ANOVA, as you can see in Mini Case 11.4.

Mini Case 11.4

Hospital Maternity Stay MaternityLOS

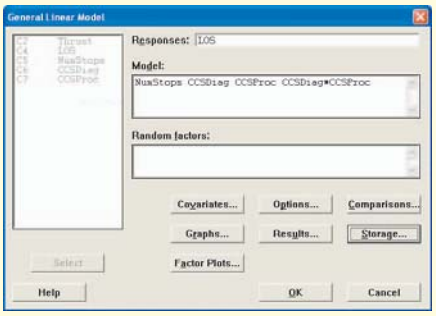
The data set consists of 4,409 maternity hospital visits whose DRG (diagnostic-related group) code is 373 (simple delivery without complicating diagnoses). The dependent variable of interest is the length of stay (LOS) in the hospital. The model contains one discrete numerical factor and two categorical factors: the number of surgical stops (*NumStops*), the CCS diagnostic code (*CCSDiag*), and the CCS procedure code (*CCSProc*). CCS codes are a medical classification scheme developed by the American Hospital Research Council to help hospitals and researchers organize medical information. The proposed model is

$$LOS = f(NumStops, CCSDiag, CCSProc)$$

Before starting the GLM analysis, a frequency tabulation was prepared for each factor. The tabulation (not shown) revealed that some factor levels were observed too rarely to be useful. Cross-tabulations (not shown) also revealed that some treatments would be empty or very sparse. Based on this preliminary data screening, the factors were recoded to avoid GLM estimation problems. *NumStops* was recoded as a binary variable (2 if there were 1 or 2 stops, 3 if there were 3 or more stops). *CCSDiag* codes with a frequency less than 100 were recoded as 999. Patients whose *CCSProc* code occurred less than 10 times (19 patients) were deleted from the sample, leaving a sample of 4,390 patients.

MINITAB's menu and GLM results are shown in Figure 11.30. You can select a variable by clicking on it, but if you want an interaction, you must type it in the Model window. The first thing shown is the number of levels for each factor and the discrete values of each factor. Frequencies of the factor values are not shown, but can be obtained from MINITAB's Tables command.

FIGURE 11.30
MINITAB menu and GLM results MaternityLOS



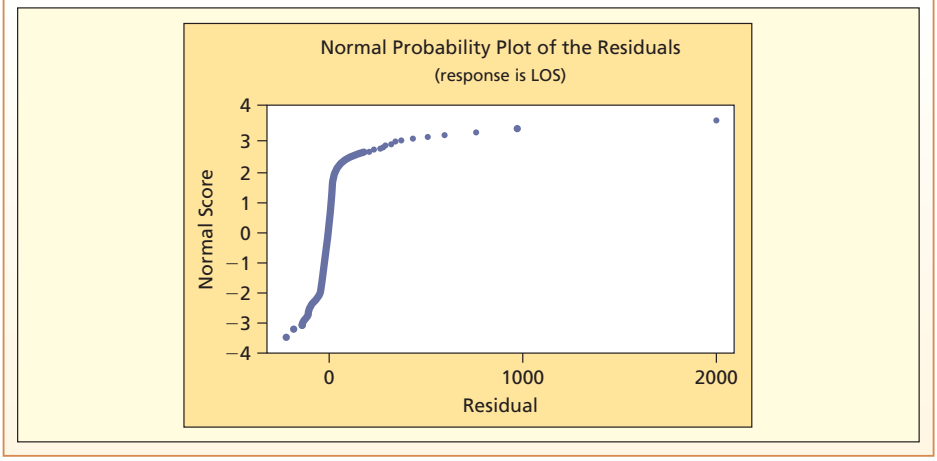

```
Factor     Type Levels Values
NumStops   fixed      2 2 3
CCSDiag    fixed      9 181 184 185 190 191 193 195 196 999
CCSProc    fixed      6 133 135 136 137 139 140

Analysis of Variance for LOS, using Adjusted SS for Tests

Source            DF     Seq SS     Adj SS     Adj MS       F      P
NumStops           1       1340       7385       7385    3.25  0.072
CCSDiag            8     446299      94498      11812    5.19  0.000
CCSProc            5      37139      45061       9012    3.96  0.001
CCSDiag*CCSProc   40     194618     194618       4865    2.14  0.000
Error           4335    9865440    9865440       2276
Total           4389   10544837
```

The *p*-values from the ANOVA table suggest that *NumStops* is significant at $\alpha = .10$ ($p = .072$) while the other two main effects *CCSDiag* and *CCSProc* ($p = .000$ and $p = .001$) are highly significant (such small *p*-values would arise less than once in 1,000 samples if there were no relationship). The interaction *CCSDiag*×*CCSProc* is also highly significant ($p = .000$). Because the sample size is large, even slight effects could be *significant,* so further analysis may be needed to see if the effects are also *important*. Unfortunately, there are outliers in the data (a violation of the normality assumption) that show up clearly on the residual normality plot in Figure 11.31. Thus, we must regard the ANOVA results with caution.

