

CHAPTER 3

Experiments with a Single Factor: The Analysis of Variance

CHAPTER OUTLINE

- 3.1 AN EXAMPLE
 - 3.2 THE ANALYSIS OF VARIANCE
 - 3.3 ANALYSIS OF THE FIXED EFFECTS MODEL
 - 3.3.1 Decomposition of the Total Sum of Squares
 - 3.3.2 Statistical Analysis
 - 3.3.3 Estimation of the Model Parameters
 - 3.3.4 Unbalanced Data
 - 3.4 MODEL ADEQUACY CHECKING
 - 3.4.1 The Normality Assumption
 - 3.4.2 Plot of Residuals in Time Sequence
 - 3.4.3 Plot of Residuals Versus Fitted Values
 - 3.4.4 Plots of Residuals Versus Other Variables
 - 3.5 PRACTICAL INTERPRETATION OF RESULTS
 - 3.5.1 A Regression Model
 - 3.5.2 Comparisons Among Treatment Means
 - 3.5.3 Graphical Comparisons of Means
 - 3.5.4 Contrasts
 - 3.5.5 Orthogonal Contrasts
 - 3.5.6 Scheffé's Method for Comparing All Contrasts
 - 3.5.7 Comparing Pairs of Treatment Means
 - 3.5.8 Comparing Treatment Means with a Control
 - 3.6 SAMPLE COMPUTER OUTPUT
 - 3.7 DETERMINING SAMPLE SIZE
 - 3.7.1 Operating Characteristic Curves
 - 3.7.2 Specifying a Standard Deviation Increase
 - 3.7.3 Confidence Interval Estimation Method
 - 3.8 OTHER EXAMPLES OF SINGLE-FACTOR EXPERIMENTS
 - 3.8.1 Chocolate and Cardiovascular Health
 - 3.8.2 A Real Economy Application of a Designed Experiment
 - 3.8.3 Analyzing Dispersion Effects
 - 3.9 THE RANDOM EFFECTS MODEL
 - 3.9.1 A Single Random Factor
 - 3.9.2 Analysis of Variance for the Random Model
 - 3.9.3 Estimating the Model Parameters
 - 3.10 THE REGRESSION APPROACH TO THE ANALYSIS OF VARIANCE
 - 3.10.1 Least Squares Estimation of the Model Parameters
 - 3.10.2 The General Regression Significance Test
 - 3.11 NONPARAMETRIC METHODS IN THE ANALYSIS OF VARIANCE
 - 3.11.1 The Kruskal–Wallis Test
 - 3.11.2 General Comments on the Rank Transformation
- SUPPLEMENTAL MATERIAL FOR CHAPTER 3
- S3.1 The Definition of Factor Effects
 - S3.2 Expected Mean Squares
 - S3.3 Confidence Interval for σ^2
 - S3.4 Simultaneous Confidence Intervals on Treatment Means
 - S3.5 Regression Models for a Quantitative Factor
 - S3.6 More about Estimable Functions
 - S3.7 Relationship Between Regression and Analysis of Variance

The supplemental material is on the textbook website www.wiley.com/college/montgomery.

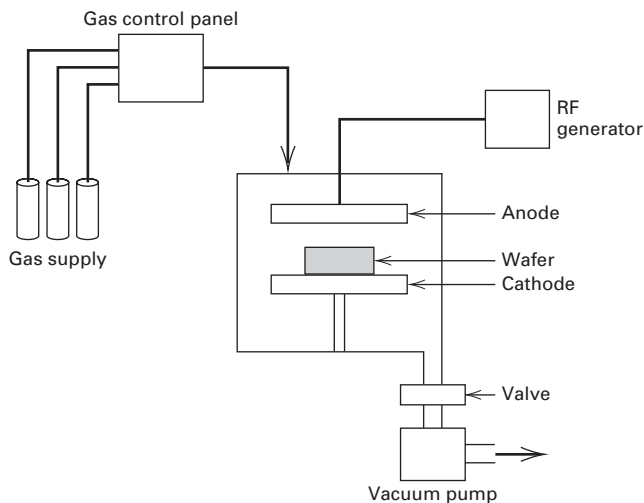
In Chapter 2, we discussed methods for comparing two conditions or treatments. For example, the Portland cement tension bond experiment involved two different mortar formulations. Another way to describe this experiment is as a single-factor experiment with two levels of the factor, where the factor is mortar formulation and the two levels are the two different formulation methods. Many experiments of this type involve more than two levels of the factor. This chapter focuses on methods for the design and analysis of single-factor experiments with an arbitrary number a levels of the factor (or a treatments). We will assume that the experiment has been completely randomized.

3.1 An Example

In many integrated circuit manufacturing steps, wafers are completely coated with a layer of material such as silicon dioxide or a metal. The unwanted material is then selectively removed by etching through a mask, thereby creating circuit patterns, electrical interconnects, and areas in which diffusions or metal depositions are to be made. A plasma etching process is widely used for this operation, particularly in small geometry applications. Figure 3.1 shows the important features of a typical single-wafer etching tool. Energy is supplied by a radio-frequency (RF) generator causing plasma to be generated in the gap between the electrodes. The chemical species in the plasma are determined by the particular gases used. Fluorocarbons, such as CF_4 (tetrafluoromethane) or C_2F_6 (hexafluoroethane), are often used in plasma etching, but other gases and mixtures of gases are relatively common, depending on the application.

An engineer is interested in investigating the relationship between the RF power setting and the etch rate for this tool. The objective of an experiment like this is to model the relationship between etch rate and RF power, and to specify the power setting that will give a desired target etch rate. She is interested in a particular gas (C_2F_6) and gap (0.80 cm) and wants to test four levels of RF power: 160, 180, 200, and 220 W. She decided to test five wafers at each level of RF power.

This is an example of a single-factor experiment with $a = 4$ levels of the factor and $n = 5$ replicates. The 20 runs should be made in random order. A very efficient way to generate the run order is to enter the 20 runs in a spreadsheet (Excel), generate a column of random numbers using the $\text{RAND}()$ function, and then sort by that column.



■ FIGURE 3.1 A single-wafer plasma etching tool

Suppose that the test sequence obtained from this process is given as below:

Test Sequence	Excel Random Number (Sorted)	Power
1	12417	200
2	18369	220
3	21238	220
4	24621	160
5	29337	160
6	32318	180
7	36481	200
8	40062	160
9	43289	180
10	49271	200
11	49813	220
12	52286	220
13	57102	160
14	63548	160
15	67710	220
16	71834	180
17	77216	180
18	84675	180
19	89323	200
20	94037	200

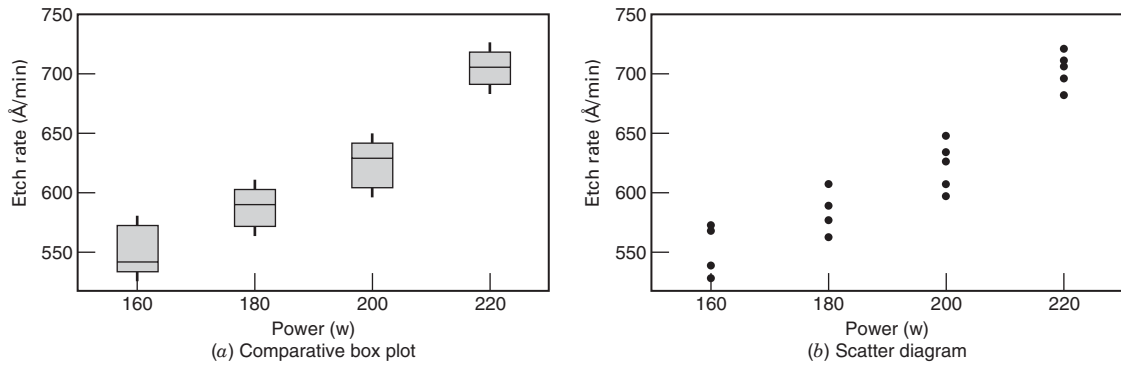
This randomized test sequence is necessary to prevent the effects of unknown nuisance variables, perhaps varying out of control during the experiment, from contaminating the results. To illustrate this, suppose that we were to run the 20 test wafers in the original nonrandomized order (that is, all five 160 W power runs are made first, all five 180 W power runs are made next, and so on). If the etching tool exhibits a warm-up effect such that the longer it is on, the lower the observed etch rate readings will be, the warm-up effect will potentially contaminate the data and destroy the validity of the experiment.

Suppose that the engineer runs the experiment that we have designed in the random order. The observations that she obtains on etch rate are shown in Table 3.1.

It is always a good idea to examine experimental data **graphically**. Figure 3.2*a* presents **box plots** for etch rate at each level of RF power, and Figure 3.2*b* a **scatter diagram** of etch rate versus RF power. Both graphs indicate that etch rate increases as the power setting increases. There

■ **TABLE 3.1**
Etch Rate Data (in Å/min) from the Plasma Etching Experiment

Power (W)	Observations					Totals	Averages
	1	2	3	4	5		
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0



■ FIGURE 3.2 Box plots and scatter diagram of the etch rate data

is no strong evidence to suggest that the variability in etch rate around the average depends on the power setting. On the basis of this simple graphical analysis, we strongly suspect that (1) RF power setting affects the etch rate and (2) higher power settings result in increased etch rate.

Suppose that we wish to be more **objective** in our analysis of the data. Specifically, suppose that we wish to test for differences between the mean etch rates at all $a = 4$ levels of RF power. Thus, we are interested in testing the equality of all four means. It might seem that this problem could be solved by performing a t -test for all six possible pairs of means. However, this is not the best solution to this problem. First of all, performing all six pairwise t -tests is inefficient. It takes a lot of effort. Second, conducting all these pairwise comparisons inflates the type I error. Suppose that all four means are equal, so if we select $\alpha = 0.05$, the probability of reaching the correct decision on any single comparison is 0.95. However, the probability of reaching the correct conclusion on all six comparisons is considerably less than 0.95, so the type I error is inflated.

The appropriate procedure for testing the equality of several means is the **analysis of variance**. However, the analysis of variance has a much wider application than the problem above. It is probably the most useful technique in the field of statistical inference.

3.2 The Analysis of Variance

Suppose we have a **treatments** or different **levels** of a **single factor** that we wish to compare. The observed response from each of the a treatments is a random variable. The data would appear as in Table 3.2. An entry in Table 3.2 (e.g., y_{ij}) represents the j th observation taken under factor level or treatment i . There will be, in general, n observations under the i th treatment. Notice that Table 3.2 is the general case of the data from the plasma etching experiment in Table 3.1.

■ TABLE 3.2
Typical Data for a Single-Factor Experiment

Treatment (Level)	Observations				Totals	Averages
1	y_{11}	y_{12}	...	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	...	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

Models for the Data. We will find it useful to describe the observations from an experiment with a **model**. One way to write this model is

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases} \quad (3.1)$$

where y_{ij} is the ij th observation, μ_i is the mean of the i th factor level or treatment, and ϵ_{ij} is a **random error** component that incorporates all other sources of variability in the experiment including measurement, variability arising from uncontrolled factors, differences between the experimental units (such as test material, etc.) to which the treatments are applied, and the general background noise in the process (such as variability over time, effects of environmental variables, and so forth). It is convenient to think of the errors as having mean zero, so that $E(y_{ij}) = \mu_i$.

Equation 3.1 is called the **means model**. An alternative way to write a model for the data is to define

$$\mu_i = \mu + \tau_i, \quad i = 1, 2, \dots, a$$

so that Equation 3.1 becomes

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases} \quad (3.2)$$

In this form of the model, μ is a parameter common to all treatments called the **overall mean**, and τ_i is a parameter unique to the i th treatment called the **i th treatment effect**. Equation 3.2 is usually called the **effects model**.

Both the means model and the effects model are **linear statistical models**; that is, the response variable y_{ij} is a linear function of the model parameters. Although both forms of the model are useful, the effects model is more widely encountered in the experimental design literature. It has some intuitive appeal in that μ is a constant and the treatment effects τ_i represent deviations from this constant when the specific treatments are applied.

Equation 3.2 (or 3.1) is also called the **one-way or single-factor analysis of variance (ANOVA)** model because only one factor is investigated. Furthermore, we will require that the experiment be performed in random order so that the environment in which the treatments are applied (often called the **experimental units**) is as uniform as possible. Thus, the experimental design is a **completely randomized design**. Our objectives will be to test appropriate hypotheses about the treatment means and to estimate them. For hypothesis testing, the model errors are assumed to be normally and independently distributed random variables with mean zero and variance σ^2 . The variance σ^2 is assumed to be constant for all levels of the factor. This implies that the observations

$$y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

and that the observations are mutually independent.

Fixed or Random Factor? The statistical model, Equation 3.2, describes two different situations with respect to the treatment effects. First, the a treatments could have been specifically chosen by the experimenter. In this situation, we wish to test hypotheses about the treatment means, and our conclusions will apply only to the factor levels considered in the analysis. The conclusions cannot be extended to similar treatments that were not explicitly considered. We may also wish to estimate the model parameters (μ, τ_i, σ^2) . This is called the **fixed effects model**. Alternatively, the a treatments could be a **random sample** from a larger population of treatments. In this situation, we should like to be able to extend the conclusions (which are based on the sample of treatments) to all treatments in the population,

whether or not they were explicitly considered in the analysis. Here, the τ_i are random variables, and knowledge about the particular ones investigated is relatively useless. Instead, we test hypotheses about the variability of the τ_i and try to estimate this variability. This is called the **random effects model** or **components of variance model**. We discuss the single-factor random effects model in Section 3.9. However, we will defer a more complete discussion of experiments with random factors to Chapter 13.

3.3 Analysis of the Fixed Effects Model

In this section, we develop the single-factor analysis of variance for the fixed effects model. Recall that y_i represents the total of the observations under the i th treatment. Let \bar{y}_i represent the average of the observations under the i th treatment. Similarly, let $y_{..}$ represent the grand total of all the observations and $\bar{y}_{..}$ represent the grand average of all the observations. Expressed symbolically,

$$\begin{aligned} y_i &= \sum_{j=1}^n y_{ij} & \bar{y}_i &= y_i/n & i &= 1, 2, \dots, a \\ y_{..} &= \sum_{i=1}^a \sum_{j=1}^n y_{ij} & \bar{y}_{..} &= y_{..}/N \end{aligned} \quad (3.3)$$

where $N = an$ is the total number of observations. We see that the “dot” subscript notation implies summation over the subscript that it replaces.

We are interested in testing the equality of the a treatment means; that is, $E(y_{ij}) = \mu + \tau_i = \mu_i$, $i = 1, 2, \dots, a$. The appropriate hypotheses are

$$\begin{aligned} H_0: \mu_1 &= \mu_2 = \dots = \mu_a \\ H_1: \mu_i &\neq \mu_j \quad \text{for at least one pair } (i, j) \end{aligned} \quad (3.4)$$

In the effects model, we break the i th treatment mean μ_i into two components such that $\mu_i = \mu + \tau_i$. We usually think of μ as an overall mean so that

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

This definition implies that

$$\sum_{i=1}^a \tau_i = 0$$

That is, the treatment or factor effects can be thought of as deviations from the overall mean.¹ Consequently, an equivalent way to write the above hypotheses is in terms of the treatment effects τ_i , say

$$\begin{aligned} H_0: \tau_1 &= \tau_2 = \dots = \tau_a = 0 \\ H_1: \tau_i &\neq 0 \quad \text{for at least one } i \end{aligned}$$

Thus, we speak of testing the equality of treatment means or testing that the treatment effects (the τ_i) are zero. The appropriate procedure for testing the equality of a treatment means is the analysis of variance.

¹ For more information on this subject, refer to the supplemental text material for Chapter 3.

3.3.1 Decomposition of the Total Sum of Squares

The name **analysis of variance** is derived from a partitioning of total variability into its component parts. The total corrected sum of squares

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$$

is used as a measure of overall variability in the data. Intuitively, this is reasonable because if we were to divide SS_T by the appropriate number of degrees of freedom (in this case, $an - 1 = N - 1$), we would have the **sample variance** of the y 's. The sample variance is, of course, a standard measure of variability.

Note that the total corrected sum of squares SS_T may be written as

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^a \sum_{j=1}^n [(\bar{y}_i - \bar{y}_{..}) + (y_{ij} - \bar{y}_i)]^2 \quad (3.5)$$

or

$$\begin{aligned} \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 &= n \sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 \\ &\quad + 2 \sum_{i=1}^a \sum_{j=1}^n (\bar{y}_i - \bar{y}_{..})(y_{ij} - \bar{y}_i) \end{aligned}$$

However, the cross-product term in this last equation is zero, because

$$\sum_{j=1}^n (y_{ij} - \bar{y}_i) = y_i - n\bar{y}_i = y_i - n(y_i/n) = 0$$

Therefore, we have

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 \quad (3.6)$$

Equation 3.6 is the fundamental ANOVA identity. It states that the total variability in the data, as measured by the total corrected sum of squares, can be partitioned into a sum of squares of the differences **between** the treatment averages and the grand average plus a sum of squares of the differences of observations **within** treatments from the treatment average. Now, the difference between the observed treatment averages and the grand average is a measure of the differences between treatment means, whereas the differences of observations within a treatment from the treatment average can be due to only random error. Thus, we may write Equation 3.6 symbolically as

$$SS_T = SS_{\text{Treatments}} + SS_E$$

where $SS_{\text{Treatments}}$ is called the sum of squares due to treatments (i.e., between treatments), and SS_E is called the sum of squares due to error (i.e., within treatments). There are $an = N$ total observations; thus, SS_T has $N - 1$ degrees of freedom. There are a levels of the factor (and a treatment means), so $SS_{\text{Treatments}}$ has $a - 1$ degrees of freedom. Finally, there are n replicates within any treatment providing $n - 1$ degrees of freedom with which to estimate the experimental error. Because there are a treatments, we have $a(n - 1) = an - a = N - a$ degrees of freedom for error.

It is instructive to examine explicitly the two terms on the right-hand side of the fundamental ANOVA identity. Consider the error sum of squares

$$SS_E = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^a \left[\sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 \right]$$

In this form, it is easy to see that the term within square brackets, if divided by $n - 1$, is the sample variance in the i th treatment, or

$$S_i^2 = \frac{\sum_{j=1}^n (y_{ij} - \bar{y}_i)^2}{n - 1} \quad i = 1, 2, \dots, a$$

Now a sample variances may be combined to give a single estimate of the common population variance as follows:

$$\begin{aligned} \frac{(n - 1)S_1^2 + (n - 1)S_2^2 + \dots + (n - 1)S_a^2}{(n - 1) + (n - 1) + \dots + (n - 1)} &= \frac{\sum_{i=1}^a \left[\sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 \right]}{\sum_{i=1}^a (n - 1)} \\ &= \frac{SS_E}{(N - a)} \end{aligned}$$

Thus, $SS_E/(N - a)$ is a **pooled estimate** of the common variance within each of the a treatments.

Similarly, if there were no differences between the a treatment means, we could use the variation of the treatment averages from the grand average to estimate σ^2 . Specifically,

$$\frac{SS_{\text{Treatments}}}{a - 1} = \frac{n \sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2}{a - 1}$$

is an estimate of σ^2 if the treatment means are equal. The reason for this may be intuitively seen as follows: The quantity $\sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2/(a - 1)$ estimates σ^2/n , the variance of the treatment averages, so $n \sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2/(a - 1)$ must estimate σ^2 if there are no differences in treatment means.

We see that the ANOVA identity (Equation 3.6) provides us with two estimates of σ^2 —one based on the inherent variability within treatments and the other based on the variability between treatments. If there are no differences in the treatment means, these two estimates should be very similar, and if they are not, we suspect that the observed difference must be caused by differences in the treatment means. Although we have used an intuitive argument to develop this result, a somewhat more formal approach can be taken.