FIGURE 11.31
MINITAB normality plot



11.8 EXPERIMENTAL DESIGN: AN OVERVIEW (OPTIONAL)

Experimental design is a specialized topic that goes far beyond this textbook. However, you may need to interact professionally with engineers or quality improvement teams that are working on product design, reliability, and product performance. It is therefore helpful to have a general idea of what experimental design is all about and to learn some of the basic terminology. If you become more deeply involved, you can ask your employer to send you to a 3-day training class in experimental design to boost your skills.

What Is Experimental Design?

Experimental design refers to the number of factors under investigation, the number of levels assigned to each factor, the way factor levels are defined, and the way observations are obtained. *Fully crossed* or *full factorial* designs include all possible combinations of factor levels. ***Fractional factorial*** designs, for reasons of economy, limit data collection to a subset of possible factor combinations. If all levels of one factor are fully contained in another, the design is ***nested*** or ***hierarchical. Balanced designs*** are characterized by an equal number of observations for each factor combination. In a *fixed-effects model* the levels of each factor are predetermined, which implies that our inferences are valid only for the specified factor levels. For example, if a firm has only three paint suppliers (S_1, S_2, S_3), these would be our factor levels. In a *random effects model* the factor levels would be chosen randomly from a population of potential factor levels. For example, if a firm has 20 paint suppliers (S_1 through S_{20}) but we only want to study three of them, we might choose three at random (say S_7, S_{11}, and S_{18}) from the 6,840 possible ways to choose 3 items from 20. Fixed effects are by far the most common models used in business analysis, where randomization and controlled experiments are not practical.

2^k Models

When there are k factors, each with two levels, we have a 2^k *factorial design*. Reducing a factor to two levels is a useful simplification that reduces the data requirements in a replicated experiment, because the data matrix will have fewer cells. Even a continuous factor (e.g., *Pressure*) can be "binarized" into roughly equal groups (*Low, High*) by cutting the data array at the median. The 2^k design is especially useful when the number of factors is very large. In automotive engineering, for example, it is not uncommon to study more than a dozen factors that are predictive of exhaust emissions. Even when each factor is limited to only two levels, full factorial 2^k experiments with replication can require substantial data-collection effort.

Fractional Factorial Designs

Unlike a full factorial design, a *fractional factorial* design, for reasons of economy, limits data collection to a subset of the possible factor combinations. Fractional factorial designs are

extremely important in real-life situations where many factors exist. For example, suppose that automobile combustion engineers are investigating 10 factors, each with two levels, to determine their effect on emissions. This would yield $2^{10} = 1{,}024$ possible factor combinations. It would be impractical and uneconomical to gather data for all 1,024 factor combinations.

By excluding some factor combinations, a fractional factorial model necessarily sacrifices some of the interaction effects. If the most important objective is to study the *main effects* (which is frequently the case, or is at least an acceptable compromise), it is possible to get by with a much smaller number of observations. It often is possible to estimate some, though not all, interaction effects in a fractional factorial experiment. Templates are published to guide experimenters in choosing the correct design and sample size for the desired number of factors (see Related Reading). The American Supplier Institute sponsors training seminars on *robust engineering* utilizing *Taguchi designs* to make efficient use of data.

Nested or Hierarchical Design

If all levels of one factor are fully contained within another, the design is *nested* or *hierarchical.* Using most computer packages, nested designs can be represented using simple notation like

$$Defects = f(Experience, Method\ (Machine))$$

In this model, *Machine* is nested within *Method* so the effect of *Machine* cannot appear as a main effect. Presumably the nature of the manufacturing process dictates that *Machine* depends on *Method*. Although the model is easy to state, this example is not intended to suggest that estimates of nested models are easy to interpret.

Random Effects Models

In a *fixed-effects model* the levels of each factor are predetermined, which implies that our inferences are valid only for the specified factor levels. In a *random effects model* the factor levels are chosen randomly from a population of potential factor levels. Computation and interpretation of random effects are more complicated, and not all tests may be feasible. Novices are advised that estimation of random effects models should be preceded by further study (see Related Reading).

Chapter Summary

ANOVA tests whether a numerical dependent variable (**response variable**) is associated with one or more categorical independent variables (**factors**) with several **levels.** Each level or combination of levels is a treatment. A **one-factor ANOVA** compares means in c columns of data. It is a generalization of a two-tailed t test for two independent sample means. Fisher's $\boldsymbol{F}$ **statistic** is a ratio of two variances (treatment versus error). It is compared with a right-tailed critical value from an F table or from Excel for appropriate numerator and denominator degrees of freedom. Alternatively, we compare the p-value for the F test statistic with the desired level of significance (p less than α is significant). An **unreplicated two-factor ANOVA** can be viewed as a **randomized block model** if only one factor is of research interest. A **replicated two-factor ANOVA** (or full factorial model) has more than one observation per treatment, permitting inclusion of an interaction test in addition to tests for the **main effects. Interaction effects** can be seen as crossing lines on plots of factor means. The **Tukey test** compares individual treatment means. We test for homogeneous variances (an assumption of ANOVA) using **Hartley's $\boldsymbol{F_{max}}$ test** or **Levene's test.** The **general linear model** (GLM) can be used when there are more than two factors. **Experimental design** helps make efficient use of limited data. Other general advice:

- ANOVA may be helpful even if those who collected the data did not utilize a formal experimental design (often the case in real-world business situations).
- ANOVA calculations are tedious because of the sums required, so computers are generally used.
- One-factor ANOVA is the most common and suffices for many business situations.
- ANOVA is an overall test. To tell which specific pairs of treatment means differ, use the Tukey test.
- Although real-life data may not perfectly meet the normality and equal-variance assumptions, ANOVA is reasonably robust (and alternative tests do exist).
- Experimental design books usually can show you an example of exactly the design you need for parsimonious use of data. Call for expert advice when you are in doubt, to save a lot of rework.

Key Terms

analysis of variance (ANOVA), *439*
balanced designs, *476*
experimental design, *476*
explained variance, *439*
factors, *439*
fixed-effects model, *457*
fractional factorial, *476*
full factorial, *464*
general linear model, *474*
Hartley's F_{max} test, *452*
hierarchical design, *476*
homogeneous variances, *452*
interaction, *441*
interaction effect, *464*
interaction plots, *473*
Levene's test, *454*
main effects, *464*
mean squares, *444*
multiple comparison, *450*
nested design, *476*
one-factor ANOVA, *442*
partitioned sum of squares, *444*
randomized block model, *457*
replication, *464*
response variable, *439*
treatment, *439*
Tukey's studentized range test, *450*
two-factor ANOVA without replication, *456*
unexplained variance, *439*

Chapter Review

Note: Questions labeled * are based on optional material from this chapter.

1. Explain each term: (a) explained variation; (b) unexplained variation; (c) factor; (d) treatment.
2. (a) Explain the difference between one-factor and two-factor ANOVA. (b) Write the linear model form of one-factor ANOVA. (c) State the hypotheses for a one-factor ANOVA in two different ways. (d) Why is one-factor ANOVA used a lot?
3. (a) State three assumptions of ANOVA. (b) What do we mean when we say that ANOVA is fairly robust to violations of these assumptions?
4. (a) Sketch the format of a one-factor ANOVA data set (completely randomized model). (b) Must group sizes be the same for one-factor ANOVA? Is it better if they are? (c) Explain the concepts of variation *between treatments* and variation *within treatments*. (d) What is the F statistic? (e) State the degrees of freedom for the F test in one-factor ANOVA.
5. (a) Sketch the format of a two-factor ANOVA data set without replication. (b) State the hypotheses for a two-factor ANOVA without replication. (c) What is the difference between a randomized block model and a two-factor ANOVA without replication? (d) What do the two F statistics represent in a two-factor ANOVA without replication? (e) What are their degrees of freedom?
6. (a) Sketch the format of a two-factor ANOVA data set with replication. (b) What is gained by replication? (c) State the hypotheses for a two-factor ANOVA with replication. (d) What do the three F statistics represent in a two-factor ANOVA with replication? (e) What are their degrees of freedom?
7. (a) What is the purpose of the Tukey test? (b) Why can't we just compare all possible pairs of group means using the two-sample t test?

*8. (a) What does a test for homogeneity of variances tell us? (b) Why should we test for homogeneity of variances? (c) Explain what Hartley's F_{max} test measures. (d) Why might we use Levene's test instead of the F_{max} test?

*9. What is the general linear model and why is it useful?

*10. (a) What is a 2^k design, and what are its advantages? (b) What is a fractional factorial design, and what are its advantages? (c) What is a nested or hierarchical design? (d) How is a random effects model different than a fixed-effects model?

CHAPTER EXERCISES

Instructions: You may use Excel, MegaStat, MINITAB, or another computer package of your choice. Attach appropriate copies of the output or capture the screens, tables, and relevant graphs and include them in a written report. Try to state your conclusions succinctly in language that would be clear to a decision

maker who is a nonstatistician. Exercises marked * are based on optional material. Answer the following questions, or those your instructor assigns.

a. Choose an appropriate ANOVA model. State the hypotheses to be tested.
b. Display the data visually (e.g., dot plots or MegaStat's line plots). What do the displays show?
c. Do the ANOVA calculations using the computer.
d. State the decision rule for $\alpha = .05$ and make the decision. Interpret the p-value.
e. In your judgment, are the observed differences in treatment means (if any) large enough to be of practical importance?
f. Do you think the sample size is sufficient? Explain. Could it be increased? Given the nature of the data, would more data collection be costly?
g. Perform Tukey multiple comparison tests and discuss the results.
*h. Perform a test for homogeneity of variances. Explain fully.