The quantities

$$MS_{\text{Treatments}} = \frac{SS_{\text{Treatments}}}{a - 1}$$

and

$$MS_E = \frac{SS_E}{N - a}$$

are called **mean squares**. We now examine the **expected values** of these mean squares. Consider

$$\begin{aligned} E(MS_E) &= E\left(\frac{SS_E}{N - a}\right) = \frac{1}{N - a} E\left[\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2\right] \\ &= \frac{1}{N - a} E\left[\sum_{i=1}^a \sum_{j=1}^n (y_{ij}^2 - 2y_{ij}\bar{y}_i + \bar{y}_i^2)\right] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{N-a} E \left[\sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - 2n \sum_{i=1}^a \bar{y}_i^2 + n \sum_{i=1}^a \bar{y}_i^2 \right] \\
&= \frac{1}{N-a} E \left[\sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - \frac{1}{n} \sum_{i=1}^a y_i^2 \right]
\end{aligned}$$

Substituting the model (Equation 3.1) into this equation, we obtain

$$E(MS_E) = \frac{1}{N-a} E \left[\sum_{i=1}^a \sum_{j=1}^n (\mu + \tau_i + \epsilon_{ij})^2 - \frac{1}{n} \sum_{i=1}^a \left(\sum_{j=1}^n \mu + \tau_i + \epsilon_{ij} \right)^2 \right]$$

Now when squaring and taking expectation of the quantity within the brackets, we see that terms involving ϵ_{ij}^2 and ϵ_i^2 are replaced by σ^2 and $n\sigma^2$, respectively, because $E(\epsilon_{ij}) = 0$. Furthermore, all cross products involving ϵ_{ij} have zero expectation. Therefore, after squaring and taking expectation, the last equation becomes

$$E(MS_E) = \frac{1}{N-a} \left[N\mu^2 + n \sum_{i=1}^a \tau_i^2 + N\sigma^2 - N\mu^2 - n \sum_{i=1}^a \tau_i^2 - a\sigma^2 \right]$$

or

$$E(MS_E) = \sigma^2$$

By a similar approach, we may also show that²

$$E(MS_{\text{Treatments}}) = \sigma^2 + \frac{n \sum_{i=1}^a \tau_i^2}{a-1}$$

Thus, as we argued heuristically, $MS_E = SS_E/(N-a)$ estimates σ^2 , and, if there are no differences in treatment means (which implies that $\tau_i = 0$), $MS_{\text{Treatments}} = SS_{\text{Treatments}}/(a-1)$ also estimates σ^2 . However, note that if treatment means do differ, the expected value of the treatment mean square is greater than σ^2 .

It seems clear that a test of the hypothesis of no difference in treatment means can be performed by comparing $MS_{\text{Treatments}}$ and MS_E . We now consider how this comparison may be made.

3.3.2 Statistical Analysis

We now investigate how a formal test of the hypothesis of no differences in treatment means ($H_0: \mu_1 = \mu_2 = \dots = \mu_a$, or equivalently, $H_0: \tau_1 = \tau_2 = \dots = \tau_a = 0$) can be performed. Because we have assumed that the errors ϵ_{ij} are normally and independently distributed with mean zero and variance σ^2 , the observations y_{ij} are normally and independently distributed with mean $\mu + \tau_i$ and variance σ^2 . Thus, SS_T is a sum of squares in normally distributed random variables; consequently, it can be shown that SS_T/σ^2 is distributed as chi-square with $N-1$ degrees of freedom. Furthermore, we can show that SS_E/σ^2 is chi-square with $N-a$ degrees of freedom and that $SS_{\text{Treatments}}/\sigma^2$ is chi-square with $a-1$ degrees of freedom if the null hypothesis $H_0: \tau_i = 0$ is true. However, all three sums of squares are not necessarily independent because $SS_{\text{Treatments}}$ and SS_E add to SS_T . The following theorem, which is a special form of one attributed to William G. Cochran, is useful in establishing the independence of SS_E and $SS_{\text{Treatments}}$.

² Refer to the supplemental text material for Chapter 3.

THEOREM 3-1 Cochran's Theorem

Let Z_i be NID(0, 1) for $i = 1, 2, \dots, \nu$ and

$$\sum_{i=1}^{\nu} Z_i^2 = Q_1 + Q_2 + \dots + Q_s$$

where $s \leq \nu$, and Q_i has ν_i degrees of freedom ($i = 1, 2, \dots, s$). Then Q_1, Q_2, \dots, Q_s are independent chi-square random variables with $\nu_1, \nu_2, \dots, \nu_s$ degrees of freedom, respectively, if and only if

$$\nu = \nu_1 + \nu_2 + \dots + \nu_s$$

Because the degrees of freedom for $SS_{\text{Treatments}}$ and SS_E add to $N - 1$, the total number of degrees of freedom, Cochran's theorem implies that $SS_{\text{Treatments}}/\sigma^2$ and SS_E/σ^2 are independently distributed chi-square random variables. Therefore, if the null hypothesis of no difference in treatment means is true, the ratio

$$F_0 = \frac{SS_{\text{Treatments}}/(a - 1)}{SS_E/(N - a)} = \frac{MS_{\text{Treatments}}}{MS_E} \quad (3.7)$$

is distributed as F with $a - 1$ and $N - a$ degrees of freedom. Equation 3.7 is the **test statistic** for the hypothesis of no differences in treatment means.

From the expected mean squares we see that, in general, MS_E is an unbiased estimator of σ^2 . Also, under the null hypothesis, $MS_{\text{Treatments}}$ is an unbiased estimator of σ^2 . However, if the null hypothesis is false, the expected value of $MS_{\text{Treatments}}$ is greater than σ^2 . Therefore, under the alternative hypothesis, the expected value of the numerator of the test statistic (Equation 3.7) is greater than the expected value of the denominator, and we should reject H_0 on values of the test statistic that are too large. This implies an upper-tail, one-tail critical region. Therefore, we should reject H_0 and conclude that there are differences in the treatment means if

$$F_0 > F_{\alpha, a-1, N-a}$$

where F_0 is computed from Equation 3.7. Alternatively, we could use the P -value approach for decision making. The table of F percentages in the Appendix (Table IV) can be used to find bounds on the P -value.

The sums of squares may be computed in several ways. One direct approach is to make use of the definition

$$y_{ij} - \bar{y}_{..} = (\bar{y}_i - \bar{y}_{..}) + (y_{ij} - \bar{y}_i)$$

Use a spreadsheet to compute these three terms for each observation. Then, sum up the squares to obtain SS_T , $SS_{\text{Treatments}}$, and SS_E . Another approach is to rewrite and simplify the definitions of $SS_{\text{Treatments}}$ and SS_T in Equation 3.6, which results in

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - \frac{y_{..}^2}{N} \quad (3.8)$$

$$SS_{\text{Treatments}} = \frac{1}{n} \sum_{i=1}^a y_i^2 - \frac{y_{..}^2}{N} \quad (3.9)$$

and

$$SS_E = SS_T - SS_{\text{Treatments}} \quad (3.10)$$

TABLE 3.3
The Analysis of Variance Table for the Single-Factor, Fixed Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$SS_{\text{Treatments}} = n \sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2$	$a - 1$	$MS_{\text{Treatments}}$	$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}$
Error (within treatments)	$SS_E = SS_T - SS_{\text{Treatments}}$	$N - a$	MS_E	
Total	$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$	$N - 1$		

This approach is nice because some calculators are designed to accumulate the sum of entered numbers in one register and the sum of the squares of those numbers in another, so each number only has to be entered once. In practice, we use computer software to do this.

The test procedure is summarized in Table 3.3. This is called an **analysis of variance** (or **ANOVA**) table.

EXAMPLE 3.1 **The Plasma Etching Experiment**

To illustrate the analysis of variance, return to the first example discussed in Section 3.1. Recall that the engineer is interested in determining if the RF power setting affects the etch rate, and she has run a completely randomized experiment with four levels of RF power and five replicates. For convenience, we repeat here the data from Table 3.1:

We will use the analysis of variance to test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ against the alternative H_1 : some means are different. The sums of squares required are computed using Equations 3.8, 3.9, and 3.10 as follows:

RF Power (W)	Observed Etch Rate (Å/min)					Totals y_i	Averages \bar{y}_i
	1	2	3	4	5		
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0
						$y_{..} = 12,355$	$\bar{y}_{..} = 617.75$

$$\begin{aligned}
 SS_T &= \sum_{i=1}^4 \sum_{j=1}^5 y_{ij}^2 - \frac{y_{..}^2}{N} \\
 &= (575)^2 + (542)^2 + \dots + (710)^2 - \frac{(12,355)^2}{20} \\
 &= 72,209.75 \\
 SS_{\text{Treatments}} &= \frac{1}{n} \sum_{i=1}^4 y_i^2 - \frac{y_{..}^2}{N} \\
 &= \frac{1}{5} [(2756)^2 + \dots + (3535)^2] - \frac{(12,355)^2}{20} \\
 &= 66,870.55
 \end{aligned}$$

$$\begin{aligned}
 SS_E &= SS_T - SS_{\text{Treatments}} \\
 &= 72,209.75 - 66,870.55 = 5339.20
 \end{aligned}$$

Usually, these calculations would be performed on a computer, using a software package with the capability to analyze data from designed experiments.

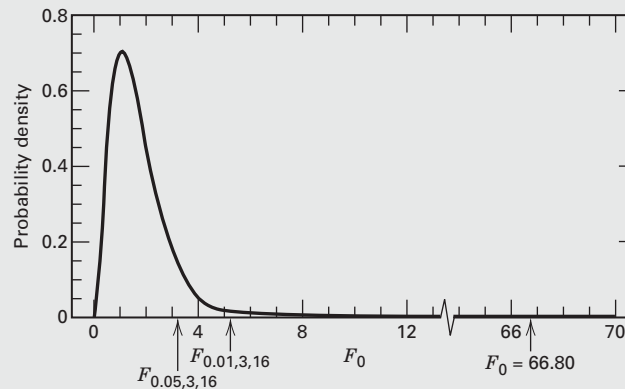
The ANOVA is summarized in Table 3.4. Note that the RF power or between-treatment mean square (22,290.18) is many times larger than the within-treatment or error mean square (333.70). This indicates that it is unlikely that the treatment means are equal. More formally, we

■ **TABLE 3.4**
ANOVA for the Plasma Etching Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P -Value
RF Power	66,870.55	3	22,290.18	$F_0 = 66.80$	<0.01
Error	5339.20	16	333.70		
Total	72,209.75	19			

can compute the F ratio $F_0 = 22,290.18/333.70 = 66.80$ and compare this to an appropriate upper-tail percentage point of the $F_{3,16}$ distribution. To use a fixed significance level approach, suppose that the experimenter has selected $\alpha = 0.05$. From Appendix Table IV we find that $F_{0.05,3,16} = 3.24$. Because $F_0 = 66.80 > 3.24$, we reject H_0 and conclude that the treatment means differ; that is, the RF power setting significantly affects the mean etch

rate. We could also compute a P -value for this test statistic. Figure 3.3 shows the reference distribution ($F_{3,16}$) for the test statistic F_0 . Clearly, the P -value is very small in this case. From Appendix Table A-4, we find that $F_{0.01,3,16} = 5.29$ and because $F_0 > 5.29$, we can conclude that an upper bound for the P -value is 0.01; that is, $P < 0.01$ (the exact P -value is $P = 2.88 \times 10^{-9}$).



■ **FIGURE 3.3** The reference distribution ($F_{3,16}$) for the test statistic F_0 in Example 3.1

Coding the Data. Generally, we need not be too concerned with computing because there are many widely available computer programs for performing the calculations. These computer programs are also helpful in performing many other analyses associated with experimental design (such as residual analysis and model adequacy checking). In many cases, these programs will also assist the experimenter in setting up the design.

However, when hand calculations are necessary, it is sometimes helpful to code the observations. This is illustrated in the next example.

EXAMPLE 3.2 Coding the Observations

The ANOVA calculations may often be made more easily or accurately by **coding** the observations. For example, consider the plasma etching data in Example 3.1. Suppose we subtract 600 from each observation. The coded data are shown in Table 3.5. It is easy to verify that

$$\begin{aligned}
 SS_T &= (-25)^2 + (-58)^2 + \cdots \\
 &\quad + (110)^2 - \frac{(355)^2}{20} = 72,209.75 \\
 SS_{\text{Treatments}} &= \frac{(-244)^2 + (-63)^2 + (127)^2 + (535)^2}{5} \\
 &\quad - \frac{(355)^2}{20} = 66,870.55
 \end{aligned}$$

and

$$SS_E = 5339.20$$

Comparing these sums of squares to those obtained in Example 3.1, we see that subtracting a constant from the original data does not change the sums of squares.

Now suppose that we multiply each observation in Example 3.1 by 2. It is easy to verify that the sums of squares for the transformed data are $SS_T = 288,839.00$, $SS_{\text{Treatments}} = 267,482.20$, and $SS_E = 21,356.80$. These sums of squares appear to differ considerably from those obtained in Example 3.1. However, if they are divided by 4 (i.e., 2^2), the results are identical. For example, for the treatment sum of squares $267,482.20/4 = 66,870.55$. Also, for the coded data, the F ratio is $F = (267,482.20/3)/(21,356.80/16) = 66.80$, which is identical to the F ratio for the original data. Thus, the ANOVAs are equivalent.

■ **TABLE 3.5**
Coded Etch Rate Data for Example 3.2

RF Power (W)	Observations					Totals y_i
	1	2	3	4	5	
160	-25	-58	-70	-61	-30	-244
180	-35	-7	-10	-21	10	-63
200	0	51	10	37	29	127
220	125	100	115	85	110	535

Randomization Tests and Analysis of Variance. In our development of the ANOVA F test, we have used the assumption that the random errors ϵ_{ij} are normally and independently distributed random variables. The F test can also be justified as an approximation to a **randomization test**. To illustrate this, suppose that we have five observations on each of two treatments and that we wish to test the equality of treatment means. The data would look like this:

<u>Treatment 1</u>	<u>Treatment 2</u>
y_{11}	y_{21}
y_{12}	y_{22}
y_{13}	y_{23}
y_{14}	y_{24}
y_{15}	y_{25}

We could use the ANOVA F test to test $H_0: \mu_1 = \mu_2$. Alternatively, we could use a somewhat different approach. Suppose we consider all the possible ways of allocating the 10 numbers in the above sample to the two treatments. There are $10!/5!5! = 252$ possible arrangements of the 10 observations. If there is no difference in treatment means, all 252 arrangements are equally likely. For each of the 252 arrangements, we calculate the value of the F statistic using Equation 3.7. The distribution of these F values is called a **randomization distribution**, and a large value of F indicates that the data are not consistent with the hypothesis $H_0: \mu_1 = \mu_2$. For example, if the value of F actually observed was exceeded by only five of the values of the randomization distribution, this would correspond to rejection of $H_0: \mu_1 = \mu_2$ at a significance level of $\alpha = 5/252 = 0.0198$ (or 1.98 percent). Notice that no normality assumption is required in this approach.

The difficulty with this approach is that, even for relatively small problems, it is computationally prohibitive to enumerate the exact randomization distribution. However, numerous studies have shown that the exact randomization distribution is well approximated by the usual normal-theory F distribution. Thus, even without the normality assumption, the ANOVA F test can be viewed as an approximation to the randomization test. For further reading on randomization tests in the analysis of variance, see Box, Hunter, and Hunter (2005).

3.3.3 Estimation of the Model Parameters

We now present estimators for the parameters in the single-factor model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

and confidence intervals on the treatment means. We will prove later that reasonable estimates of the overall mean and the treatment effects are given by

$$\begin{aligned}\hat{\mu} &= \bar{y}_{..} \\ \hat{\tau}_i &= \bar{y}_{i.} - \bar{y}_{..}, \quad i = 1, 2, \dots, a\end{aligned}\quad (3.11)$$

These estimators have considerable intuitive appeal; note that the overall mean is estimated by the grand average of the observations and that any treatment effect is just the difference between the treatment average and the grand average.

A **confidence interval** estimate of the i th treatment mean may be easily determined. The mean of the i th treatment is

$$\mu_i = \mu + \tau_i$$

A point estimator of μ_i would be $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$ Now, if we assume that the errors are normally distributed, each treatment average $\bar{y}_{i.}$ is distributed $\text{NID}(\mu_i, \sigma^2/n)$. Thus, if σ^2 were known, we could use the normal distribution to define the confidence interval. Using the MS_E as an estimator of σ^2 , we would base the confidence interval on the t distribution. Therefore, a $100(1 - \alpha)$ percent confidence interval on the i th treatment mean μ_i is

$$\bar{y}_{i.} - t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n}} \leq \mu_i \leq \bar{y}_{i.} + t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n}} \quad (3.12)$$

Differences in treatments are frequently of great practical interest. A $100(1 - \alpha)$ percent confidence interval on the difference in any two treatments means, say $\mu_i - \mu_j$, would be

$$\bar{y}_{i.} - \bar{y}_{j.} - t_{\alpha/2, N-a} \sqrt{\frac{2MS_E}{n}} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + t_{\alpha/2, N-a} \sqrt{\frac{2MS_E}{n}} \quad (3.13)$$

EXAMPLE 3.3

Using the data in Example 3.1, we may find the estimates of the overall mean and the treatment effects as $\hat{\mu} = 12,355/20 = 617.75$ and

$$\hat{\tau}_1 = \bar{y}_1 - \bar{y}_{..} = 551.20 - 617.75 = -66.55$$

$$\hat{\tau}_2 = \bar{y}_2 - \bar{y}_{..} = 587.40 - 617.75 = -30.35$$

$$\hat{\tau}_3 = \bar{y}_3 - \bar{y}_{..} = 625.40 - 617.75 = 7.65$$

$$\hat{\tau}_4 = \bar{y}_4 - \bar{y}_{..} = 707.00 - 617.75 = 89.25$$

A 95 percent confidence interval on the mean of treatment 4 (220W of RF power) is computed from Equation 3.12 as

$$707.00 - 2.120 \sqrt{\frac{333.70}{5}} \leq \mu_4 \leq 707.00 + 2.120 \sqrt{\frac{333.70}{5}}$$

or

$$707.00 - 17.32 \leq \mu_4 \leq 707.00 + 17.32$$

Thus, the desired 95 percent confidence interval is $689.68 \leq \mu_4 \leq 724.32$.

Simultaneous Confidence Intervals. The confidence interval expressions given in Equations 3.12 and 3.13 are **one-at-a-time** confidence intervals. That is, the confidence level $1 - \alpha$ applies to only one particular estimate. However, in many problems, the experimenter may wish to calculate several confidence intervals, one for each of a number of means or differences between means. If there are r such $100(1 - \alpha)$ percent confidence intervals of interest, the probability that the r intervals will **simultaneously** be correct is at least $1 - r\alpha$. The probability $r\alpha$ is often called the **experimentwise error rate** or overall confidence coefficient. The number of intervals r does not have to be large before the set of confidence intervals becomes relatively uninformative. For example, if there are $r = 5$ intervals and $\alpha = 0.05$ (a typical choice), the simultaneous confidence level for the set of five confidence intervals is at least 0.75, and if $r = 10$ and $\alpha = 0.05$, the simultaneous confidence level is at least 0.50.

One approach to ensuring that the simultaneous confidence level is not too small is to replace $\alpha/2$ in the one-at-a-time confidence interval Equations 3.12 and 3.13 with $\alpha/(2r)$. This is called the **Bonferroni method**, and it allows the experimenter to construct a set of r simultaneous confidence intervals on treatment means or differences in treatment means for which the overall confidence level is at least $100(1 - \alpha)$ percent. When r is not too large, this is a very nice method that leads to reasonably short confidence intervals. For more information, refer to the **supplemental text material** for Chapter 3.

3.3.4 Unbalanced Data

In some single-factor experiments, the number of observations taken within each treatment may be different. We then say that the design is **unbalanced**. The analysis of variance described above may still be used, but slight modifications must be made in the sum of squares formulas. Let n_i observations be taken under treatment i ($i = 1, 2, \dots, a$) and $N = \sum_{i=1}^a n_i$. The manual computational formulas for SS_T and $SS_{\text{Treatments}}$ become

$$SS_T = \sum_{i=1}^a \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N} \quad (3.14)$$

and

$$SS_{\text{Treatments}} = \sum_{i=1}^a \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} \quad (3.15)$$

No other changes are required in the analysis of variance.

There are two advantages in choosing a balanced design. First, the test statistic is relatively insensitive to small departures from the assumption of equal variances for the a treatments if the sample sizes are equal. This is not the case for unequal sample sizes. Second, the power of the test is maximized if the samples are of equal size.

3.4 Model Adequacy Checking

The decomposition of the variability in the observations through an analysis of variance identity (Equation 3.6) is a purely algebraic relationship. However, the use of the partitioning to test formally for no differences in treatment means requires that certain assumptions be satisfied. Specifically, these assumptions are that the observations are adequately described by the model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

and that the errors are normally and independently distributed with mean zero and constant but unknown variance σ^2 . If these assumptions are valid, the analysis of variance procedure is an exact test of the hypothesis of no difference in treatment means.

In practice, however, these assumptions will usually not hold exactly. Consequently, it is usually unwise to rely on the analysis of variance until the validity of these assumptions has been checked. Violations of the basic assumptions and model adequacy can be easily investigated by the examination of **residuals**. We define the residual for observation j in treatment i as

$$e_{ij} = y_{ij} - \hat{y}_{ij} \quad (3.16)$$

where \hat{y}_{ij} is an estimate of the corresponding observation y_{ij} obtained as follows:

$$\begin{aligned} \hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i \\ &= \bar{y}_{..} + (\bar{y}_i - \bar{y}_{..}) \\ &= \bar{y}_i. \end{aligned} \quad (3.17)$$

Equation 3.17 gives the intuitively appealing result that the estimate of any observation in the i th treatment is just the corresponding treatment average.

Examination of the residuals should be an automatic part of any analysis of variance. If the model is adequate, the residuals should be **structureless**; that is, they should contain no obvious patterns. Through analysis of residuals, many types of model inadequacies and violations of the underlying assumptions can be discovered. In this section, we show how model diagnostic checking can be done easily by graphical analysis of residuals and how to deal with several commonly occurring abnormalities.

3.4.1 The Normality Assumption

A check of the normality assumption could be made by plotting a histogram of the residuals. If the $NID(0, \sigma^2)$ assumption on the errors is satisfied, this plot should look like a sample from a normal distribution centered at zero. Unfortunately, with small samples, considerable fluctuation in the shape of a histogram often occurs, so the appearance of a moderate departure from normality does not necessarily imply a serious violation of the assumptions. Gross deviations from normality are potentially serious and require further analysis.

An extremely useful procedure is to construct a **normal probability plot** of the residuals. Recall from Chapter 2 that we used a normal probability plot of the raw data to check the assumption of normality when using the t -test. In the analysis of variance, it is usually more effective (and straightforward) to do this with the **residuals**. If the underlying error distribution is normal, this plot will resemble a straight line. In visualizing the straight line, place more emphasis on the central values of the plot than on the extremes.

■ **TABLE 3.6**
Etch Rate Data and Residuals from Example 3.1^a

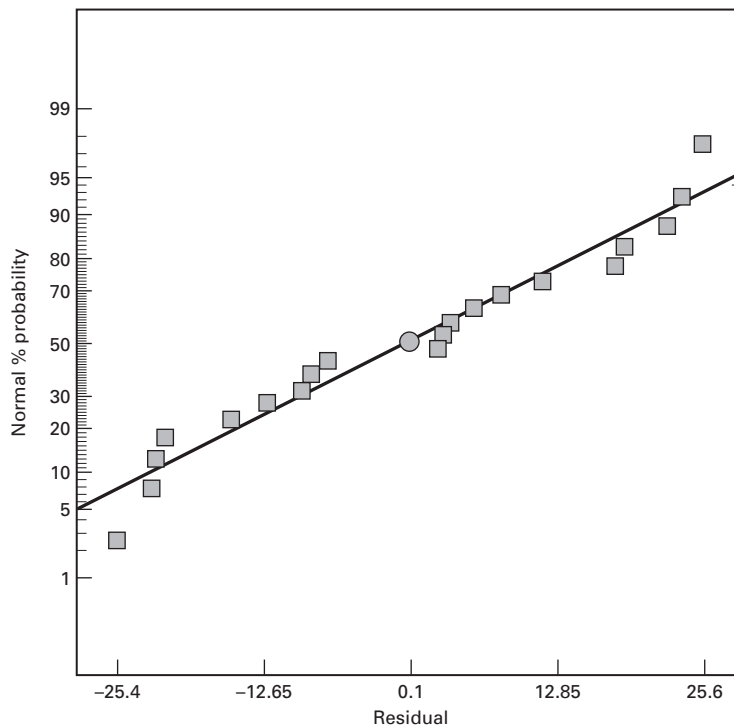
Power (w)	Observations (<i>j</i>)					$\hat{y}_{ij} = \bar{y}_i$
	1	2	3	4	5	
160	23.8 575 (13)	-9.2 542 (14)	-21.2 530 (8)	-12.2 539 (5)	18.8 570 (4)	551.2
180	-22.4 565 (18)	5.6 593 (9)	2.6 590 (6)	-8.4 579 (16)	22.6 610 (17)	587.4
200	-25.4 600 (7)	25.6 651 (19)	-15.4 610 (10)	11.6 637 (20)	3.6 629 (1)	625.4
220	18.0 725 (2)	-7.0 700 (3)	8.0 715 (15)	-22.0 685 (11)	3.0 710 (12)	707.0

^aThe residuals are shown in the box in each cell. The numbers in parentheses indicate the order in which each experimental run was made.

Table 3.6 shows the original data and the residuals for the etch rate data in Example 3.1. The normal probability plot is shown in Figure 3.4. The general impression from examining this display is that the error distribution is approximately normal. The tendency of the normal probability plot to bend down slightly on the left side and upward slightly on the right side implies that the tails of the error distribution are somewhat *thinner* than would be anticipated in a normal distribution; that is, the largest residuals are not quite as large (in absolute value) as expected. This plot is not grossly nonnormal, however.

In general, moderate departures from normality are of little concern in the fixed effects analysis of variance (recall our discussion of randomization tests in Section 3.3.2). An error distribution that has considerably thicker or thinner tails than the normal is of more concern than a skewed distribution. Because the *F* test is only slightly affected, we say that the analysis of

■ **FIGURE 3.4**
Normal probability plot of residuals for Example 3.1



variance (and related procedures such as multiple comparisons) is **robust** to the normality assumption. Departures from normality usually cause both the true significance level and the power to differ slightly from the advertised values, with the power generally being lower. The random effects model that we will discuss in Section 3.9 and Chapter 13 is more severely affected by nonnormality.

A very common defect that often shows up on normal probability plots is one residual that is very much larger than any of the others. Such a residual is often called an **outlier**. The presence of one or more outliers can seriously distort the analysis of variance, so when a potential outlier is located, careful investigation is called for. Frequently, the cause of the outlier is a mistake in calculations or a data coding or copying error. If this is not the cause, the experimental circumstances surrounding this run must be carefully studied. If the outlying response is a particularly desirable value (high strength, low cost, etc.), the outlier may be more informative than the rest of the data. We should be careful not to reject or discard an outlying observation unless we have reasonably nonstatistical grounds for doing so. At worst, you may end up with two analyses; one with the outlier and one without.

Several formal statistical procedures may be used for detecting outliers [e.g., see Stefansky (1972), John and Prescott (1975), and Barnett and Lewis (1994)]. Some statistical software packages report the results of a statistical test for normality (such as the Anderson-Darling test) on the normal probability plot of residuals. This should be viewed with caution as those tests usually assume that the data to which they are applied are independent and residuals are not independent.

A rough check for outliers may be made by examining the **standardized residuals**

$$d_{ij} = \frac{e_{ij}}{\sqrt{MS_E}} \quad (3.18)$$

If the errors ϵ_{ij} are $N(0, \sigma^2)$, the standardized residuals should be approximately normal with mean zero and unit variance. Thus, about 68 percent of the standardized residuals should fall within the limits ± 1 , about 95 percent of them should fall within ± 2 , and virtually all of them should fall within ± 3 . A residual bigger than 3 or 4 standard deviations from zero is a potential outlier.

For the tensile strength data of Example 3.1, the normal probability plot gives no indication of outliers. Furthermore, the largest standardized residual is

$$d_1 = \frac{e_1}{\sqrt{MS_E}} = \frac{25.6}{\sqrt{333.70}} = \frac{25.6}{18.27} = 1.40$$

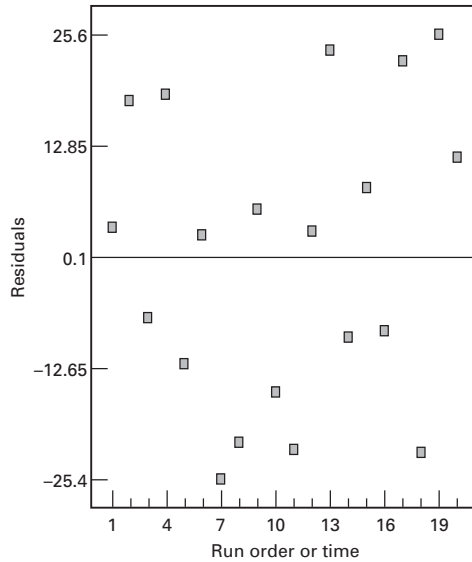
which should cause no concern.

3.4.2 Plot of Residuals in Time Sequence

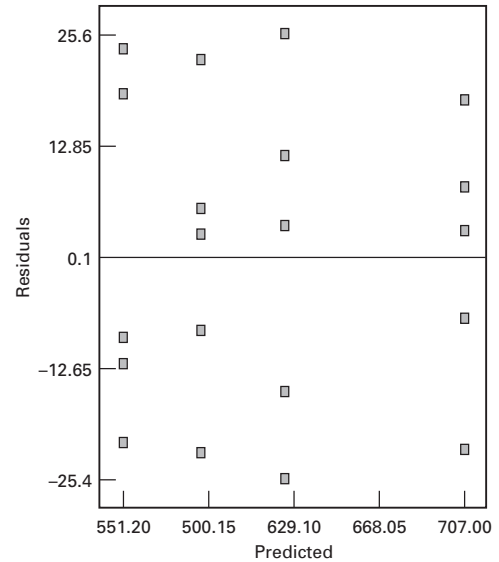
Plotting the residuals in time order of data collection is helpful in detecting strong **correlation** between the residuals. A tendency to have runs of positive and negative residuals indicates positive correlation. This would imply that the **independence assumption** on the errors has been violated. This is a potentially serious problem and one that is difficult to correct, so it is important to prevent the problem if possible when the data are collected. Proper randomization of the experiment is an important step in obtaining independence.

Sometimes the skill of the experimenter (or the subjects) may change as the experiment progresses, or the process being studied may “drift” or become more erratic. This will often result in a change in the error variance over time. This condition often leads to a plot of residuals versus time that exhibits more spread at one end than at the other. Nonconstant variance is a potentially serious problem. We will have more to say on the subject in Sections 3.4.3 and 3.4.4.

Table 3.6 displays the residuals and the time sequence of data collection for the tensile strength data. A plot of these residuals versus run order or time is shown in Figure 3.5. There is no reason to suspect any violation of the independence or constant variance assumptions.



■ FIGURE 3.5 Plot of residuals versus run order or time



■ FIGURE 3.6 Plot of residuals versus fitted values

3.4.3 Plot of Residuals Versus Fitted Values

If the model is correct and the assumptions are satisfied, the residuals should be structureless; in particular, they should be unrelated to any other variable including the predicted response. A simple check is to plot the residuals versus the fitted values \hat{y}_{ij} . (For the single-factor experiment model, remember that $\hat{y}_{ij} = \bar{y}_i$, the i th treatment average.) This plot should not reveal any obvious pattern. Figure 3.6 plots the residuals versus the fitted values for the tensile strength data of Example 3.1. No unusual structure is apparent.

A defect that occasionally shows up on this plot is **nonconstant variance**. Sometimes the variance of the observations increases as the magnitude of the observation increases. This would be the case if the error or background noise in the experiment was a constant percentage of the size of the observation. (This commonly happens with many measuring instruments—error is a percentage of the scale reading.) If this were the case, the residuals would get larger as y_{ij} gets larger, and the plot of residuals versus \hat{y}_{ij} would look like an outward-opening funnel or megaphone. Nonconstant variance also arises in cases where the data follow a nonnormal, skewed distribution because in skewed distributions the variance tends to be a function of the mean.

If the assumption of homogeneity of variances is violated, the F test is only slightly affected in the balanced (equal sample sizes in all treatments) fixed effects model. However, in unbalanced designs or in cases where one variance is very much larger than the others, the problem is more serious. Specifically, if the factor levels having the larger variances also have the smaller sample sizes, the actual type I error rate is larger than anticipated (or confidence intervals have lower actual confidence levels than were specified). Conversely, if the factor levels with larger variances also have the larger sample sizes, the significance levels are smaller than anticipated (confidence levels are higher). This is a good reason for choosing **equal sample sizes** whenever possible. For random effects models, unequal error variances can significantly disturb inferences on variance components even if balanced designs are used.

Inequality of variance also shows up occasionally on the plot of residuals versus run order. An outward-opening funnel pattern indicates that variability is increasing over time. This could result from operator/subject fatigue, accumulated stress on equipment, changes in material properties such as catalyst degradation, or tool wear, or any of a number of causes.

The usual approach to dealing with nonconstant variance when it occurs for the above reasons is to apply a **variance-stabilizing transformation** and then to run the analysis of variance on the transformed data. In this approach, one should note that the conclusions of the analysis of variance apply to the *transformed* populations.

Considerable research has been devoted to the selection of an appropriate transformation. If experimenters know the theoretical distribution of the observations, they may utilize this information in choosing a transformation. For example, if the observations follow the Poisson distribution, the **square root transformation** $y_{ij}^* = \sqrt{y_{ij}}$ or $y_{ij}^* = \sqrt{1 + y_{ij}}$ would be used. If the data follow the lognormal distribution, the **logarithmic transformation** $y_{ij}^* = \log y_{ij}$ is appropriate. For binomial data expressed as fractions, the **arcsin transformation** $y_{ij}^* = \arcsin \sqrt{y_{ij}}$ is useful. When there is no obvious transformation, the experimenter usually *empirically* seeks a transformation that equalizes the variance regardless of the value of the mean. We offer some guidance on this at the conclusion of this section. In factorial experiments, which we introduce in Chapter 5, another approach is to select a transformation that minimizes the interaction mean square, resulting in an experiment that is easier to interpret. In Chapter 15, we discuss in more detail methods for analytically selecting the form of the transformation. Transformations made for inequality of variance also affect the form of the error distribution. In most cases, the transformation brings the error distribution closer to normal. For more discussion of transformations, refer to Bartlett (1947), Dolby (1963), Box and Cox (1964), and Draper and Hunter (1969).

Statistical Tests for Equality of Variance. Although residual plots are frequently used to diagnose inequality of variance, several statistical tests have also been proposed. These tests may be viewed as formal tests of the hypotheses

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

$$H_1: \text{above not true for at least one } \sigma_i^2$$

A widely used procedure is **Bartlett's test**. The procedure involves computing a statistic whose sampling distribution is closely approximated by the chi-square distribution with $a - 1$ degrees of freedom when the a random samples are from independent normal populations. The test statistic is

$$\chi_0^2 = 2.3026 \frac{q}{c} \quad (3.19)$$

where

$$q = (N - a) \log_{10} S_p^2 - \sum_{i=1}^a (n_i - 1) \log_{10} S_i^2$$

$$c = 1 + \frac{1}{3(a - 1)} \left(\sum_{i=1}^a (n_i - 1)^{-1} - (N - a)^{-1} \right)$$

$$S_p^2 = \frac{\sum_{i=1}^a (n_i - 1) S_i^2}{N - a}$$

and S_i^2 is the sample variance of the i th population.

The quantity q is large when the sample variances S_i^2 differ greatly and is equal to zero when all S_i^2 are equal. Therefore, we should reject H_0 on values of χ_0^2 that are too large; that is, we reject H_0 only when

$$\chi_0^2 > \chi_{\alpha, a-1}^2$$

where $\chi_{\alpha, a-1}^2$ is the upper α percentage point of the chi-square distribution with $a - 1$ degrees of freedom. The P -value approach to decision making could also be used.

Bartlett's test is very sensitive to the normality assumption. Consequently, when the validity of this assumption is doubtful, Bartlett's test should not be used.

EXAMPLE 3.4

In the plasma etch experiment, the normality assumption is not in question, so we can apply Bartlett's test to the etch rate data. We first compute the sample variances in each treatment and find that $S_1^2 = 400.7$, $S_2^2 = 280.3$, $S_3^2 = 421.3$, and $S_4^2 = 232.5$. Then

$$S_p^2 = \frac{4(400.7) + 4(280.3) + 4(421.3) + 4(232.5)}{16} = 333.7$$

$$q = 16 \log_{10}(333.7) - 4[\log_{10}400.7 + \log_{10}280.3 + \log_{10}421.3 + \log_{10}232.5] = 0.21$$

$$c = 1 + \frac{1}{3(3)} \left(\frac{4}{4} - \frac{1}{16} \right) = 1.10$$

and the test statistic is

$$\chi_0^2 = 2.3026 \frac{(0.21)}{(1.10)} = 0.43$$

From Appendix Table III, we find that $\chi_{0.05,3}^2 = 7.81$ (the P -value is $P = 0.934$), so we cannot reject the null hypothesis. There is no evidence to counter the claim that all five variances are the same. This is the same conclusion reached by analyzing the plot of residuals versus fitted values.

Because Bartlett's test is sensitive to the normality assumption, there may be situations where an alternative procedure would be useful. Anderson and McLean (1974) present a useful discussion of statistical tests for equality of variance. The **modified Levene test** [see Levene (1960) and Conover, Johnson, and Johnson (1981)] is a very nice procedure that is robust to departures from normality. To test the hypothesis of equal variances in all treatments, the modified Levene test uses the absolute deviation of the observations y_{ij} in each treatment from the treatment median, say, \tilde{y}_i . Denote these deviations by

$$d_{ij} = |y_{ij} - \tilde{y}_i| \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n_i \end{cases}$$

The modified Levene test then evaluates whether or not the means of these deviations are equal for all treatments. It turns out that if the mean deviations are equal, the variances of the observations in all treatments will be the same. The test statistic for Levene's test is simply the usual ANOVA F statistic for testing equality of means applied to the absolute deviations.

EXAMPLE 3.5

A civil engineer is interested in determining whether four different methods of estimating flood flow frequency produce equivalent estimates of peak discharge when applied to the same watershed. Each procedure is used six times on the watershed, and the resulting discharge data (in cubic feet per second) are shown in the upper panel of Table 3.7. The analysis of variance for the data, summarized in Table 3.8, implies that there is a difference in mean peak discharge estimates given by the four procedures. The plot of residuals versus fitted values, shown in Figure 3.7, is disturbing because the outward-opening funnel shape indicates that the constant variance assumption is not satisfied.

We will apply the modified Levene test to the peak discharge data. The upper panel of Table 3.7 contains the treatment medians \tilde{y}_i and the lower panel contains the deviations d_{ij} around the medians. Levene's test consists of conducting a standard analysis of variance on the d_{ij} . The F test statistic that results from this is $F_0 = 4.55$, for which the P -value is $P = 0.0137$. Therefore, Levene's test rejects the null hypothesis of equal variances, essentially confirming the diagnosis we made from visual examination of Figure 3.7. The peak discharge data are a good candidate for data transformation.