11.21 Below are grade point averages for 25 randomly chosen university business students during a recent semester. *Research question:* Are the mean grade point averages the same for students in these four class levels? **GPA2**

Grade Point Averages of 25 Business Students

Freshman (5 students)	*Sophomore (7 students)*	*Junior (7 students)*	*Senior (6 students)*
1.91	3.89	3.01	3.32
2.14	2.02	2.89	2.45
3.47	2.96	3.45	3.81
2.19	3.32	3.67	3.02
2.71	2.29	3.33	3.01
	2.82	2.98	3.17
	3.11	3.26	

11.22 The XYZ Corporation is interested in possible differences in days worked by salaried employees in three departments in the financial area. A survey of 23 randomly chosen employees reveals the data shown below. Because of the casual sampling methodology in this survey, the sample sizes are unequal. *Research question:* Are the mean annual attendance rates the same for employees in these three departments? **DaysWorked**

Days Worked Last Year by 23 Employees

Department	*Days Worked*									
Budgets (5 workers)	278	260	265	245	258					
Payables (10 workers)	205	270	220	240	255	217	266	239	240	228
Pricing (8 workers)	240	258	233	256	233	242	244	249		

11.23 Mean output of solar cells of three types are measured six times under random light intensity over a period of 5 minutes, yielding the results shown. *Research question:* Is the mean solar cell output the same for all cell types? **SolarWatts**

Solar Cell Output (watts)

Cell Type	*Output (watts)*					
A	123	121	123	124	125	127
B	125	122	122	121	122	126
C	126	128	125	129	131	128

11.24 In a bumper test, three types of autos were deliberately crashed into a barrier at 5 mph, and the resulting damage (in dollars) was estimated. Five test vehicles of each type were crashed, with the results shown below. *Research question:* Are the mean crash damages the same for these three vehicles? **Crash1**

Crash Damage ($)

Goliath	*Varmint*	*Weasel*
1,600	1,290	1,090
760	1,400	2,100
880	1,390	1,830
1,950	1,850	1,250
1,220	950	1,920

11.25 The waiting time (in minutes) for emergency room patients with non-life-threatening injuries was measured at four hospitals for all patients who arrived between 6:00 and 6:30 PM on a certain Wednesday. The results are shown below. *Research question:* Are the mean waiting times the same for emergency patients in these four hospitals? **ERWait**

Emergency Room Waiting Time (minutes)

Hospital A (5 patients)	*Hospital B (4 patients)*	*Hospital C (7 patients)*	*Hospital D (6 patients)*
10	8	5	0
19	25	11	20
5	17	24	9
26	36	16	5
11		18	10
		29	12
		15	

11.26 The results shown below are mean productivity measurements (average number of assemblies completed per hour) for a random sample of workers at each of three plants. *Research question:* Are the mean hourly productivity levels the same for workers in these three plants? **Productivity**

Hourly Productivity of Assemblers in Plants

Plant	*Finished Units Produced Per Hour*									
A (9 workers)	3.6	5.1	2.8	4.6	4.7	4.1	3.4	2.9	4.5	
B (6 workers)	2.7	3.1	5.0	1.9	2.2	3.2				
C (10 workers)	6.8	2.5	5.4	6.7	4.6	3.9	5.4	4.9	7.1	8.4

11.27 Below are results of braking tests of the Ford Explorer on glare ice, packed snow, and split traction (one set of wheels on ice, the other on dry pavement), using three braking methods. *Research questions:* Is the mean stopping distance affected by braking method and/or by surface type? **Brake2**

Stopping Distance from 40 mph to 0 mph

Method	*Ice*	*Split Traction*	*Packed Snow*
Pumping	441	223	149
Locked	455	148	146
ABS	460	183	167

Source: *Popular Science* 252, no. 6 (June 1998), p. 78.