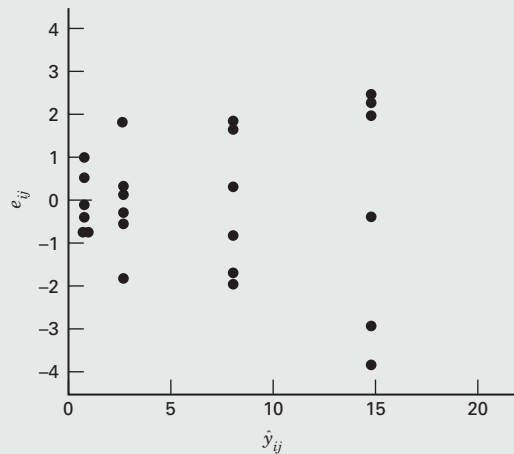
■ **TABLE 3.7**
Peak Discharge Data

Estimation Method		Observations					\bar{y}_i	\tilde{y}_i	S_i
1	0.34	0.12	1.23	0.70	1.75	0.12	0.71	0.520	0.66
2	0.91	2.94	2.14	2.36	2.86	4.55	2.63	2.610	1.09
3	6.31	8.37	9.75	6.09	9.82	7.24	7.93	7.805	1.66
4	17.15	11.82	10.95	17.20	14.35	16.82	14.72	15.59	2.77

Estimation Method		Deviations d_{ij} for the Modified Levene Test				
1	0.18	0.40	0.71	0.18	1.23	0.40
2	1.70	0.33	0.47	0.25	0.25	1.94
3	1.495	0.565	1.945	1.715	2.015	0.565
4	1.56	3.77	4.64	1.61	1.24	1.23

■ **TABLE 3.8**
Analysis of Variance for Peak Discharge Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P -Value
Methods	708.3471	3	236.1157	76.07	<0.001
Error	62.0811	20	3.1041		
Total	770.4282	23			



■ **FIGURE 3.7** Plot of residuals versus \hat{y}_{ij} for Example 3.5

Empirical Selection of a Transformation. We observed above that if experimenters knew the relationship between the variance of the observations and the mean, they could use this information to guide them in selecting the form of the transformation. We now elaborate on this point and show one method for empirically selecting the form of the required transformation from the data.

Let $E(y) = \mu$ be the mean of y , and suppose that the standard deviation of y is proportional to a power of the mean of y such that

$$\sigma_y \propto \mu^\alpha$$

We want to find a transformation on y that yields a constant variance. Suppose that the transformation is a power of the original data, say

$$y^* = y^\lambda \tag{3.20}$$

Then it can be shown that

$$\sigma_{y^*} \propto \mu^{\lambda + \alpha - 1} \tag{3.21}$$

Clearly, if we set $\lambda = 1 - \alpha$, the variance of the transformed data y^* is constant.

Several of the common transformations discussed previously are summarized in Table 3.9. Note that $\lambda = 0$ implies the log transformation. These transformations are arranged in order of increasing **strength**. By the strength of a transformation, we mean the amount of curvature it induces. A mild transformation applied to data spanning a narrow range has little effect on the analysis, whereas a strong transformation applied over a large range may have dramatic results. Transformations often have little effect unless the ratio y_{\max}/y_{\min} is larger than 2 or 3.

In many experimental design situations where there is replication, we can empirically estimate α from the data. Because in the i th treatment combination $\sigma_{y_i} \propto \mu_i^\alpha = \theta \mu_i^\alpha$, where θ is a constant of proportionality, we may take logs to obtain

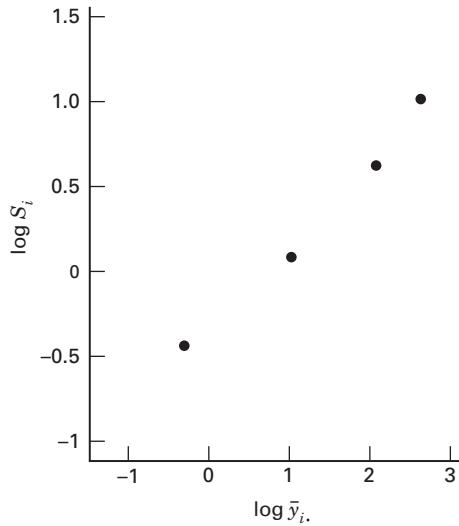
$$\log \sigma_{y_i} = \log \theta + \alpha \log \mu_i \tag{3.22}$$

Therefore, a plot of $\log \sigma_{y_i}$ versus $\log \mu_i$ would be a straight line with slope α . Because we don't know σ_{y_i} and μ_i , we may substitute reasonable estimates of them in Equation 3.22 and use the slope of the resulting straight line fit as an estimate of α . Typically, we would use the standard deviation S_i and the average \bar{y}_i of the i th treatment (or, more generally, the i th treatment combination or set of experimental conditions) to estimate σ_{y_i} and μ_i .

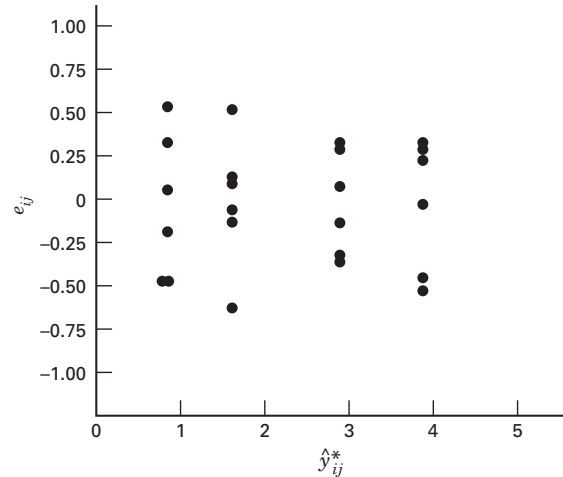
To investigate the possibility of using a variance-stabilizing transformation on the peak discharge data from Example 3.5, we plot $\log S_i$ versus $\log \bar{y}_i$ in Figure 3.8. The slope of a straight line passing through these four points is close to 1/2 and from Table 3.9 this implies that the square root transformation may be appropriate. The analysis of variance for

■ **TABLE 3.9**
Variance-Stabilizing Transformations

Relationship Between σ_y and μ	α	$\lambda = 1 - \alpha$	Transformation	Comment
$\sigma_y \propto \text{constant}$	0	1	No transformation	
$\sigma_y \propto \mu^{1/2}$	1/2	1/2	Square root	Poisson (count) data
$\sigma_y \propto \mu$	1	0	Log	
$\sigma_y \propto \mu^{3/2}$	3/2	-1/2	Reciprocal square root	
$\sigma_y \propto \mu^2$	2	-1	Reciprocal	



■ FIGURE 3.8 Plot of $\log S_i$ versus $\log \bar{y}_i$ for the peak discharge data from Example 3.5



■ FIGURE 3.9 Plot of residuals from transformed data versus \hat{y}_{ij}^* for the peak discharge data in Example 3.5

the transformed data $y^* = \sqrt{y}$ is presented in Table 3.10, and a plot of residuals versus the predicted response is shown in Figure 3.9. This residual plot is much improved in comparison to Figure 3.7, so we conclude that the square root transformation has been helpful. Note that in Table 3.10 we have reduced the degrees of freedom for error and total by 1 to account for the use of the data to estimate the transformation parameter α .

In practice, many experimenters select the form of the transformation by simply trying several alternatives and observing the effect of each transformation on the plot of residuals versus the predicted response. The transformation that produced the most satisfactory residual plot is then selected. Alternatively, there is a formal method called the **Box-Cox Method** for selecting a variance-stability transformation. In chapter 15 we discuss and illustrate this procedure. It is widely used and implemented in many software packages.

3.4.4 Plots of Residuals Versus Other Variables

If data have been collected on any other variables that might possibly affect the response, the residuals should be plotted against these variables. For example, in the tensile strength experiment of Example 3.1, strength may be significantly affected by the thickness of the fiber, so the residuals should be plotted versus fiber thickness. If different testing machines were used to collect the data, the residuals should be plotted against machines. Patterns in such residual plots imply that the variable affects the response. This suggests that the variable should be either controlled more carefully in future experiments or included in the analysis.

■ TABLE 3.10 Analysis of Variance for Transformed Peak Discharge Data, $y^* = \sqrt{y}$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P -Value
Methods	32.6842	3	10.8947	76.99	<0.001
Error	2.6884	19	0.1415		
Total	35.3726	22			

3.5 Practical Interpretation of Results

After conducting the experiment, performing the statistical analysis, and investigating the underlying assumptions, the experimenter is ready to draw practical conclusions about the problem he or she is studying. Often this is relatively easy, and certainly in the simple experiments we have considered so far, this might be done somewhat informally, perhaps by inspection of graphical displays such as the box plots and scatter diagram in Figures 3.1 and 3.2. However, in some cases, more formal techniques need to be applied. We will present some of these techniques in this section.

3.5.1 A Regression Model

The factors involved in an experiment can be either **quantitative** or **qualitative**. A quantitative factor is one whose levels can be associated with points on a numerical scale, such as temperature, pressure, or time. Qualitative factors, on the other hand, are factors for which the levels cannot be arranged in order of magnitude. Operators, batches of raw material, and shifts are typical qualitative factors because there is no reason to rank them in any particular numerical order.

Insofar as the initial design and analysis of the experiment are concerned, both types of factors are treated identically. The experimenter is interested in determining the differences, if any, between the levels of the factors. In fact, the analysis of variance treat the design factor as if it were qualitative or categorical. If the factor is really qualitative, such as operators, it is meaningless to consider the response for a subsequent run at an intermediate level of the factor. However, with a quantitative factor such as time, the experimenter is usually interested in the entire range of values used, particularly the response from a subsequent run at an intermediate factor level. That is, if the levels 1.0, 2.0, and 3.0 hours are used in the experiment, we may wish to predict the response at 2.5 hours. Thus, the experimenter is frequently interested in developing an interpolation equation for the response variable in the experiment. This equation is an **empirical model** of the process that has been studied.

The general approach to fitting empirical models is called **regression analysis**, which is discussed extensively in Chapter 10. See also the **supplemental text material** for this chapter. This section briefly illustrates the technique using the etch rate data of Example 3.1.

Figure 3.10 presents scatter diagrams of etch rate y versus the power x for the experiment in Example 3.1. From examining the scatter diagram, it is clear that there is a strong relationship between etch rate and power. As a first approximation, we could try fitting a **linear model** to the data, say

$$y = \beta_0 + \beta_1 x + \epsilon$$

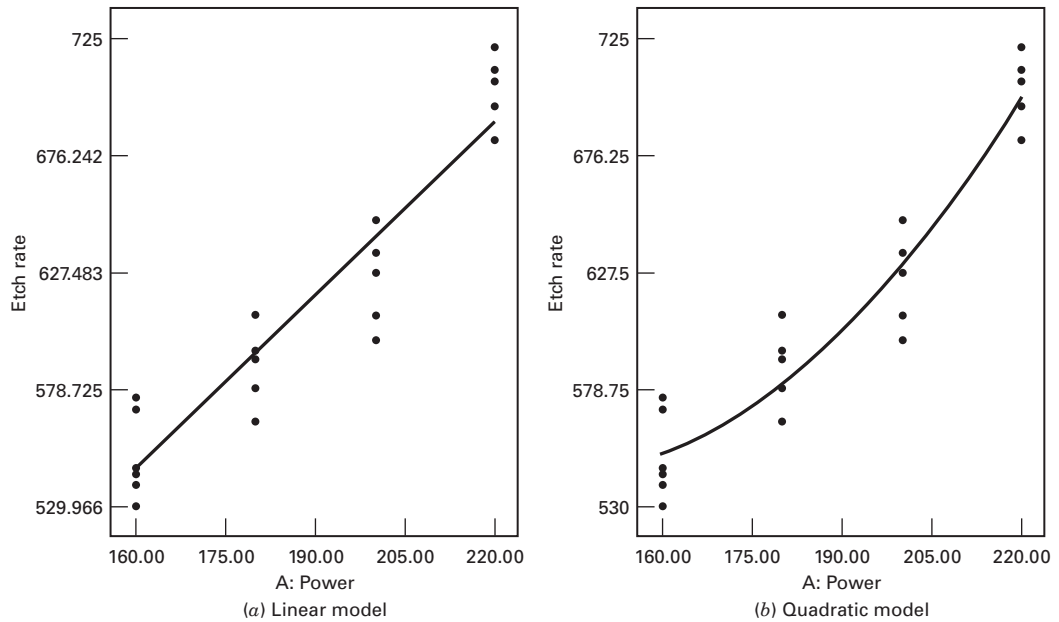
where β_0 and β_1 are unknown parameters to be estimated and ϵ is a random error term. The method often used to estimate the parameters in a model such as this is the **method of least squares**. This consists of choosing estimates of the β 's such that the sum of the squares of the errors (the ϵ 's) is minimized. The least squares fit in our example is

$$\hat{y} = 137.62 + 2.527x$$

(If you are unfamiliar with regression methods, see Chapter 10 and the supplemental text material for this chapter.)

This linear model is shown in Figure 3.10a. It does not appear to be very satisfactory at the higher power settings. Perhaps an improvement can be obtained by adding a quadratic term in x . The resulting **quadratic model** fit is

$$\hat{y} = 1147.77 - 8.2555x + 0.028375x^2$$



■ **FIGURE 3.10** Scatter diagrams and regression models for the etch rate data of Example 3.1

This quadratic fit is shown in Figure 3.10*b*. The quadratic model appears to be superior to the linear model because it provides a better fit at the higher power settings.

In general, we would like to fit the lowest order polynomial that adequately describes the system or process. In this example, the quadratic polynomial seems to fit better than the linear model, so the extra complexity of the quadratic model is justified. Selecting the order of the approximating polynomial is not always easy, however, and it is relatively easy to overfit, that is, to add high-order polynomial terms that do not really improve the fit but increase the complexity of the model and often damage its usefulness as a predictor or interpolation equation.

In this example, the empirical model could be used to predict etch rate at power settings within the region of experimentation. In other cases, the empirical model could be used for **process optimization**, that is, finding the levels of the design variables that result in the best values of the response. We will discuss and illustrate these problems extensively later in the book.

3.5.2 Comparisons Among Treatment Means

Suppose that in conducting an analysis of variance for the fixed effects model the null hypothesis is rejected. Thus, there are differences between the treatment means but exactly *which* means differ is not specified. Sometimes in this situation, further comparisons and analysis among **groups** of treatment means may be useful. The *i*th treatment mean is defined as $\mu_i = \mu + \tau_i$, and μ_i is estimated by \bar{y}_i . Comparisons between treatment means are made in terms of either the treatment totals $\{y_i\}$ or the treatment averages $\{\bar{y}_i\}$. The procedures for making these comparisons are usually called **multiple comparison methods**. In the next several sections, we discuss methods for making comparisons among individual treatment means or groups of these means.

3.5.3 Graphical Comparisons of Means

It is very easy to develop a graphical procedure for the comparison of means following an analysis of variance. Suppose that the factor of interest has a levels and that $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_a$ are the treatment averages. If we know σ , any treatment average would have a standard deviation σ/\sqrt{n} . Consequently, if all factor level means are identical, the observed sample means \bar{y}_i would behave as if they were a set of observations drawn at random from a normal distribution with mean \bar{y}_\cdot and standard deviation σ/\sqrt{n} . Visualize a normal distribution capable of being slid along an axis below which the $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_a$ are plotted. If the treatment means are all equal, there should be some position for this distribution that makes it obvious that the \bar{y}_i values were drawn from the same distribution. If this is not the case, the \bar{y}_i values that appear *not* to have been drawn from this distribution are associated with factor levels that produce different mean responses.

The only flaw in this logic is that σ is unknown. Box, Hunter, and Hunter (2005) point out that we can replace σ with $\sqrt{MS_E}$ from the analysis of variance and use a t distribution with a scale factor $\sqrt{MS_E/n}$ instead of the normal. Such an arrangement for the etch rate data of Example 3.1 is shown in Figure 3.11. Focus on the t distribution shown as a solid line curve in the middle of the display.

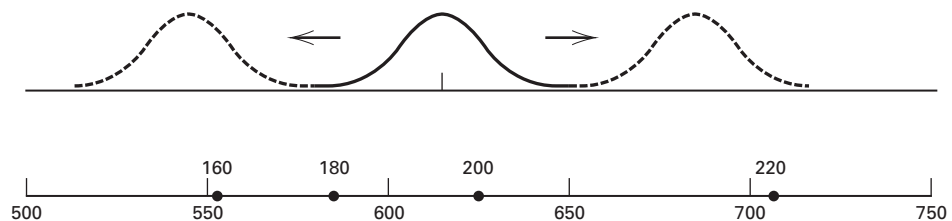
To sketch the t distribution in Figure 3.11, simply multiply the abscissa t value by the scale factor

$$\sqrt{MS_E/n} = \sqrt{330.70/5} = 8.13$$

and plot this against the ordinate of t at that point. Because the t distribution looks much like the normal, except that it is a little flatter near the center and has longer tails, this sketch is usually easily constructed by eye. If you wish to be more precise, there is a table of abscissa t values and the corresponding ordinates in Box, Hunter, and Hunter (2005). The distribution can have an arbitrary origin, although it is usually best to choose one in the region of the \bar{y}_i values to be compared. In Figure 3.11, the origin is 615 Å/min.

Now visualize sliding the t distribution in Figure 3.11 along the horizontal axis as indicated by the dashed lines and examine the four means plotted in the figure. Notice that there is no location for the distribution such that all four averages could be thought of as typical, randomly selected observations from the distribution. This implies that all four means are not equal; thus, the figure is a graphical display of the ANOVA results. Furthermore, the figure indicates that all four levels of power (160, 180, 200, 220 W) produce mean etch rates that differ from each other. In other words, $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$.

This simple procedure is a rough but effective technique for many multiple comparison problems. However, there are more formal methods. We now give a brief discussion of some of these procedures.



■ **FIGURE 3.11** Etch rate averages from Example 3.1 in relation to a t distribution with scale factor $\sqrt{MS_E/n} = \sqrt{330.70/5} = 8.13$

3.5.4 Contrasts

Many multiple comparison methods use the idea of a **contrast**. Consider the plasma etching experiment of Example 3.1. Because the null hypothesis was rejected, we know that some power settings produce different etch rates than others, but which ones actually cause this difference? We might suspect at the outset of the experiment that 200 W and 220 W produce the same etch rate, implying that we would like to test the hypothesis

$$\begin{aligned}H_0: \mu_3 &= \mu_4 \\H_1: \mu_3 &\neq \mu_4\end{aligned}$$

or equivalently

$$\begin{aligned}H_0: \mu_3 - \mu_4 &= 0 \\H_1: \mu_3 - \mu_4 &\neq 0\end{aligned}\tag{3.23}$$

If we had suspected at the start of the experiment that the *average* of the lowest levels of power did not differ from the *average* of the highest levels of power, then the hypothesis would have been

$$\begin{aligned}H_0: \mu_1 + \mu_2 &= \mu_3 + \mu_4 \\H_1: \mu_1 + \mu_2 &\neq \mu_3 + \mu_4\end{aligned}$$

or

$$\begin{aligned}H_0: \mu_1 + \mu_2 - \mu_3 - \mu_4 &= 0 \\H_1: \mu_1 + \mu_2 - \mu_3 - \mu_4 &\neq 0\end{aligned}\tag{3.24}$$

In general, a **contrast** is a linear combination of parameters of the form

$$\Gamma = \sum_{i=1}^a c_i \mu_i$$

where the **contrast constants** c_1, c_2, \dots, c_a sum to zero; that is, $\sum_{i=1}^a c_i = 0$. Both of the above hypotheses can be expressed in terms of contrasts:

$$\begin{aligned}H_0: \sum_{i=1}^a c_i \mu_i &= 0 \\H_1: \sum_{i=1}^a c_i \mu_i &\neq 0\end{aligned}\tag{3.25}$$

The contrast constants for the hypotheses in Equation 3.23 are $c_1 = c_2 = 0, c_3 = +1$, and $c_4 = -1$, whereas for the hypotheses in Equation 3.24, they are $c_1 = c_2 = +1$ and $c_3 = c_4 = -1$.