11.28 As an independent project, students went to grocery stores and noted the fat grams per serving of various types of bread (from the product label). This was a convenience sample. Manufacturers

with only one product are omitted. *Research question:* Are there differences in the mean fat content (fat per gram) among these seven manufacturers? **BreadFat**

Fat Content of Various Bread Products

Manufacturer	*Product Name*	*Serving Size (grams)*	*Fat Grams Per Serving*	*Fat Grams Per Gram*
Aunt Millie's	ButterMilk—White	34	1	0.0294
Aunt Millie's	Split Top—White	28	1	0.0357
Brownberry	Natural Wheat	36	1	0.0278
Brownberry	Soft Wheat	32	2	0.0625
Brownberry	Whole Wheat	38	1	0.0263
Brownberry	Country Wheat	38	1.5	0.0395
Brownberry	White	38	1	0.0263
Compass Food	America's Choice Light Wheat	21.5	0.5	0.0233
Compass Food	America's Choice Split Top	28	1	0.0357
Interstate Brand Co.	Home Pride Butter Top Wheat	28	1	0.0357
Interstate Brand Co.	Wonder Whole Wheat	34	1.5	0.0441
Interstate Brand Co.	Wonder White	26	1	0.0385
Koepplinger's Bakery	Natural Wheat	38	0	0.0000
Koepplinger's Bakery	Whole Wheat	23	0.5	0.0217
Koepplinger's Bakery	White	26	1	0.0385
Metz Baking Co.	Taystee Wheat	22.5	0.75	0.0333
Metz Baking Co.	Roman Meal Whole Wheat	32	1	0.0313
Metz Baking Co.	Taystee White	26	1	0.0385
Pepperidge Farm	Whole Wheat A	25	1	0.0400
Pepperidge Farm	Light Wheat	19	1	0.0526
Pepperidge Farm	Whole Wheat B	34	1	0.0294
Pepperidge Farm	Pepperidge Farm	32	1.5	0.0469

Source: Class project by statistics students Madonna Klippstein, Nancy Kadarman, Katrina Gagnon, and Bryce Clark.
Note: Data are for educational use and should not be viewed as a guide to current products.

11.29 Is a state's income related to its high school dropout rate? *Research question:* Do the high school dropout rates differ among the five income quintiles? **Dropout**

State High School Dropout Rates by Income Groups

Lowest Income Quintile		2nd Income Quintile		3rd Income Quintile		4th Income Quintile		Highest Income Quintile	
State	*Dropout %*	*State*	*Dropout %*	*State*	*Dropout %*	*State*	*Dropout %*	*State*	*Dropout %*
Mississippi	40.0	Kentucky	34.3	N. Carolina	39.5	Oregon	26.0	Minnesota	15.3
W. Virginia	24.2	S. Carolina	44.5	Wyoming	23.3	Ohio	30.5	Illinois	24.6
New Mexico	39.8	N. Dakota	15.5	Missouri	27.6	Pennsylvania	25.1	California	31.7
Arkansas	27.3	Arizona	39.2	Kansas	25.5	Michigan	27.2	Colorado	28.0
Montana	21.5	Maine	24.4	Nebraska	12.1	Rhode Island	31.3	N. Hampshire	27.0
Louisiana	43.0	S. Dakota	28.1	Texas	39.4	Alaska	33.2	Maryland	27.4
Alabama	39.0	Tennessee	40.1	Georgia	44.2	Nevada	26.3	New York	39.0
Oklahoma	26.9	Iowa	16.8	Florida	42.2	Virginia	25.7	New Jersey	20.4
Utah	16.3	Vermont	19.5	Hawaii	36.0	Delaware	35.9	Massachusetts	25.0
Idaho	22.0	Indiana	28.8	Wisconsin	21.9	Washington	25.9	Connecticut	28.2

Source: *Statistical Abstract of the United States, 2002.*

11.30 In a bumper test, three test vehicles of each of three types of autos were crashed into a barrier at 5 mph, and the resulting damage was estimated. Crashes were from three angles: head-on, slanted, and rear-end. The results are shown on page 482. *Research questions:* Is the mean repair cost affected by crash type and/or vehicle type? Are the observed effects (if any) large enough to be of practical importance (as opposed to statistical significance)? **Crash2**

5 mph Collision Damage ($)

Crash Type	*Goliath*	*Varmint*	*Weasel*
Head-On	700	1,700	2,280
	1,400	1,650	1,670
	850	1,630	1,740
Slant	1,430	1,850	2,000
	1,740	1,700	1,510
	1,240	1,650	2,480
Rear-end	700	860	1,650
	1,250	1,550	1,650
	970	1,250	1,240

11.31 Repeat exercise 11.30 using 6 crashes (not 3) of each type of vehicle and crash angle, as shown below. This is an exact "doubling" of the data set. What effect does doubling the sample size have on your test statistics and conclusions, *ceteris paribus?* Explain fully. **Crash3**

5 mph Collision Damage ($)

Crash Type	*Goliath*	*Varmint*	*Weasel*
Head-On	700	1,700	2,280
	1,400	1,650	1,670
	850	1,630	1,740
	700	1,700	2,280
	1,400	1,650	1,670
	850	1,630	1,740
Slant	1,430	1,850	2,000
	1,740	1,700	1,510
	1,240	1,650	2,480
	1,430	1,850	2,000
	1,740	1,700	1,510
	1,240	1,650	2,480
Rear-end	700	860	1,650
	1,250	1,550	1,650
	970	1,250	1,240
	700	860	1,650
	1,250	1,550	1,650
	970	1,250	1,240