Testing hypotheses involving contrasts can be done in two basic ways. The first method uses a *t*-test. Write the contrast of interest in terms of the **treatment averages**, giving

$$C = \sum_{i=1}^a c_i \bar{y}_i.$$

The variance of C is

$$V(C) = \frac{\sigma^2}{n} \sum_{i=1}^a c_i^2\tag{3.26}$$

when the sample sizes in each treatment are equal. If the null hypothesis in Equation 3.25 is true, the ratio

$$\frac{\sum_{i=1}^a c_i \bar{y}_i}{\sqrt{\frac{\sigma^2}{n} \sum_{i=1}^a c_i^2}}$$

has the $N(0, 1)$ distribution. Now we would replace the unknown variance σ^2 by its estimate, the mean square error MS_E and use the statistic

$$t_0 = \frac{\sum_{i=1}^a c_i \bar{y}_i}{\sqrt{\frac{MS_E}{n} \sum_{i=1}^a c_i^2}} \quad (3.27)$$

to test the hypotheses in Equation 3.25. The null hypothesis would be rejected if $|t_0|$ in Equation 3.27 exceeds $t_{\alpha/2, N-a}$.

The second approach uses an F test. Now the square of a t random variable with ν degrees of freedom is an F random variable with 1 numerator and ν denominator degrees of freedom. Therefore, we can obtain

$$F_0 = t_0^2 = \frac{\left(\sum_{i=1}^a c_i \bar{y}_i \right)^2}{\frac{MS_E}{n} \sum_{i=1}^a c_i^2} \quad (3.28)$$

as an F statistic for testing Equation 3.25. The null hypothesis would be rejected if $F_0 > F_{\alpha, 1, N-a}$. We can write the test statistic of Equation 3.28 as

$$F_0 = \frac{MS_C}{MS_E} = \frac{SS_C/1}{MS_E}$$

where the single degree of freedom contrast sum of squares is

$$SS_C = \frac{\left(\sum_{i=1}^a c_i \bar{y}_i \right)^2}{\frac{1}{n} \sum_{i=1}^a c_i^2} \quad (3.29)$$

Confidence Interval for a Contrast. Instead of testing hypotheses about a contrast, it may be more useful to construct a confidence interval. Suppose that the contrast of interest is

$$\Gamma = \sum_{i=1}^a c_i \mu_i$$

Replacing the treatment means with the treatment averages yields

$$C = \sum_{i=1}^a c_i \bar{y}_i.$$

Because

$$E\left(\sum_{i=1}^a c_i \bar{y}_i \right) = \sum_{i=1}^a c_i \mu_i \quad \text{and} \quad V(C) = \sigma^2/n \sum_{i=1}^a c_i^2$$

the $100(1 - \alpha)$ percent confidence interval on the contrast $\sum_{i=1}^a c_i \mu_i$ is

$$\sum_{i=1}^a c_i \bar{y}_i - t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n} \sum_{i=1}^a c_i^2} \leq \sum_{i=1}^a c_i \mu_i \leq \sum_{i=1}^a c_i \bar{y}_i + t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n} \sum_{i=1}^a c_i^2} \quad (3.30)$$

Note that we have used MS_E to estimate σ^2 . Clearly, if the confidence interval in Equation 3.30 includes zero, we would be unable to reject the null hypothesis in Equation 3.25.

Standardized Contrast. When more than one contrast is of interest, it is often useful to evaluate them on the same scale. One way to do this is to standardize the contrast so that it has variance σ^2 . If the contrast $\sum_{i=1}^a c_i \mu_i$ is written in terms of treatment averages as $\sum_{i=1}^a c_i \bar{y}_i$, dividing it by $\sqrt{(1/n)\sum_{i=1}^a c_i^2}$ will produce a standardized contrast with variance σ^2 . Effectively, then, the **standardized contrast** is

$$\sum_{i=1}^a c_i^* \bar{y}_i$$

where

$$c_i^* = \frac{c_i}{\sqrt{\frac{1}{n} \sum_{i=1}^a c_i^2}}$$

Unequal Sample Sizes. When the sample sizes in each treatment are different, minor modifications are made in the above results. First, note that the definition of a contrast now requires that

$$\sum_{i=1}^a n_i c_i = 0$$

Other required changes are straightforward. For example, the t statistic in Equation 3.27 becomes

$$t_0 = \frac{\sum_{i=1}^a c_i \bar{y}_i}{\sqrt{MS_E \sum_{i=1}^a \frac{c_i^2}{n_i}}}$$

and the contrast sum of squares from Equation 3.29 becomes

$$SS_C = \frac{\left(\sum_{i=1}^a c_i \bar{y}_i \right)^2}{\sum_{i=1}^a \frac{c_i^2}{n_i}}$$

3.5.5 Orthogonal Contrasts

A useful special case of the procedure in Section 3.5.4 is that of **orthogonal contrasts**. Two contrasts with coefficients $\{c_i\}$ and $\{d_i\}$ are orthogonal if

$$\sum_{i=1}^a c_i d_i = 0$$

or, for an unbalanced design, if

$$\sum_{i=1}^a n_i c_i d_i = 0$$

For a treatments, the set of $a - 1$ orthogonal contrasts partition the sum of squares due to treatments into $a - 1$ independent single-degree-of-freedom components. Thus, tests performed on orthogonal contrasts are independent.

There are many ways to choose the orthogonal contrast coefficients for a set of treatments. Usually, something in the nature of the experiment should suggest which comparisons will be of interest. For example, if there are $a = 3$ treatments, with treatment 1 a control and treatments 2 and 3 actual levels of the factor of interest to the experimenter, appropriate orthogonal contrasts might be as follows:

Treatment	Coefficients for Orthogonal Contrasts	
1 (control)	-2	0
2 (level 1)	1	-1
3 (level 2)	1	1

Note that contrast 1 with $c_i = -2, 1, 1$ compares the average effect of the factor with the control, whereas contrast 2 with $d_i = 0, -1, 1$ compares the two levels of the factor of interest.

Generally, the method of contrasts (or orthogonal contrasts) is useful for what are called **preplanned comparisons**. That is, the contrasts are specified prior to running the experiment and examining the data. The reason for this is that if comparisons are selected after examining the data, most experimenters would construct tests that correspond to large observed differences in means. These large differences could be the result of the presence of real effects, or they could be the result of random error. If experimenters consistently pick the largest differences to compare, they will inflate the type I error of the test because it is likely that, in an unusually high percentage of the comparisons selected, the observed differences will be the result of error. Examining the data to select comparisons of potential interest is often called **data snooping**. The Scheffé method for all comparisons, discussed in the next section, permits data snooping.

EXAMPLE 3.6

Consider the plasma etching experiment in Example 3.1. There are four treatment means and three degrees of freedom between these treatments. Suppose that prior to running the experiment the following set of comparisons among the treatment means (and their associated contrasts) were specified:

Hypothesis	Contrast
$H_0: \mu_1 = \mu_2$	$C_1 = \bar{y}_1 - \bar{y}_2$
$H_0: \mu_1 + \mu_2 = \mu_3 + \mu_4$	$C_2 = \bar{y}_1 + \bar{y}_2 - \bar{y}_3 - \bar{y}_4$
$H_0: \mu_3 = \mu_4$	$C_3 = \bar{y}_3 - \bar{y}_4$

Notice that the contrast coefficients are orthogonal. Using the data in Table 3.4, we find the numerical values of the contrasts and the sums of squares to be as follows:

$$C_1 = +1(551.2) - 1(587.4) = -36.2$$

$$SS_{C_1} = \frac{(-36.2)^2}{\frac{1}{5}(2)} = 3276.10$$

$$C_2 = +1(551.2) + 1(587.4) - 1(625.4) - 1(707.0) = -193.8$$

$$SS_{C_2} = \frac{(-193.8)^2}{\frac{1}{5}(4)} = 46,948.05$$

$$C_3 = +1(625.4) - 1(707.6) = -81.6$$

$$SS_{C_3} = \frac{(-81.6)^2}{\frac{1}{5}(2)} = 16,646.40$$

These contrast sums of squares completely partition the treatment sum of squares. The tests on such orthogonal contrasts are usually incorporated in the ANOVA, as shown in Table 3.11. We conclude from the P -values that there are significant differences in mean etch rates between levels 1 and 2 and between levels 3 and 4 of the power settings, and that the *average* of levels 1 and 2 does differ significantly from the average of levels 3 and 4 at the $\alpha = 0.05$ level.

■ TABLE 3.11
Analysis of Variance for the Plasma Etching Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P -Value
Power setting	66,870.55	3	22,290.18	66.80	<0.001
Orthogonal contrasts					
$C_1: \mu_1 = \mu_2$	(3276.10)	1	3276.10	9.82	<0.01
$C_2: \mu_1 + \mu_3 = \mu_3 + \mu_4$	(46,948.05)	1	46,948.05	140.69	<0.001
$C_3: \mu_3 = \mu_4$	(16,646.40)	1	16,646.40	49.88	<0.001
Error	5,339.20	16	333.70		
Total	72,209.75	19			

3.5.6 Scheffé's Method for Comparing All Contrasts

In many situations, experimenters may not know in advance which contrasts they wish to compare, or they may be interested in more than $a - 1$ possible comparisons. In many exploratory experiments, the comparisons of interest are discovered only after preliminary examination of the data. Scheffé (1953) has proposed a method for comparing any and all possible contrasts between treatment means. In the Scheffé method, the type I error is at most α for any of the possible comparisons.

Suppose that a set of m contrasts in the treatment means

$$\Gamma_u = c_{1u}\mu_1 + c_{2u}\mu_2 + \cdots + c_{au}\mu_a \quad u = 1, 2, \dots, m \quad (3.31)$$

of interest have been determined. The corresponding contrast in the treatment averages \bar{y}_i is

$$C_u = c_{1u}\bar{y}_1 + c_{2u}\bar{y}_2 + \cdots + c_{au}\bar{y}_a \quad u = 1, 2, \dots, m \quad (3.32)$$

and the **standard error** of this contrast is

$$S_{C_u} = \sqrt{MS_E \sum_{i=1}^a (c_{iu}^2/n_i)} \quad (3.33)$$

where n_i is the number of observations in the i th treatment. It can be shown that the critical value against which C_u should be compared is

$$S_{\alpha,u} = S_{C_u} \sqrt{(a-1)F_{\alpha,a-1,N-a}} \quad (3.34)$$

To test the hypothesis that the contrast Γ_u differs significantly from zero, refer C_u to the critical value. If $|C_u| > S_{\alpha,u}$, the hypothesis that the contrast Γ_u equals zero is rejected.

The Scheffé procedure can also be used to form confidence intervals for all possible contrasts among treatment means. The resulting intervals, say $C_u - S_{\alpha,u} \leq \Gamma_u \leq C_u + S_{\alpha,u}$, are **simultaneous confidence intervals** in that the probability that all of them are simultaneously true is at least $1 - \alpha$.

To illustrate the procedure, consider the data in Example 3.1 and suppose that the contrasts of interests are

$$\Gamma_1 = \mu_1 + \mu_2 - \mu_3 - \mu_4$$

and

$$\Gamma_2 = \mu_1 - \mu_4$$

The numerical values of these contrasts are

$$\begin{aligned} C_1 &= \bar{y}_1 + \bar{y}_2 - \bar{y}_3 - \bar{y}_4 \\ &= 551.2 + 587.4 - 625.4 - 707.0 = -193.80 \end{aligned}$$

and

$$\begin{aligned} C_2 &= \bar{y}_1 - \bar{y}_4 \\ &= 551.2 - 707.0 = -155.8 \end{aligned}$$

The standard errors are found from Equation 3.33 as

$$S_{C_1} = \sqrt{MS_E \sum_{i=1}^5 (c_{i1}^2/n_i)} = \sqrt{333.70(1 + 1 + 1 + 1)/5} = 16.34$$

and

$$S_{C_2} = \sqrt{MS_E \sum_{i=1}^5 (c_{i2}^2/n_i)} = \sqrt{333.70(1 + 1)/5} = 11.55$$

From Equation 3.34, the 1 percent critical values are

$$S_{0.01,1} = S_{C_1} \sqrt{(a-1)F_{0.01,a-1,N-a}} = 16.34 \sqrt{3(5.29)} = 65.09$$

and

$$S_{0.01,2} = S_{C_2} \sqrt{(a-1)F_{0.01,a-1,N-a}} = 11.55 \sqrt{3(5.29)} = 45.97$$

Because $|C_1| > S_{0.01,1}$, we conclude that the contrast $\Gamma_1 = \mu_1 + \mu_2 - \mu_3 - \mu_4$ does not equal zero; that is, we conclude that the mean etch rates of power settings 1 and 2 as a group differ from the means of power settings 3 and 4 as a group. Furthermore, because $|C_2| > S_{0.01,2}$, we conclude that the contrast $\Gamma_2 = \mu_1 - \mu_4$ does not equal zero; that is, the mean etch rates of treatments 1 and 4 differ significantly.

3.5.7 Comparing Pairs of Treatment Means

In many practical situations, we will wish to compare only **pairs of means**. Frequently, we can determine which means differ by testing the differences between *all* pairs of treatment means. Thus, we are interested in contrasts of the form $\Gamma = \mu_i - \mu_j$ for all $i \neq j$. Although the Scheffé method described in the previous section could be easily applied to this problem, it is not the most sensitive procedure for such comparisons. We now turn to a consideration of methods specifically designed for pairwise comparisons between all a population means.

Suppose that we are interested in comparing all pairs of a treatment means and that the null hypotheses that we wish to test are $H_0: \mu_i = \mu_j$ for all $i \neq j$. There are numerous procedures available for this problem. We now present two popular methods for making such comparisons.

Tukey’s Test. Suppose that, following an ANOVA in which we have rejected the null hypothesis of equal treatment means, we wish to test all pairwise mean comparisons:

$$\begin{aligned} H_0: \mu_i &= \mu_j \\ H_1: \mu_i &\neq \mu_j \end{aligned}$$

for all $i \neq j$. Tukey (1953) proposed a procedure for testing hypotheses for which the overall significance level is exactly α when the sample sizes are equal and at most α when the sample sizes are unequal. His procedure can also be used to construct confidence intervals on the differences in all pairs of means. For these intervals, the simultaneous confidence level is $100(1 - \alpha)$ percent when the sample sizes are equal and at least $100(1 - \alpha)$ percent when sample sizes are unequal. In other words, the Tukey procedure controls the **experimentwise** or “family” error rate at the selected level α . This is an excellent data snooping procedure when interest focuses on pairs of means.

Tukey’s procedure makes use of the distribution of the **studentized range statistic**

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{MS_E/n}}$$

where \bar{y}_{\max} and \bar{y}_{\min} are the largest and smallest sample means, respectively, out of a group of p sample means. Appendix Table VII contains values of $q_\alpha(p, f)$, the upper α percentage points of q , where f is the number of degrees of freedom associated with the MS_E . For equal sample sizes, Tukey’s test declares two means significantly different if the absolute value of their sample differences exceeds

$$T_\alpha = q_\alpha(a, f) \sqrt{\frac{MS_E}{n}} \tag{3.35}$$

Equivalently, we could construct a set of $100(1 - \alpha)$ percent confidence intervals for all pairs of means as follows:

$$\begin{aligned} \bar{y}_i - \bar{y}_j - q_\alpha(a, f) \sqrt{\frac{MS_E}{n}} &\leq \mu_i - \mu_j \\ &\leq \bar{y}_i - \bar{y}_j + q_\alpha(a, f) \sqrt{\frac{MS_E}{n}}, \quad i \neq j. \end{aligned} \tag{3.36}$$

When sample sizes are not equal, Equations 3.35 and 3.36 become

$$T_\alpha = \frac{q_\alpha(a, f)}{\sqrt{2}} \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \tag{3.37}$$

and

$$\begin{aligned} \bar{y}_i - \bar{y}_j - \frac{q_\alpha(a, f)}{\sqrt{2}} \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} &\leq \mu_i - \mu_j \\ &\leq \bar{y}_i - \bar{y}_j + \frac{q_\alpha(a, f)}{\sqrt{2}} \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}, \quad i \neq j \end{aligned} \tag{3.38}$$

respectively. The unequal sample size version is sometimes called the **Tukey–Kramer procedure**.

EXAMPLE 3.7

To illustrate Tukey's test, we use the data from the plasma etching experiment in Example 3.1. With $\alpha = 0.05$ and $f = 16$ degrees of freedom for error, Appendix Table VII gives $q_{0.05}(4, 16) = 4.05$. Therefore, from Equation 3.35,

$$T_{0.05} = q_{0.05}(4, 16) \sqrt{\frac{MS_E}{n}} = 4.05 \sqrt{\frac{333.70}{5}} = 33.09$$

Thus, any pairs of treatment averages that differ in absolute value by more than 33.09 would imply that the corresponding pair of population means are significantly different. The four treatment averages are

$$\begin{aligned} \bar{y}_1 &= 551.2 & \bar{y}_2 &= 587.4 \\ \bar{y}_3 &= 625.4 & \bar{y}_4 &= 707.0 \end{aligned}$$

and the differences in averages are

$$\begin{aligned} \bar{y}_1 - \bar{y}_2 &= 551.2 - 587.4 = -36.20^* \\ \bar{y}_1 - \bar{y}_3 &= 551.2 - 625.4 = -74.20^* \\ \bar{y}_1 - \bar{y}_4 &= 551.2 - 707.0 = -155.8^* \\ \bar{y}_2 - \bar{y}_3 &= 587.4 - 625.4 = -38.0^* \\ \bar{y}_2 - \bar{y}_4 &= 587.4 - 707.0 = -119.6^* \\ \bar{y}_3 - \bar{y}_4 &= 625.4 - 707.0 = -81.60^* \end{aligned}$$

The starred values indicate the pairs of means that are significantly different. Note that the Tukey procedure indicates that all pairs of means differ. Therefore, each power setting results in a mean etch rate that differs from the mean etch rate at any other power setting.

When using any procedure for pairwise testing of means, we occasionally find that the overall F test from the ANOVA is significant, but the pairwise comparison of means fails to reveal any significant differences. This situation occurs because the F test is simultaneously considering all possible contrasts involving the treatment means, not just pairwise comparisons. That is, in the data at hand, the significant contrasts may not be of the form $\mu_i - \mu_j$.

The derivation of the Tukey confidence interval of Equation 3.36 for equal sample sizes is straightforward. For the studentized range statistic q , we have

$$P\left(\frac{\max(\bar{y}_i - \mu_i) - \min(\bar{y}_i - \mu_i)}{\sqrt{MS_E/n}} \leq q_\alpha(a, f)\right) = 1 - \alpha$$

If $\max(\bar{y}_i - \mu_i) - \min(\bar{y}_i - \mu_i)$ is less than or equal to $q_\alpha(a, f)\sqrt{MS_E/n}$, it must be true that $|(\bar{y}_i - \mu_i) - (\bar{y}_j - \mu_j)| \leq q_\alpha(a, f)\sqrt{MS_E/n}$ for every pair of means. Therefore

$$P\left(-q_\alpha(a, f)\sqrt{\frac{MS_E}{n}} \leq \bar{y}_i - \bar{y}_j - (\mu_i - \mu_j) \leq q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}\right) = 1 - \alpha$$

Rearranging this expression to isolate $\mu_i - \mu_j$ between the inequalities will lead to the set of $100(1 - \alpha)$ percent simultaneous confidence intervals given in Equation 3.38.

The Fisher Least Significant Difference (LSD) Method. The Fisher method for comparing all pairs of means controls the error rate α for each individual pairwise comparison but does not control the experimentwise or family error rate. This procedure uses the t statistic for testing $H_0: \mu_i = \mu_j$

$$t_0 = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad (3.39)$$

Assuming a two-sided alternative, the pair of means μ_i and μ_j would be declared significantly different if $|\bar{y}_i - \bar{y}_j| > t_{\alpha/2, N-a} \sqrt{MS_E(1/n_i + 1/n_j)}$. The quantity

$$\text{LSD} = t_{\alpha/2, N-a} \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (3.40)$$

is called the **least significant difference**. If the design is balanced, $n_1 = n_2 = \dots = n_a = n$, and

$$\text{LSD} = t_{\alpha/2, N-a} \sqrt{\frac{2MS_E}{n}} \quad (3.41)$$

To use the Fisher LSD procedure, we simply compare the observed difference between each pair of averages to the corresponding LSD. If $|\bar{y}_i - \bar{y}_j| > \text{LSD}$, we conclude that the population means μ_i and μ_j differ. The t statistic in Equation 3.39 could also be used.