11.32 Three samples of each of three types of PVC pipe of equal wall thickness are tested to failure under three temperature conditions, yielding the results shown below. *Research questions:* Is mean burst strength affected by temperature and/or by pipe type? Is there a "best" brand of PVC pipe? Explain. **PVCPipe**

Burst Strength of PVC Pipes (psi)

Temperature	*PVC1*	*PVC2*	*PVC3*
Hot (70° C)	250	301	235
	273	285	260
	281	275	279
Warm (40° C)	321	342	302
	322	322	315
	299	339	301
Cool (10° C)	358	375	328
	363	355	336
	341	354	342

11.33 The percent of tax returns audited by income taxpayer class is shown below for 6 years. *Research question:* Is the mean tax audit rate affected by taxpayer class and/or by year? **Audits**

IRS Audit Rates by Taxpayer Class

Taxpayer Class	*1990*	*1991*	*1992*	*1993*	*1994*	*1995*
1040A TPI	0.55	0.94	0.78	0.74	1.04	1.96
1040 TPI < $25,000	0.91	1.03	0.92	0.66	0.88	1.30
1040 TPI $25,000–49,999	0.97	0.77	0.70	0.58	0.53	0.90
1040 TPI $50,000–100,000	1.38	1.24	1.10	0.88	0.72	1.05
1040 TPI > $100,000	5.55	5.64	5.28	4.03	2.94	2.79
C-GR < $25,000	1.84	1.91	1.89	2.24	4.39	5.85
C-GR $25,000–100,000	2.35	2.22	2.28	2.41	3.01	3.08
C-GR > $100,000	3.84	3.92	4.17	3.91	3.57	3.47
F-GR < $100,000	1.67	1.53	1.28	1.06	1.16	1.23
F-GR > $100,000	3.09	3.98	2.40	2.06	1.74	2.51

Source: H. Cecil Wayne, "The Real Audit Rates for Individual Taxpayers," 75 *Tax Notes* 831, May 12, 1997.

11.34 To check pain-relieving medications for potential side effects on blood pressure, it is decided to give equal doses of each of four medications to test subjects. To control for the potential effect of weight, subjects are classified by weight groups. Subjects are approximately the same age and are in general good health. Two subjects in each category are chosen at random from a large group of male prison volunteers. Subjects' blood pressures 15 minutes after the dose are shown below. *Research question:* Is mean blood pressure affected by body weight and/or by medication type? **Systolic**

Systolic Blood Pressure of Subjects (mmHg)

Percent of Normal Weight	*Medication M1*	*Medication M2*	*Medication M3*	*Medication M4*
Under 1.1	131	146	140	130
	135	136	132	125
1.1 to 1.3	136	138	134	131
	145	145	147	133
1.3 to 1.5	145	149	146	139
	152	157	151	141

11.35 To assess the effects of instructor and student gender on student course scores, an experiment was conducted in 11 sections of managerial accounting classes ranging in size from 25 to 66 students. The factors were instructor gender (M, F) and student gender (M, F). There were 11 instructors (7 male, 4 female). Steps were taken to eliminate subjectivity in grading, such as common exams and sharing exam grading responsibility among all instructors so no one instructor could influence exam grades unduly. (a) What type of ANOVA is this? (b) What conclusions can you draw? (c) Discuss sample size and raise any questions you think may be important.

Analysis of Variance for Students' Course Scores

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	F *Ratio*	p-*Value*
Instructor gender (I)	97.84	1	97.84	0.61	0.43
Student gender (S)	218.23	1	218.23	1.37	0.24
Interaction ($I \times S$)	743.84	1	743.84	4.66	0.03
Error	63,358.90	397	159.59		
Total	64,418.81	400			

Source: Marlys Gascho Lipe, "Further Evidence on the Performance of Female Versus Male Accounting Students," *Issues in Accounting Education* 4, no. 1 (Spring 1989), pp. 144–50.

11.36 In a market research study, members of a consumer test panel are asked to rate the visual appeal (on a 1 to 10 scale) of the texture of dashboard plastic trim in a mockup of a new fuel cell car. The manufacturer is testing four finish textures. Panelists are assigned randomly to evaluate each texture. The test results are shown below. Each cell shows the average rating by panelists who evaluated each texture. *Research question:* Is mean rating affected by age group and/or by surface type?
Texture

Mean Ratings of Dashboard Surface Texture

Age Group	*Shiny*	*Satin*	*Pebbled*	*Pattern*
Youth (under 21)	6.7	6.6	5.5	4.3
Adult (21 to 39)	5.5	5.3	6.2	5.9
Middle-Age (40 to 61)	4.5	5.1	6.7	5.5
Senior (62 and over)	3.9	4.5	6.1	4.1