EXAMPLE 3.8

To illustrate the procedure, if we use the data from the experiment in Example 3.1, the LSD at $\alpha = 0.05$ is

$$\text{LSD} = t_{0.025, 16} \sqrt{\frac{2MS_E}{n}} = 2.120 \sqrt{\frac{2(333.70)}{5}} = 24.49$$

Thus, any pair of treatment averages that differ in absolute value by more than 24.49 would imply that the corresponding pair of population means are significantly different. The differences in averages are

$$\bar{y}_1 - \bar{y}_2 = 551.2 - 587.4 = -36.2^*$$

$$\bar{y}_1 - \bar{y}_3 = 551.2 - 625.4 = -74.2^*$$

$$\bar{y}_1 - \bar{y}_4 = 551.2 - 707.0 = -155.8^*$$

$$\bar{y}_2 - \bar{y}_3 = 587.4 - 625.4 = -38.0^*$$

$$\bar{y}_2 - \bar{y}_4 = 587.4 - 707.0 = -119.6^*$$

$$\bar{y}_3 - \bar{y}_4 = 625.4 - 707.0 = -81.6^*$$

The starred values indicate pairs of means that are significantly different. Clearly, all pairs of means differ significantly.

Note that the overall α risk may be considerably inflated using this method. Specifically, as the number of treatments a gets larger, the experimentwise or family type I error rate (the ratio of the number of experiments in which at least one type I error is made to the total number of experiments) becomes large.

Which Pairwise Comparison Method Do I Use? Certainly, a logical question at this point is, Which one of these procedures should I use? Unfortunately, there is no clear-cut answer to this question, and professional statisticians often disagree over the utility of the various procedures. Carmer and Swanson (1973) have conducted Monte Carlo simulation studies of a number of multiple comparison procedures, including others not discussed here. They report that the least significant difference method is a very effective test for detecting true differences in means if it is applied *only after* the F test in the ANOVA is significant at 5 percent. However, this method does not contain the experimentwise error rate. Because the Tukey method does control the overall error rate, many statisticians prefer to use it.

As indicated above, there are several other multiple comparison procedures. For articles describing these methods, see O'Neill and Wetherill (1971), Miller (1977), and Nelson (1989). The books by Miller (1991) and Hsu (1996) are also recommended.

3.5.8 Comparing Treatment Means with a Control

In many experiments, one of the treatments is a **control**, and the analyst is interested in comparing each of the other $a - 1$ treatment means with the control. Thus, only $a - 1$ comparisons are to be made. A procedure for making these comparisons has been developed by Dunnett (1964). Suppose that treatment a is the control and we wish to test the hypotheses

$$H_0: \mu_i = \mu_a$$

$$H_1: \mu_i \neq \mu_a$$

for $i = 1, 2, \dots, a - 1$. Dunnett's procedure is a modification of the usual t -test. For each hypothesis, we compute the observed differences in the sample means

$$|\bar{y}_i - \bar{y}_a| \quad i = 1, 2, \dots, a - 1$$

The null hypothesis $H_0: \mu_i = \mu_a$ is rejected using a type I error rate α if

$$|\bar{y}_i - \bar{y}_a| > d_\alpha(a - 1, f) \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_a} \right)} \quad (3.42)$$

where the constant $d_\alpha(a - 1, f)$ is given in Appendix Table VIII. (Both two- and one-sided tests are possible.) Note that α is the **joint significance level** associated with all $a - 1$ tests.

EXAMPLE 3.9

To illustrate Dunnett's test, consider the experiment from Example 3.1 with treatment 4 considered as the control. In this example, $a = 4$, $a - 1 = 3$, $f = 16$, and $n_i = n = 5$. At the 5 percent level, we find from Appendix Table VIII that $d_{0.05}(3, 16) = 2.59$. Thus, the critical difference becomes

$$d_{0.05}(3, 16) \sqrt{\frac{2MS_E}{n}} = 2.59 \sqrt{\frac{2(333.70)}{5}} = 29.92$$

(Note that this is a simplification of Equation 3.42 resulting from a balanced design.) Thus, any treatment mean that dif-

fers in absolute value from the control by more than 29.92 would be declared significantly different. The observed differences are

$$1 \text{ vs. } 4: \bar{y}_1 - \bar{y}_4 = 551.2 - 707.0 = -155.8$$

$$2 \text{ vs. } 4: \bar{y}_2 - \bar{y}_4 = 587.4 - 707.0 = -119.6$$

$$3 \text{ vs. } 4: \bar{y}_3 - \bar{y}_4 = 625.4 - 707.0 = -81.6$$

Note that all differences are significant. Thus, we would conclude that all power settings are different from the control.

When comparing treatments with a control, it is a good idea to use more observations for the control treatment (say n_a) than for the other treatments (say n), assuming equal numbers of observations for the remaining $a - 1$ treatments. The ratio n_a/n should be chosen to be approximately equal to the square root of the total number of treatments. That is, choose $n_a/n = \sqrt{a}$.

3.6 Sample Computer Output

Computer programs for supporting experimental design and performing the analysis of variance are widely available. The output from one such program, Design-Expert, is shown in Figure 3.12, using the data from the plasma etching experiment in Example 3.1. The sum of squares corresponding to the “Model” is the usual $SS_{\text{Treatments}}$ for a single-factor design. That source is further identified as “A.” When there is more than one factor in the experiment, the model sum of squares will be decomposed into several sources (A , B , etc.). Notice that the analysis of variance summary at the top of the computer output contains the usual sums of squares, degrees of freedom, mean squares, and test statistic F_0 . The column “Prob > F” is the P -value (actually, the upper bound on the P -value because probabilities less than 0.0001 are defaulted to 0.0001).

In addition to the basic analysis of variance, the program displays some other useful information. The quantity “R-squared” is defined as

$$R^2 = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = \frac{66,870.55}{72,209.75} = 0.9261$$

and is loosely interpreted as the proportion of the variability in the data “explained” by the ANOVA model. Thus, in the plasma etching experiment, the factor “power” explains about 92.61 percent of the variability in etch rate. Clearly, we must have $0 \leq R^2 \leq 1$, with larger values being more desirable. There are also some other R^2 -like statistics displayed in the output. The “adjusted” R^2 is a variation of the ordinary R^2 statistic that reflects the number of factors in the model. It can be a useful statistic for more complex experiments with several design factors when we wish to evaluate the impact of increasing or decreasing the number of model terms. “Std. Dev.” is the square root of the error mean square, $\sqrt{333.70} = 18.27$, and “C.V.” is the coefficient of variation, defined as $(\sqrt{MS_E/\bar{y}})100$. The coefficient of variation measures the unexplained or residual variability in the data as a percentage of the mean of the response variable. “PRESS” stands for “prediction error sum of squares,” and it is a measure of how well the model for the experiment is likely to predict the responses in a *new experiment*. Small values of PRESS are desirable. Alternatively, one can calculate an R^2 for prediction based on PRESS (we will show how to do this later). This R^2_{Pred} in our problem is 0.8845, which is not unreasonable, considering that the model accounts for about 93 percent of the variability in the current experiment. The “adequate precision” statistic is computed by dividing the difference between the maximum predicted response and the minimum predicted response by the average standard deviation of all predicted responses. Large values of this quantity are desirable, and values that exceed four usually indicate that the model will give reasonable performance in prediction.

Treatment means are estimated, and the standard error (or sample standard deviation of each treatment mean, $\sqrt{MS_E/n}$) is displayed. Differences between pairs of treatment means are investigated by using a hypothesis testing version of the Fisher LSD method described in Section 3.5.7.

The computer program also calculates and displays the residuals, as defined in Equation 3.16. The program will also produce all of the residual plots that we discussed in Section 3.4. There are also several other residual diagnostics displayed in the output. Some of these will be discussed later. Design-Expert also displays the studentized residual (called “Student Residual” in the output), calculate as

$$r_{ij} = \frac{e_{ij}}{\sqrt{MS_E(1 - \text{Leverage}_{ij})}}$$

where Leverage_{ij} is a measure of the influence of the ij^{th} observation on the model. We will discuss leverage in more detail and show how it is calculated in chapter 10. Studentized residuals

Response: Etch Rate

ANOVA for Selected Factorial Model
Analysis of variance table [Partial sum of squares]

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F
Model	66870.55	3	22290.18	66.80	<0.0001 significant
A	66870.55	3	22290.18	66.80	<0.0001
Pure Error	5338.20	16	333.70		
Cor Total	72209.75	19			

The Model F-value of 66.80 implies that the model is significant. There is only a 0.01% chance that a "Model F-Value" this large could occur due to noise.

Values of "Prob > F" less than 0.0500 indicate that model terms are significant. In this case, A are significant model terms. Values greater than 0.1000 indicate that the model terms are not significant. If there are many insignificant model terms (not counting those required to support hierarchy), model reduction may improve your model.

Std. Dev.	18.27	R-Squared	0.9261
Mean	617.75	Adj R-Squared	0.9122
C.V.	2.96	Pred R-Squared	0.8846
PRESS	8342.50	Adeq Precision	19.071

The "Pred R-Squared" of 0.8845 is in reasonable agreement with the "Adj R-Squared" of 0.9122.

"Adeq Precision" measures the signal-to-noise ratio. A ratio greater than four is desirable. Your ratio of 19.071 indicates an adequate signal. This model can be used to navigate the design space.

Treatment Means (Adjusted, if Necessary)

	Estimated Mean	Standard Error
1-160	551.20	8.17
2-180	587.40	8.17
3-200	625.40	8.17
4-220	707.00	8.17

Treatment	Mean Difference	DF	Standard Error	t for H ₀ Coeff=0	Prob > t
1 vs 2	-36.20	1	11.55	-3.13	0.0064
1 vs 3	-74.20	1	11.55	-6.42	<0.0001
1 vs 4	-155.80	1	11.55	-13.49	<0.0001
2 vs 3	-38.00	1	11.55	-3.29	0.0046
2 vs 4	-119.60	1	11.55	-10.35	<0.0001
3 vs 4	-81.60	1	11.55	-7.06	<0.0001

Values of "Prob > |t|" less than 0.0500 indicate that the difference in the treatment means is significant. Values of "Prob > |t|" greater than 0.1000 indicate that the difference in the two treatment means is not significant.

Diagnostics Case Statistics

Standard Order	Actual Value	Predicted Value	Residual	Leverage	Student Residual	Cook's Distance	Outlier t	Run Order
1	575.00	551.20	23.80	0.200	1.457	0.133	1.514	13
2	542.00	551.20	-9.20	0.200	-0.563	0.020	-0.551	14
3	530.00	551.20	-21.20	0.200	-1.298	0.105	-1.328	8
4	539.00	551.20	-12.20	0.200	-0.747	0.035	-0.736	5
5	570.00	551.20	18.80	0.200	1.151	0.083	1.163	4
6	565.00	587.40	-22.40	0.200	-1.371	0.117	-1.413	18
7	593.00	587.40	5.60	0.200	0.343	0.007	0.333	9
8	590.00	587.40	2.60	0.200	0.159	0.002	0.154	6
9	579.00	587.40	-8.40	0.200	-0.514	0.017	-0.502	16
10	610.00	587.40	22.60	0.200	1.383	0.120	1.427	17
11	600.00	625.40	-25.40	0.200	-1.555	0.151	-1.634	7
12	651.00	625.40	25.60	0.200	1.567	0.153	1.649	19
13	610.00	625.40	-15.40	0.200	-0.943	0.056	-0.939	10
14	637.00	625.40	11.60	0.200	0.710	0.032	0.699	20
15	629.00	625.40	3.60	0.200	0.220	0.003	0.214	1
16	725.00	707.00	18.00	0.200	1.102	0.076	1.110	2
17	700.00	707.00	-7.00	0.200	-0.428	0.011	-0.417	3
18	715.00	707.00	8.00	0.200	0.490	0.015	0.478	15
19	685.00	707.00	-22.00	0.200	-1.346	0.113	-1.385	11
20	710.00	707.00	3.00	0.200	0.184	0.002	0.178	12

- Proceed to Diagnostic Plots (the next icon in progression). Be sure to look at the
- (1) Normal probability plot of the studentized residuals to check for normality of residuals.
 - (2) Studentized residuals versus predicted values to check for constant error.
 - (3) Outlier t versus run order to look for outliers, i.e., influential values.
 - (4) Box-Cox plot for power transformations.

If all the model statistics and diagnostic plots are OK, finish up with the Model Graphs icon.

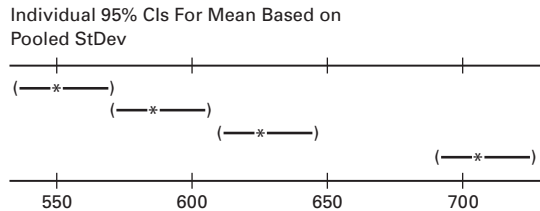
■ **FIGURE 3.12** Design-Expert computer output for Example 3.1

One-way ANOVA: Etch Rate versus Power

Source	DF	SS	MS	F	P
Power	3	66871	22290	66.80	0.000
Error	16	5339	334		
Total	19	72210			

S = 18.27 R-Sq = 92.61% R-Sq (adj) = 91.22%

Level	N	Mean	Std.Dev.
160	5	551.20	20.02
180	5	587.40	16.74
200	5	625.40	20.53
220	5	707.00	15.25



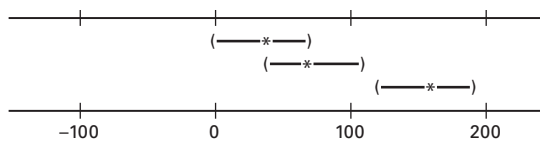
Pooled Std. Dev. = 18.27

Turkey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Power

Individual confidence level = 98.87%

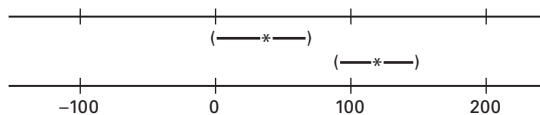
Power = 160 subtracted from

Power	Lower	Center	Upper
180	3.11	36.20	69.29
200	41.11	74.20	107.29
220	122.71	155.80	188.89



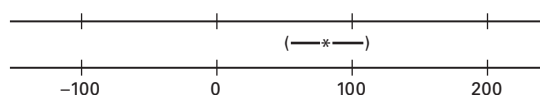
Power = 180 subtracted from

Power	Lower	Center	Upper
200	4.91	38.00	71.09
220	86.51	119.60	152.69



Power = 200 subtracted from

Power	Lower	Center	Upper
220	48.51	81.60	114.69



■ **FIGURE 3.13** Minitab computer output for Example 3.1

are considered to be more effective in identifying potential rather than either the ordinary residuals or standardized residuals.

Finally, notice that the computer program also has some interpretative guidance embedded in the output. This “advisory” information is fairly standard in many PC-based statistics packages. Remember in reading such guidance that it is written in very general terms and may not exactly suit the report writing requirements of any specific experimenter. This advisory output may be hidden upon request by the user.

Figure 3.13 presents the output from Minitab for the plasma etching experiment. The output is very similar to the Design-Expert output in Figure 3.12. Note that confidence intervals on each individual treatment mean are provided and that the pairs of means are compared using Tukey’s method. However, the Tukey method is presented using the confidence interval format instead of the hypothesis-testing format that we used in Section 3.5.7. None of the Tukey confidence intervals includes zero, so we would conclude that all of the means are different.

Figure 3.14 is the output from JMP for the plasma etch experiment in Example 3.1. The output information is very similar to that from Design-Expert and Minitab. The plots of actual observations versus the predicted values and residuals versus the predicted values are default output. There is an option in JMP to provide the Fisher LSD procedure or Tukey's method to compare all pairs of means.

3.7 Determining Sample Size

In any experimental design problem, a critical decision is the choice of sample size—that is, determining the number of replicates to run. Generally, if the experimenter is interested in detecting small effects, more replicates are required than if the experimenter is interested in detecting large effects. In this section, we discuss several approaches to determining sample size. Although our discussion focuses on a single-factor design, most of the methods can be used in more complex experimental situations.

3.7.1 Operating Characteristic Curves

Recall that an **operating characteristic (OC) curve** is a plot of the type II error probability of a statistical test for a particular sample size versus a parameter that reflects the extent to which the null hypothesis is false. These curves can be used to guide the experimenter in selecting the number of replicates so that the design will be sensitive to important potential differences in the treatments.

We consider the probability of type II error of the fixed effects model for the case of equal sample sizes per treatment, say

$$\begin{aligned}\beta &= 1 - P\{\text{Reject } H_0 | H_0 \text{ is false}\} \\ &= 1 - P\{F_0 > F_{\alpha, a-1, N-a} | H_0 \text{ is false}\}\end{aligned}\quad (3.43)$$

To evaluate the probability statement in Equation 3.43, we need to know the distribution of the test statistic F_0 if the null hypothesis is false. It can be shown that, if H_0 is false, the statistic $F_0 = MS_{\text{Treatments}}/MS_E$ is distributed as a **noncentral F** random variable with $a - 1$ and $N - a$ degrees of freedom and the noncentrality parameter δ . If $\delta = 0$, the noncentral F distribution becomes the usual (central) F distribution.

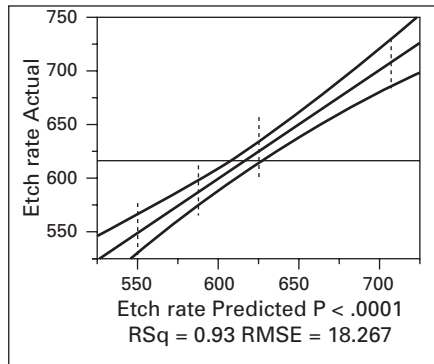
Operating characteristic curves given in Chart V of the Appendix are used to evaluate the probability statement in Equation 3.43. These curves plot the probability of type II error (β) against a parameter Φ , where

$$\Phi^2 = \frac{n \sum_{i=1}^a \tau_i^2}{a\sigma^2}\quad (3.44)$$

The quantity Φ^2 is related to the noncentrality parameter δ . Curves are available for $\alpha = 0.05$ and $\alpha = 0.01$ and a range of degrees of freedom for numerator and denominator.

In using the OC curves, the experimenter must specify the parameter Φ and the value of σ^2 . This is often difficult to do in practice. One way to determine Φ is to choose the actual values of the treatment means for which we would like to reject the null hypothesis with high probability. Thus, if $\mu_1, \mu_2, \dots, \mu_a$ are the specified treatment means, we find the τ_i in Equation 3.48 as $\tau_i = \mu_i - \bar{\mu}$, where $\bar{\mu} = (1/a)\sum_{i=1}^a \mu_i$ is the average of the individual treatment means. The estimate of σ^2 may be available from prior experience, a previous experiment or a preliminary test (as suggested in Chapter 1), or a judgment estimate. When we are uncertain about the value of σ^2 , sample sizes could be determined for a range of likely values of σ^2 to study the effect of this parameter on the required sample size before a final choice is made.