11.37 In a call center, the average waiting time for an answer (in seconds) is shown below by time of day. *Research question:* Is mean waiting time affected by time of day and/or by day of the week?
CallWait

Average Waiting Time for Answer (seconds)

Time	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*
06:00	34	71	33	39	39
06:30	52	70	88	53	49
07:00	36	103	47	32	91
07:30	52	97	55	101	37
08:00	46	76	67	74	66
08:30	60	96	46	51	73
09:00	83	34	42	51	79
09:30	32	117	71	57	27
10:00	88	60	54	37	61
10:30	62	152	121	59	47
11:00	45	34	37	79	33
11:30	34	42	28	38	30
12:00	91	37	62	110	51
12:30	42	37	77	63	46
13:00	71	42	49	40	33
13:30	39	125	28	99	57
14:00	132	73	108	81	40
14:30	34	35	38	43	40
15:00	36	34	34	36	36
15:30	34	61	41	46	34
16:00	25	37	52	0	30
16:30	33	26	33	9	36
17:00	27	29	34	38	35
17:30	28	31	27	22	26
18:00	35	14	115	26	22
18:30	25	34	9	5	47

11.38 Several friends go bowling several times per month. They keep track of their scores over several months. An ANOVA was performed. (a) What kind of ANOVA is this (one-factor, two-factor, etc.)? (b) How many friends were there? How many months were observed? How many observations per bowler per month? Explain how you know. (c) What are your conclusions about bowling scores? Explain, referring either to the F tests or p-values.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Month	1702.389	2	851.194	11.9793	0.0002	3.4028
Bowler	4674.000	3	1558.000	21.9265	0.0000	3.0088
Interaction	937.167	6	156.194	2.1982	0.0786	2.5082
Within	1705.333	24	71.056			
Total	9018.889	35				

11.39 Air pollution (micrograms of particulate per ml of air) was measured along four freeways at each of five different times of day, with the results shown below. (a) What kind of ANOVA is this (one-factor, two-factor, etc.)? (b) What is your conclusion about air pollution? Explain, referring either to the *F* tests or *p*-values. (c) Do you think the variances can be assumed equal? Explain your reasoning. Why does it matter? *(d) Perform an F_{max} test to test for unequal variances.

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Chrysler	5	1584	316.8	14333.7
Davidson	5	1047	209.4	3908.8
Reuther	5	714	142.8	2926.7
Lodge	5	1514	302.8	11947.2
12:00A-6:00A	4	505	126.25	872.9
6:00A-10:00A	4	1065	266.25	11060.3
10:00A-3:00P	4	959	239.75	5080.3
3:00P-7:00P	4	1451	362.75	14333.6
7:00P-12:00A	4	879	219.75	7710.9

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Freeway	100957.4	3	33652.45	24.903	0.000	3.490
Time of Day	116249.2	4	29062.3	21.506	0.000	3.259
Error	16216.4	12	1351.367			
Total	233423	19				

11.40 A company has several suppliers of office supplies. It receives several shipments each quarter from each supplier. The time (days) between order and delivery was recorded for several randomly chosen shipments from each supplier in each quarter, and an ANOVA was performed. (a) What kind of ANOVA is this (one-factor, two-factor, etc.)? (b) How many suppliers were there? How many quarters? How many observations per supplier per quarter? Explain how you know. (c) What are your conclusions about shipment time? Explain, referring either to the *F* tests or *p*-values.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Quarter	148.04	3	49.34667	6.0326	0.0009	2.7188
Supplier	410.14	4	102.535	12.5348	0.0000	2.4859
Interaction	247.06	12	20.5883	2.5169	0.0073	1.8753
Within	654.40	80	8.180			
Total	1459.64	99				

11.41 Several friends go bowling several times per month. They keep track of their scores over several months. An ANOVA was performed. (a) What kind of ANOVA is this (one-factor, two-factor, etc.)? (b) How could you tell how many friends there were in the sample just from the ANOVA table? Explain. (c) What are your conclusions about bowling scores? Explain, referring either to the *F* test or *p*-value. (d) Do you think the variances can be assumed equal? Explain your reasoning.

SUMMARY

Bowler	*Count*	*Sum*	*Average*	*Variance*
Mary	15	1856	123.733	77.067
Bill	14	1599	114.214	200.797
Sally	12	1763	146.917	160.083
Robert	15	2211	147.400	83.686
Tom	11	1267	115.182	90.164

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	14465.63	4	3616.408	29.8025	0.0000	2.5201
Within Groups	7523.444	62	121.3459			
Total	21989.07	66				

11.42 Are large companies more profitable *per dollar of assets?* The largest 500 companies in the world in 2000 were ranked according to their number of employees, with groups defined as follows: Small = Under 25,000 employees, Medium = 25,000 to 49,999 employees, Large = 50,000 to 99,000 employees, Huge = 100,000 employees or more. An ANOVA was performed using the company's profit-to-assets ratio (percent) as the dependent variable. (a) What kind of ANOVA is this (one-factor, two-factor, etc.)? (b) What is your conclusion about the research question? Explain, referring either to the *F* test or *p*-value. (c) What can you learn from the plots that compare the groups? (d) Do you think the variances can be assumed equal? Explain your reasoning. *(e) Perform an F_{max} test to test for unequal variances. (f) Which groups of companies have significantly different means? Explain.

Mean	*n*	*Std. Dev*	
1.054	119	2.8475	Small
3.058	113	4.8265	Medium
3.855	147	5.8610	Large
3.843	120	4.3125	Huge
3.004	499	4.7925	Total

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Treatment	643.9030	3	214.63433	9.84	2.58E-06
Error	10,794.2720	495	21.80661		
Total	11,438.1750	498			

Post hoc analysis

Tukey simultaneous comparison t-values (d.f. = 495)

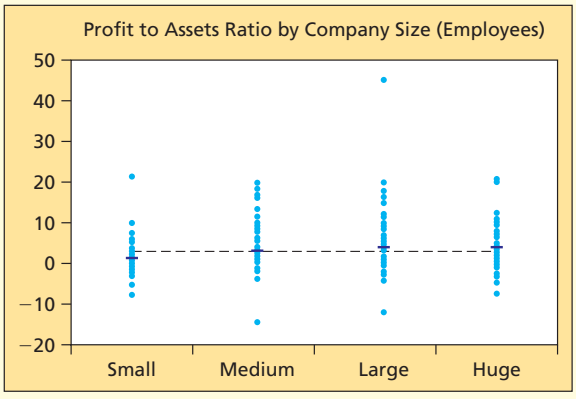
		Small 1.054	Medium 3.058	Huge 3.843	Large 3.855
Small	1.054				
Medium	3.058	3.27			
Huge	3.843	4.62	1.28		
Large	3.855	4.86	1.37	0.02	

critical values for experimentwise error rate:

0.05	2.60
0.01	3.18

p-values for pairwise t-tests

		Small 1.054	Medium 3.058	Huge 3.843	Large 3.855
Small	1.054				
Medium	3.058	.0012			
Huge	3.843	4.97E-06	.2000		
Large	3.855	1.54E-06	.1729	.9833	



Related Reading

Box, George E.; J. Stuart Hunter; and William G. Hunter. *Statistics for Experimenters.* 2nd ed. John Wiley & Sons, 2005.

Hilbe, Joseph M. "Generalized Linear Models." *The American Statistician* 48, no. 3 (August 1994), pp. 255–65.

Miller, Rupert G. *Simultaneous Statistical Inference.* 2nd ed. Springer-Verlag, 1981.

Montgomery, Douglas C. *Design and Analysis of Experiments.* 5th ed. John Wiley & Sons, 2000.

Nachtsheim, Christopher; Michael H. Kutner; and John Neter. *Applied Linear Statistical Models.* 5th ed. McGraw-Hill, 2005.

LearningStats Unit 11 Analysis of Variance

LearningStats Unit 11 gives examples of the three most common ANOVA tests (one-factor, two-factor, full factorial), including a simulation and tables of critical values. Your instructor may assign specific modules, or you may pursue those that sound interesting.

Topic	*LearningStats Modules*
Overview	Overview of ANOVA ANOVA Illustrations
Format and Excel examples	Examples: ANOVA Tests Stacked versus Unstacked Data
Simulation	One-Factor ANOVA ANOVA Data Set Generator
Case studies	One Factor: Car Braking and Noise Two Factors: Car Braking and Noise Two-Factor Replicated: ATM Data Student Project: Call Center Times One Factor: Drug Prices (details) Two Factors: Car Noise (details) Two-Factor Replicated: Braking (details)
General linear model	Insurance Claims Case Study
Tables	Appendix F—Critical Values of *F*

Key: = PowerPoint = Word = Excel

Visual Statistics

Visual Statistics Modules on Analysis of Variance

Module	*Module Name*
12	Visualizing Analysis of Variance

Visual Statistics Module 12 (included on your CD) is designed to help you

- Become familiar with situations in which one-factor ANOVA is applicable.
- Understand how much difference must exist between groups to be detected using an *F* test.
- Appreciate the role of sample size in determining power.
- Know the ANOVA assumptions and the effects of violating them.

The worktext (included on the CD in .PDF format) contains lists of concepts covered, objectives of the modules, overviews of concepts, illustrations of concepts, orientations to module features, learning exercises (basic, intermediate, advanced), learning projects (individual, team), self-evaluation quizzes, glossaries of terms, and solutions to self-evaluation quizzes.