Response Etch rate
Whole Model
Actual by Predicted Plot



Summary of Fit

RSquare	0.92606
RSquare Adj	0.912196
Root Mean Square Error	18.26746
Mean of Response	617.75
Observations (or Sum Wgts)	20

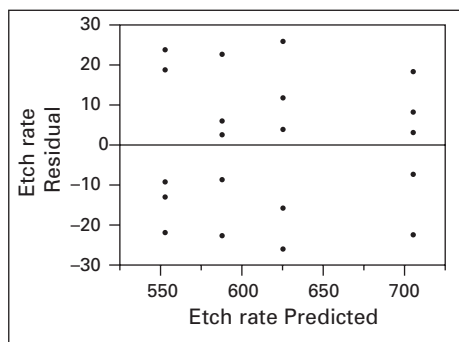
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	66870.550	22290.2	66.7971
Error	16	5339.200	333.7	Prob > F
C. Total	19	72209.750		<.0001

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
RF power	3	3	66870.550	66.7971	<.0001

Residual by Predicted Plot



RF power

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
160	551.20000	8.1694553	551.200
180	587.40000	8.1694553	587.400
200	625.40000	8.1694553	625.400
220	707.00000	8.1694553	707.000

■ **FIGURE 3.14** JMP output from Example 3.1

EXAMPLE 3.10

Consider the plasma etching experiment described in Example 3.1. Suppose that the experimenter is interested in rejecting the null hypothesis with a probability of at least 0.90 if the four treatment means are

$\mu_1 = 575$ $\mu_2 = 600$ $\mu_3 = 650$ and $\mu_4 = 675$
 She plans to use $\alpha = 0.01$. In this case, because $\sum_{i=1}^4 \mu_i = 2500$, we have $\bar{\mu} = (1/4)2500 = 625$ and

$$\tau_1 = \mu_1 - \bar{\mu} = 575 - 625 = -50$$

$$\tau_2 = \mu_2 - \bar{\mu} = 600 - 625 = -25$$

$$\tau_3 = \mu_3 - \bar{\mu} = 650 - 625 = 25$$

$$\tau_4 = \mu_4 - \bar{\mu} = 675 - 625 = 50$$

Thus, $\sum_{i=1}^4 \tau_i^2 = 6250$. Suppose the experimenter feels that the standard deviation of etch rate at any particular level of

power will be no larger than $\sigma = 25$ Å/min. Then, by using Equation 3.44, we have

$$\Phi^2 = \frac{n \sum_{i=1}^4 \tau_i^2}{a\sigma^2} = \frac{n(6,250)}{4(25)^2} = 2.5n$$

We use the OC curve for $a - 1 = 4 - 1 = 3$ with $N - a = a(n - 1) = 4(n - 1)$ error degrees of freedom and $\alpha = 0.01$ (see Appendix Chart V). As a first guess at the required sample size, try $n = 3$ replicates. This yields $\Phi^2 = 2.5n = 2.5(3) = 7.5$, $\Phi = 2.74$, and $4(2) = 8$ error degrees of freedom. Consequently, from Chart V, we find that $\beta \approx 0.25$. Therefore, the power of the test is approximately $1 - \beta = 1 - 0.25 = 0.75$, which is less than the required 0.90, and so we conclude that $n = 3$ replicates are not sufficient. Proceeding in a similar manner, we can construct the following display:

n	Φ^2	Φ	$a(n - 1)$	β	Power ($1 - \beta$)
3	7.5	2.74	8	0.25	0.75
4	10.0	3.16	12	0.04	0.96
5	12.5	3.54	16	<0.01	>0.99

Thus, 4 or 5 replicates are sufficient to obtain a test with the required power.

A significant problem with this approach to using OC curves is that it is usually difficult to select a set of treatment means on which the sample size decision should be based. An alternate approach is to select a sample size such that if the difference between any two treatment means exceeds a specified value, the null hypothesis should be rejected. If the difference between any two treatment means is as large as D , it can be shown that the minimum value of Φ^2 is

$$\Phi^2 = \frac{nD^2}{2a\sigma^2} \quad (3.45)$$

Because this is a minimum value of Φ^2 , the corresponding sample size obtained from the operating characteristic curve is a conservative value; that is, it provides a power at least as great as that specified by the experimenter.

To illustrate this approach, suppose that in the plasma etching experiment from Example 3.1, the experimenter wished to reject the null hypothesis with probability at least 0.90 if any two treatment means differed by as much as 75 Å/min and $\alpha = 0.01$. Then, assuming that $\sigma = 25$ psi, we find the minimum value of Φ^2 to be

$$\Phi^2 = \frac{n(75)^2}{2(4)(25^2)} = 1.125n$$

Now we can use the OC curves exactly as in Example 3.10. Suppose we try $n = 4$ replicates. This results in $\Phi^2 = 1.125(4) = 4.5$, $\Phi = 2.12$, and $4(3) = 12$ degrees of freedom for error. From the OC curve, we find that the power is approximately 0.65. For $n = 5$ replicates, we have $\Phi^2 = 5.625$, $\Phi = 2.37$, and $4(4) = 16$ degrees of freedom for error. From the OC curve, the power is approximately 0.8. For $n = 6$ replicates, we have $\Phi^2 = 6.75$, $\Phi = 2.60$, and $4(5) = 20$ degrees of freedom for error. From the OC curve, the power exceeds 0.90, so $n = 6$ replicates are required.

Minitab uses this approach to perform power calculations and find sample sizes for single-factor ANOVAs. Consider the following display:

```

Power and Sample Size

One-way ANOVA

Alpha = 0.01 Assumed standard deviation = 25
Number of Levels = 4

      SS Means      Sample      Power      Maximum
      2812.5        Size 5      0.804838      Difference
                        75

The sample size is for each level.

Power and Sample Size

One-way ANOVA

Alpha = 0.01 Assumed standard deviation = 25
Number of Levels 5 4

      SS Means      Sample      Target      Maximum
      2812.5        Size 6      Power      Actual Power      Difference
                        0.9      0.915384      75

The sample size is for each level.

```

In the upper portion of the display, we asked Minitab to calculate the power for $n = 5$ replicates when the maximum difference in treatment means is 75. Notice that the results closely match those obtained from the OC curves. The bottom portion of the display the output when the experimenter requests the sample size to obtain a target power of at least 0.90. Once again, the results agree with those obtained from the OC curve.

3.7.2 Specifying a Standard Deviation Increase

This approach is occasionally helpful in choosing the sample size. If the treatment means do not differ, the standard deviation of an observation chosen at random is σ . If the treatment means are different, however, the standard deviation of a randomly chosen observation is

$$\sqrt{\sigma^2 + \left(\sum_{i=1}^a \tau_i^2 / a\right)}$$

If we choose a percentage P for the increase in the standard deviation of an observation beyond which we wish to reject the hypothesis that all treatment means are equal, this is

equivalent to choosing

$$\frac{\sqrt{\sigma^2 + \left(\sum_{i=1}^a \tau_i^2/a\right)}}{\sigma} = 1 + 0.01P \quad (P = \text{percent})$$

or

$$\frac{\sqrt{\sum_{i=1}^a \tau_i^2/a}}{\sigma} = \sqrt{(1 + 0.01P)^2 - 1}$$

so that

$$\Phi = \frac{\sqrt{\sum_{i=1}^a \tau_i^2/a}}{\sigma/\sqrt{n}} = \sqrt{(1 + 0.01P)^2 - 1}(\sqrt{n}) \quad (3.46)$$

Thus, for a specified value of P , we may compute Φ from Equation 3.46 and then use the operating characteristic curves in Appendix Chart V to determine the required sample size.

For example, in the plasma etching experiment from Example 3.1, suppose that we wish to detect a standard deviation increase of 20 percent with a probability of at least 0.90 and $\alpha = 0.05$. Then

$$\Phi = \sqrt{(1.2)^2 - 1}(\sqrt{n}) = 0.66\sqrt{n}$$

Reference to the operating characteristic curves shows that $n = 10$ replicates would be required to give the desired sensitivity.

3.7.3 Confidence Interval Estimation Method

This approach assumes that the experimenter wishes to express the final results in terms of confidence intervals and is willing to specify in advance how wide he or she wants these confidence intervals to be. For example, suppose that in the plasma etching experiment from Example 3.1, we wanted a 95 percent confidence interval on the difference in mean etch rate for any two power settings to be ± 30 Å/min and a prior estimate of σ is 25. Then, using Equation 3.13, we find that the accuracy of the confidence interval is

$$\pm t_{\alpha/2, N-a} \sqrt{\frac{2MS_E}{n}}$$

Suppose that we try $n = 5$ replicates. Then, using $\sigma^2 = (25)^2 = 625$ as an estimate of MS_E , the accuracy of the confidence interval becomes

$$\pm 2.120 \sqrt{\frac{2(625)}{5}} = \pm 33.52$$

which does not meet the requirement. Trying $n = 6$ gives

$$\pm 2.086 \sqrt{\frac{2(625)}{6}} = \pm 30.11$$

Trying $n = 7$ gives

$$\pm 2.064 \sqrt{\frac{2(625)}{7}} = \pm 27.58$$

Clearly, $n = 7$ is the smallest sample size that will lead to the desired accuracy.

The quoted level of significance in the above illustration applies only to one confidence interval. However, the same general approach can be used if the experimenter wishes to prespecify a *set* of confidence intervals about which a **joint** or **simultaneous confidence statement** is made (see the comments about simultaneous confidence intervals in Section 3.3.3). Furthermore, the confidence intervals could be constructed about more general contrasts in the treatment means than the pairwise comparison illustrated above.

3.8 Other Examples of Single-Factor Experiments

3.8.1 Chocolate and Cardiovascular Health

An article in *Nature* describes an experiment to investigate the effect of consuming chocolate on cardiovascular health (“Plasma Antioxidants from Chocolate,” *Nature*, Vol. 424, 2003, pp. 1013). The experiment consisted of using three different types of chocolates: 100 g of dark chocolate, 100 g of dark chocolate with 200 mL of full-fat milk, and 200 g of milk chocolate. Twelve subjects were used, 7 women and 5 men, with an average age range of 32.2 ± 1 years, an average weight of 65.8 ± 3.1 kg, and body-mass index of 21.9 ± 0.4 kg m⁻². On different days a subject consumed one of the chocolate-factor levels and one hour later the total antioxidant capacity of their blood plasma was measured in an assay. Data similar to that summarized in the article are shown in Table 3.12.

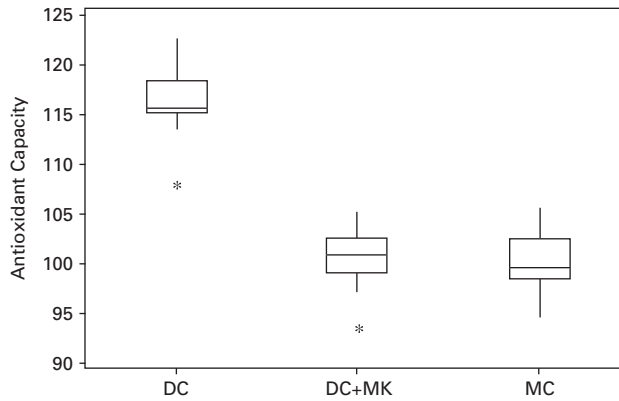
Figure 3.15 presents box plots for the data from this experiment. The result is an indication that the blood antioxidant capacity one hour after eating the dark chocolate is higher than for the other two treatments. The variability in the sample data from all three treatments seems very similar. Table 3.13 is the Minitab ANOVA output. The test statistic is highly significant (Minitab reports a *P*-value of 0.000, which is clearly wrong because *P*-values cannot be zero; this means that the *P*-value is less than 0.001), indicating that some of the treatment means are different. The output also contains the Fisher LSD analysis for this experiment. This indicates that the mean antioxidant capacity after consuming dark chocolate is higher than after consuming dark chocolate plus milk or milk chocolate alone, and the mean antioxidant capacity after consuming dark chocolate plus milk or milk chocolate alone are equal. Figure 3.16 is the normal probability plot of the residual and Figure 3.17 is the plot of residuals versus predicted values. These plots do not suggest any problems with model assumptions. We conclude that consuming dark chocolate results in higher mean blood antioxidant capacity after one hour than consuming either dark chocolate plus milk or milk chocolate alone.

3.8.2 A Real Economy Application of a Designed Experiment

Designed experiments have had tremendous impact on manufacturing industries, including the design of new products and the improvement of existing ones, development of new

■ **TABLE 3.12**
Blood Plasma Levels One Hour Following Chocolate Consumption

Factor	Subjects (Observations)											
	1	2	3	4	5	6	7	8	9	10	11	12
DC	118.8	122.6	115.6	113.6	119.5	115.9	115.8	115.1	116.9	115.4	115.6	107.9
DC+MK	105.4	101.1	102.7	97.1	101.9	98.9	100.0	99.8	102.6	100.9	104.5	93.5
MC	102.1	105.8	99.6	102.7	98.8	100.9	102.8	98.7	94.7	97.8	99.7	98.6



■ **FIGURE 3.15** Box plots of the blood antioxidant capacity data from the chocolate consumption experiment

■ **TABLE 3.13**

Minitab ANOVA Output, Chocolate Consumption Experiment

One-way ANOVA: DC, DC+MK, MC

Source	DF	SS	MS	F	P
Factor	2	1952.6	976.3	93.58	0.000
Error	33	344.3	10.4		
Total	35	2296.9			

S = 3.230 R-Sq = 85.01% R-Sq(adj) = 84.10%

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev	CI
DC	12	116.06	3.53	(---*---)
DC+MK	12	100.70	3.24	(---*---)
MC	12	100.18	2.89	(---*---)

Pooled StDev = 3.23

Fisher 95% Individual Confidence Intervals

All Pairwise Comparisons

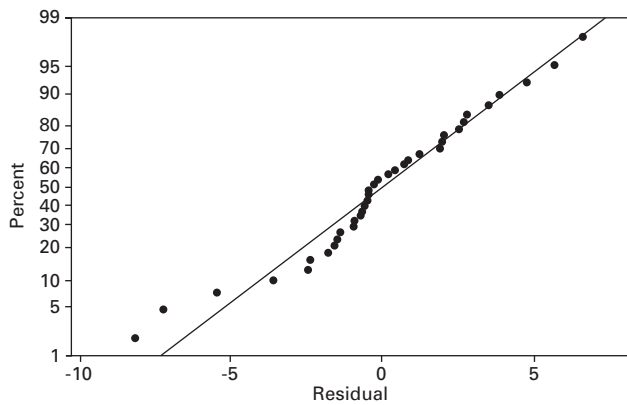
Simultaneous confidence level = 88.02

DC subtracted from:

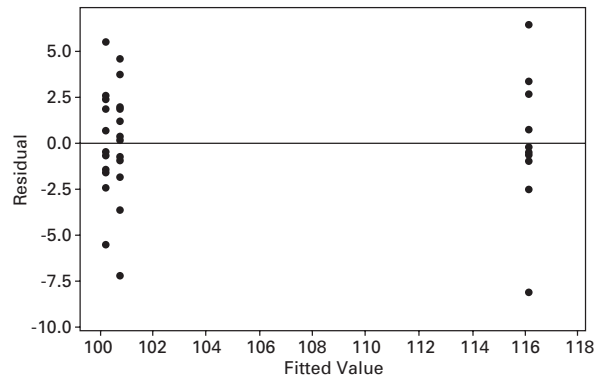
	Lower	Center	Upper	CI
DC+MK	-18.041	-15.358	-12.675	(---*---)
MC	-18.558	-15.875	-13.192	(---*---)

DC+MK subtracted from:

	Lower	Center	Upper	CI
MC	-3.200	-0.517	2.166	(---*---)



■ **FIGURE 3.16** Normal probability plot of the residuals from the chocolate consumption experiment



■ **FIGURE 3.17** Plot of residuals versus the predicted values from the chocolate consumption experiment

manufacturing processes, and process improvement. In the last 15 years, designed experiments have begun to be widely used outside of this traditional environment. These applications are in financial services, telecommunications, health care, e-commerce, legal services, marketing, logistics and transportation, and many of the nonmanufacturing components of manufacturing businesses. These types of businesses are sometimes referred to as the real economy. It has been estimated that manufacturing accounts for only about 20 percent of the total US economy, so applications of experimental design in the real economy are of growing importance. In this section, we present an example of a designed experiment in marketing.

A soft drink distributor knows that end-aisle displays are an effective way to increase sales of the product. However, there are several ways to design these displays: by varying the text displayed, the colors used, and the visual images. The marketing group has designed three new end-aisle displays and wants to test their effectiveness. They have identified 15 stores of similar size and type to participate in the study. Each store will test one of the displays for a period of one month. The displays are assigned at random to the stores, and each display is tested in five stores. The response variable is the percentage increase in sales activity over the typical sales for that store when the end-aisle display is not in use. The data from this experiment are shown in Table 3.13.

Table 3.14 shows the analysis of the end-aisle display experiment. This analysis was conducted using JMP. The *P*-value for the model *F* statistic in the ANOVA indicates that there is a difference in the mean percentage increase in sales between the three display types. In this application, we had JMP use the Fisher LSD procedure to compare the pairs of treatment means (JMP labels these as the least squares means). The results of this comparison are presented as confidence intervals on the difference in pairs of means. For pairs of means where the confidence interval includes zero, we would not declare that pair of means are different. The JMP output indicates that display designs 1 and 2 are similar in that they result in the same mean increase in sales, but that

■ **TABLE 3.13**
The End-Aisle Display Experimental Design

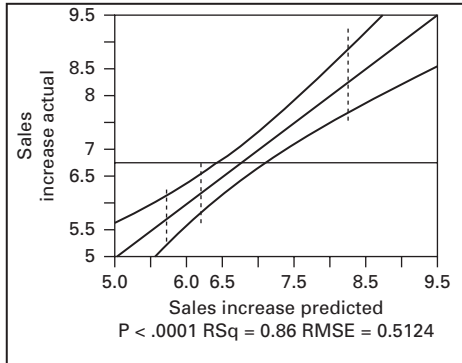
Display Design	Sample Observations, Percent Increase in Sales					
1	5.43	5.71	6.22	6.01	5.29	
2	6.24	6.71	5.98	5.66	6.60	
3	8.79	9.20	7.90	8.15	7.55	

display design 3 is different from both designs 1 and 2 and that the mean increase in sales for display 3 exceeds that of both designs 1 and 2. Notice that JMP automatically includes some useful graphics in the output, a plot of the actual observations versus the predicted values from the model, and a plot of the residuals versus the predicted values. There is some mild indication that display design 3 may exhibit more variability in sales increase than the other two designs.

■ **TABLE 3.14**
JMP Output for the End-Aisle Display Experiment

Response Sales Increase
 Whole Model

Actual by Predicted Plot



Summary of Fit

RSquare	0.856364
RSquare Adj	0.832425
Root Mean Square Error	0.512383
Mean of Response	6.762667
Observations (or Sum Wgts)	15

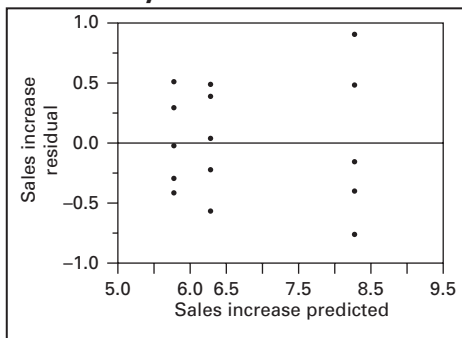
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	18.783053	9.39153	35.7722
Error	12	3.150440	0.26254	Prob>F
C.Total	14	21.933493		<.0001

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Display	2	2	18.783053	35.7722	<.001

Residual by Predicted Plot



■ TABLE 3.14 (Continued)

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
1	5.7320000	0.22914479	5.73200
2	6.2380000	0.22914479	6.23800
3	8.3180000	0.22914479	8.31800

LSMeans Differences Student's *t* $\alpha = 0.050$ $t = 2.17881$

LSMean[i] By LSMean [i]

Mean[i]-Mean [i] Std Err Dif Lower CL Dif Upper CL Dif	1	2	3
1	0	-0.506	-2.586
	0	0.32406	0.32406
	0	-1.2121	-3.2921
	0	0.20007	-1.8799
2	0.506	0	-2.08
	0.32406	0	0.32406
	-0.2001	0	-2.7861
	1.21207	0	-1.3739
3	2.586	2.08	0
	0.32406	0.32406	0
	1.87993	1.37393	0
	3.29207	2.78607	0

Level		Least Sq Mean
3	A	8.3180000
2	B	6.2380000
1	B	5.7320000

Levels not connected by same letter are significantly different.

3.8.3 Discovering Dispersion Effects

We have focused on using the analysis of variance and related methods to determine which factor levels result in differences among treatment or factor level means. It is customary to refer to these effects as **location effects**. If there was inequality of variance at the different factor levels, we used transformations to stabilize the variance to improve our inference on the location effects. In some problems, however, we are interested in discovering whether the different factor levels affect **variability**; that is, we are interested in discovering potential **dispersion effects**. This will occur whenever the standard deviation, variance, or some other measure of variability is used as a response variable.

To illustrate these ideas, consider the data in Table 3.15, which resulted from a designed experiment in an aluminum smelter. Aluminum is produced by combining alumina with other ingredients in a reaction cell and applying heat by passing electric current through the cell. Alumina is added continuously to the cell to maintain the proper ratio of alumina to other ingredients. Four different ratio control algorithms were investigated in this experiment. The response variables studied were related to cell voltage. Specifically, a sensor scans cell voltage several times each second, producing thousands of voltage measurements during each run of the experiment. The process engineers decided to use the average voltage and the standard deviation of

■ **TABLE 3.15**
Data for the Smelting Experiment

Ratio Control Algorithm	Observations					
	1	2	3	4	5	6
1	4.93(0.05)	4.86(0.04)	4.75(0.05)	4.95(0.06)	4.79(0.03)	4.88(0.05)
2	4.85(0.04)	4.91(0.02)	4.79(0.03)	4.85(0.05)	4.75(0.03)	4.85(0.02)
3	4.83(0.09)	4.88(0.13)	4.90(0.11)	4.75(0.15)	4.82(0.08)	4.90(0.12)
4	4.89(0.03)	4.77(0.04)	4.94(0.05)	4.86(0.05)	4.79(0.03)	4.76(0.02)

■ **TABLE 3.16**
Analysis of Variance for the Natural Logarithm of Pot Noise

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P -Value
Ratio control algorithm	6.166	3	2.055	21.96	<0.001
Error	1.872	20	0.094		
Total	8.038	23			

cell voltage (shown in parentheses) over the run as the response variables. The average voltage is important because it affects cell temperature, and the standard deviation of voltage (called “pot noise” by the process engineers) is important because it affects the overall cell efficiency.

An analysis of variance was performed to determine whether the different ratio control algorithms affect average cell voltage. This revealed that the ratio control algorithm had no **location effect**; that is, changing the ratio control algorithms does not change the average cell voltage. (Refer to Problem 3.38.)

To investigate dispersion effects, it is usually best to use

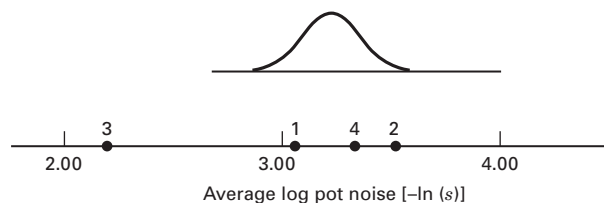
$$\log(s) \text{ or } \log(s^2)$$

as a response variable since the log transformation is effective in stabilizing variability in the distribution of the sample standard deviation. Because all sample standard deviations of pot voltage are less than unity, we will use

$$y = -\ln(s)$$

as the response variable. Table 3.16 presents the analysis of variance for this response, the natural logarithm of “pot noise.” Notice that the choice of a ratio control algorithm affects pot noise; that is, the ratio control algorithm has a **dispersion effect**. Standard tests of model adequacy, including normal probability plots of the residuals, indicate that there are no problems with experimental validity. (Refer to Problem 3.39.)

Figure 3.18 plots the average log pot noise for each ratio control algorithm and also presents a scaled t distribution for use as a **reference distribution** in discriminating between ratio control algorithms. This plot clearly reveals that ratio control algorithm 3 produces



■ **FIGURE 3.18** Average log pot noise $[-\ln(s)]$ for four ratio control algorithms relative to a scaled t distribution with scale factor $\sqrt{MS_E/n} = \sqrt{0.094/6} = 0.125$

greater pot noise or greater cell voltage standard deviation than the other algorithms. There does not seem to be much difference between algorithms 1, 2, and 4.

3.9 The Random Effects Model

3.9.1 A Single Random Factor

An experimenter is frequently interested in a factor that has a large number of possible levels. If the experimenter randomly selects a of these levels from the population of factor levels, then we say that the factor is **random**. Because the levels of the factor actually used in the experiment were chosen randomly, inferences are made about the entire population of factor levels. We assume that the population of factor levels is either of infinite size or is large enough to be considered infinite. Situations in which the population of factor levels is small enough to employ a finite population approach are not encountered frequently. Refer to Bennett and Franklin (1954) and Searle and Fawcett (1970) for a discussion of the finite population case.

The linear statistical model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases} \quad (3.47)$$

where both the treatment effects τ_i and ϵ_{ij} are random variables. We will assume that the treatment effects τ_i are NID $(0, \sigma_\tau^2)$ random variables¹ and that the errors are NID $(0, \sigma^2)$, random variables, and that the τ_i and ϵ_{ij} are independent. Because τ_i is independent of ϵ_{ij} , the variance of any observation is

$$V(y_{ij}) = \sigma_\tau^2 + \sigma^2$$

The variances σ_τ^2 and σ^2 are called **variance components**, and the model (Equation 3.47) is called the **components of variance** or **random effects model**. The observations in the random effects model are normally distributed because they are linear combinations of the two normally and independently distributed random variables τ_i and ϵ_{ij} . However, unlike the fixed effects case in which all of the observations y_{ij} are independent, in the random model the observations y_{ij} are only independent if they come from different factor levels. Specifically, we can show that the covariance of any two observations is

$$\begin{aligned} \text{Cov}(y_{ij}, y_{i'j'}) &= \sigma_\tau^2 & j \neq j' \\ \text{Cov}(y_{ij}, y_{i'j'}) &= 0 & i \neq i' \end{aligned}$$

Note that the observations within a specific factor level all have the same covariance, because before the experiment is conducted, we expect the observations at that factor level to be similar because they all have the same random component. Once the experiment has been conducted, we can assume that all observations can be assumed to be independent, because the parameter τ_i has been determined and the observations in that treatment differ only because of random error.

We can express the covariance structure of the observations in the single-factor random effects model through the **covariance matrix** of the observations. To illustrate, suppose that we have $a = 3$ treatments and $n = 2$ replicates. There are $N = 6$ observations, which we can write as a vector

¹ The assumption that the $[\tau_i]$ are independent random variables implies that the usual assumption of $\sum_{i=1}^a \tau_i = 0$ from the fixed effects model does not apply to the random effects model.

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix}$$

and the 6×6 covariance matrix of these observations is

$$\text{Cov}(\mathbf{y}) = \begin{bmatrix} \sigma_\tau^2 + \sigma^2 & \sigma_\tau^2 & 0 & 0 & 0 & 0 \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\tau^2 + \sigma^2 & \sigma_\tau^2 & 0 & 0 \\ 0 & 0 & \sigma_\tau^2 & \sigma_\tau^2 + \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\tau^2 + \sigma^2 & \sigma^2 \\ 0 & 0 & 0 & 0 & \sigma_\tau^2 & \sigma_\tau^2 + \sigma^2 \end{bmatrix}$$

The main diagonals of this matrix are the variances of each individual observation and every off-diagonal element is the covariance of a pair of observations.

3.9.2 Analysis of Variance for the Random Model

The basic ANOVA sum of squares identity

$$SS_T = SS_{\text{Treatments}} + SS_E \tag{3.48}$$

is still valid. That is, we partition the total variability in the observations into a component that measures the variation between treatments ($SS_{\text{Treatments}}$) and a component that measures the variation within treatments (SS_E). Testing hypotheses about individual treatment effects is not very meaningful because they were selected randomly, we are more interested in the **population** of treatments, so we test hypotheses about the variance component σ_τ^2 .

$$\begin{aligned} H_0: \sigma_\tau^2 &= 0 \\ H_1: \sigma_\tau^2 &> 0 \end{aligned} \tag{3.49}$$

If $\sigma_\tau^2 = 0$, all treatments are identical; but if $\sigma_\tau^2 > 0$, variability exists between treatments. As before, SS_E/σ^2 is distributed as chi-square with $N - a$ degrees of freedom and, under the null hypothesis, $SS_{\text{Treatments}}/\sigma^2$ is distributed as chi-square with $a - 1$ degrees of freedom. Both random variables are independent. Thus, under the null hypothesis $\sigma_\tau^2 = 0$, the ratio

$$F_0 = \frac{\frac{SS_{\text{Treatments}}}{a - 1}}{\frac{SS_E}{N - a}} = \frac{MS_{\text{Treatments}}}{MS_E} \tag{3.50}$$

is distributed as F with $a - 1$ and $N - a$ degrees of freedom. However, we need to examine the expected mean squares to fully describe the test procedure.

Consider

$$\begin{aligned} E(MS_{\text{Treatments}}) &= \frac{1}{a - 1} E(SS_{\text{Treatments}}) = \frac{1}{a - 1} E\left[\sum_{i=1}^a \frac{y_i^2}{n} - \frac{y_{..}^2}{N}\right] \\ &= \frac{1}{a - 1} E\left[\frac{1}{n} \sum_{i=1}^a \left(\sum_{j=1}^n \mu + \tau_i + \epsilon_{ij}\right)^2 - \frac{1}{N} \left(\sum_{i=1}^a \sum_{j=1}^n \mu + \tau_i + \epsilon_{ij}\right)^2\right] \end{aligned}$$

When squaring and taking expectation of the quantities in brackets, we see that terms involving τ_i^2 are replaced by σ_τ^2 as $E(\tau_i) = 0$. Also, terms involving $\epsilon_i^2, \epsilon_{.i}^2$, and $\sum_{i=1}^a \sum_{j=1}^n \tau_i^2$ are replaced by $n\sigma^2, an\sigma^2$, and an^2 , respectively. Furthermore, all cross-product terms involving τ_i and ϵ_{ij} have zero expectation. This leads to

$$E(MS_{\text{Treatments}}) = \frac{1}{a-1} [N\mu^2 + N\sigma_\tau^2 + a\sigma^2 - N\mu^2 - n\sigma_\tau^2 - \sigma^2]$$

or

$$E(MS_{\text{Treatments}}) = \sigma^2 + n\sigma_\tau^2 \tag{3.51}$$

Similarly, we may show that

$$E(MS_E) = \sigma^2 \tag{3.52}$$

From the expected mean squares, we see that under H_0 both the numerator and denominator of the test statistic (Equation 3.50) are unbiased estimators of σ^2 , whereas under H_1 the expected value of the numerator is greater than the expected value of the denominator. Therefore, we should reject H_0 for values of F_0 that are too large. This implies an upper-tail, one-tail critical region, so we reject H_0 if $F_0 > F_{\alpha, a-1, N-a}$.

The computational procedure and ANOVA for the random effects model are identical to those for the fixed effects case. The conclusions, however, are quite different because they apply to the entire population of treatments.

3.9.3 Estimating the Model Parameters

We are usually interested in estimating the variance components (σ^2 and σ_τ^2) in the model. One very simple procedure that we can use to estimate σ^2 and σ_τ^2 is called the **analysis of variance method** because it makes use of the lines in the analysis of variance table. The procedure consists of equating the expected mean squares to their observed values in the ANOVA table and solving for the variance components. In equating observed and expected mean squares in the single-factor random effects model, we obtain

$$MS_{\text{Treatments}} = \sigma^2 + n\sigma_\tau^2$$

and

$$MS_E = \sigma^2$$

Therefore, the estimators of the variance components are

$$\hat{\sigma}^2 = MS_E \tag{3.53}$$

and

$$\hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatments}} - MS_E}{n} \tag{3.54}$$

For unequal sample sizes, replace n in Equation 13.8 by

$$n_0 = \frac{1}{a-1} \left[\sum_{i=1}^a n_i - \frac{\sum_{i=1}^a n_i^2}{\sum_{i=1}^a n_i} \right] \tag{3.55}$$

The analysis of variance method of variance component estimation is a **method of moments procedure**. It does not require the normality assumption. It does yield estimators of σ^2 and σ_τ^2 that are best quadratic unbiased (i.e., of all unbiased quadratic functions of the observations, these estimators have minimum variance). There is a different method based on maximum likelihood that can be used to estimate the variance components that will be introduced later.

Occasionally, the analysis of variance method produces a negative estimate of a variance component. Clearly, variance components are by definition nonnegative, so a negative estimate of a variance component is viewed with some concern. One course of action is to accept the estimate and use it as evidence that the true value of the variance component is zero, assuming that sampling variation led to the negative estimate. This has intuitive appeal, but it suffers from some theoretical difficulties. For instance, using zero in place of the negative estimate can disturb the statistical properties of other estimates. Another alternative is to reestimate the negative variance component using a method that always yields nonnegative estimates. Still another alternative is to consider the negative estimate as evidence that the assumed linear model is incorrect and reexamine the problem. Comprehensive treatment of variance component estimation is given by Searle (1971a, 1971b), Searle, Casella, and McCulloch (1992), and Burdick and Graybill (1992).

EXAMPLE 3.11

A textile company weaves a fabric on a large number of looms. It would like the looms to be homogeneous so that it obtains a fabric of uniform strength. The process engineer suspects that, in addition to the usual variation in strength within samples of fabric from the same loom, there may also

be significant variations in strength between looms. To investigate this, she selects four looms at random and makes four strength determinations on the fabric manufactured on each loom. This experiment is run in random order, and the data obtained are shown in Table 3.17. The ANOVA is con-

■ **TABLE 3.17**
Strength Data for Example 3.11

Looms	Observations				y_i
	1	2	3	4	
1	98	97	99	96	390
2	91	90	93	92	366
3	96	95	97	95	383
4	95	96	99	98	388

$$1527 = y_{..}$$

ducted and is shown in Table 3.18. From the ANOVA, we conclude that the looms in the plant differ significantly.

The variance components are estimated by $\hat{\sigma}^2 = 1.90$ and

$$\hat{\sigma}_\tau^2 = \frac{29.73 - 1.90}{4} = 6.96$$

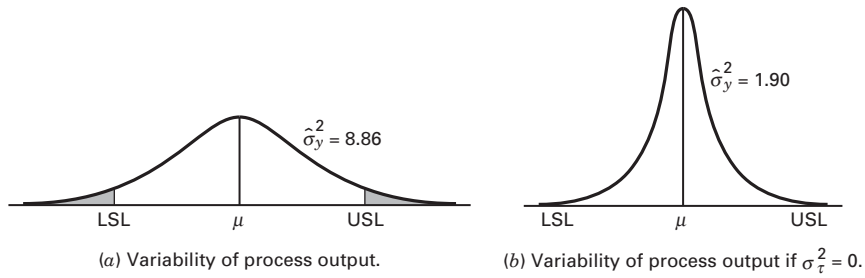
Therefore, the variance of any observation on strength is estimated by

$$\hat{\sigma}_y^2 = \hat{\sigma}^2 + \hat{\sigma}_\tau^2 = 1.90 + 6.96 = 8.86.$$

Most of this variability is attributable to differences *between* looms.

■ **TABLE 3.18**
Analysis of Variance for the Strength Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P -Value
Looms	89.19	3	29.73	15.68	<0.001
Error	22.75	12	1.90		
Total	111.94	15			



■ FIGURE 3.19 Process output in the fiber strength problem

This example illustrates an important use of variance components—isolating different sources of variability that affect a product or system. The problem of product variability frequently arises in quality assurance, and it is often difficult to isolate the sources of variability. For example, this study may have been motivated by an observation that there is too much variability in the strength of the fabric, as illustrated in Figure 3.19a. This graph displays the process output (fiber strength) modeled as a normal distribution with variance $\hat{\sigma}_y^2 = 8.86$. (This is the estimate of the variance of any observation on strength from Example 3.11.) Upper and lower specifications on strength are also shown in Figure 3.19a, and it is relatively easy to see that a fairly large proportion of the process output is outside the specifications (the shaded tail areas in Figure 3.19a). The process engineer has asked why so much fabric is defective and must be scrapped, reworked, or downgraded to a lower quality product. The answer is that most of the product strength variability is the result of differences between looms. Different loom performance could be the result of faulty setup, poor maintenance, ineffective supervision, poorly trained operators, defective input fiber, and so forth.

The process engineer must now try to isolate the specific causes of the differences in loom performance. If she could identify and eliminate these sources of between-loom variability, the variance of the process output could be reduced considerably, perhaps to as low as $\hat{\sigma}_y^2 = 1.90$, the estimate of the within-loom (error) variance component in Example 3.11. Figure 3.19b shows a normal distribution of fiber strength with $\hat{\sigma}_y^2 = 1.90$. Note that the proportion of defective product in the output has been dramatically reduced. Although it is unlikely that *all* of the between-loom variability can be eliminated, it is clear that a significant reduction in this variance component would greatly increase the quality of the fiber produced.

We may easily find a confidence interval for the variance component σ^2 . If the observations are normally and independently distributed, then $(N - a)MS_E/\sigma^2$ is distributed as χ^2_{N-a} . Thus,

$$P\left[\chi^2_{1-(\alpha/2), N-a} \leq \frac{(N - a)MS_E}{\sigma^2} \leq \chi^2_{\alpha/2, N-a}\right] = 1 - \alpha$$

and a $100(1 - \alpha)$ percent confidence interval for σ^2 is

$$\frac{(N - a)MS_E}{\chi^2_{\alpha/2, N-a}} \leq \sigma^2 \leq \frac{(N - a)MS_E}{\chi^2_{1-(\alpha/2), N-a}} \tag{3.56}$$

Since $MS_E = 190$, $N = 16$, $a = 4$, $\chi^2_{0.025, 12} = 23.3367$ and $\chi^2_{0.975, 12} = 4.4038$, the 95% CI on σ^2 is $0.9770 \leq \sigma^2 \leq 5.1775$.

Now consider the variance component σ_τ^2 . The point estimator of σ_τ^2 is

$$\hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatments}} - MS_E}{n}$$

The random variable $(a - 1)MS_{\text{Treatments}}/(\sigma^2 + n\sigma_\tau^2)$ is distributed as χ^2_{a-1} , and $(N - a)MS_E/\sigma^2$ is distributed as χ^2_{N-a} . Thus, the probability distribution of $\hat{\sigma}_\tau^2$ is a linear combination of two chi-square random variables, say

$$u_1 \chi_{a-1}^2 - u_2 \chi_{N-a}^2$$

where

$$u_1 = \frac{\sigma^2 + n\sigma_\tau^2}{n(a-1)} \quad \text{and} \quad u_2 = \frac{\sigma^2}{n(N-a)}$$

Unfortunately, a closed-form expression for the distribution of this linear combination of chi-square random variables cannot be obtained. Thus, an exact confidence interval for σ_τ^2 cannot be constructed. Approximate procedures are given in Graybill (1961) and Searle (1971a). Also see Section 13.6 of Chapter 13.

It is easy to find an exact expression for a confidence interval on the ratio $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$. This ratio is called the **intraclass correlation coefficient**, and it reflects the *proportion* of the variance of an observation [recall that $V(y_{ij}) = \sigma_\tau^2 + \sigma^2$] that is the result of differences between treatments. To develop this confidence interval for the case of a balanced design, note that $MS_{\text{Treatments}}$ and MS_E are independent random variables and, furthermore, it can be shown that

$$\frac{MS_{\text{Treatments}}/(n\sigma_\tau^2 + \sigma^2)}{MS_E/\sigma^2} \sim F_{a-1, N-a}$$

Thus,

$$\left(F_{1-\alpha/2, a-1, N-a} \leq \frac{MS_{\text{Treatments}}}{MS_E} \frac{\sigma^2}{n\sigma_\tau^2 + \sigma^2} \leq F_{\alpha/2, a-1, N-a} \right) = 1 - \alpha \quad (3.57)$$

By rearranging Equation 13.11, we may obtain the following:

$$P\left(L \leq \frac{\sigma_\tau^2}{\sigma^2} \leq U\right) = 1 - \alpha \quad (3.58)$$

where

$$L = \frac{1}{n} \left(\frac{MS_{\text{Treatments}}}{MS_E} \frac{1}{F_{\alpha/2, a-1, N-a}} - 1 \right) \quad (3.59a)$$

and

$$U = \frac{1}{n} \left(\frac{MS_{\text{Treatments}}}{MS_E} \frac{1}{F_{1-\alpha/2, a-1, N-a}} - 1 \right) \quad (3.59b)$$

Note that L and U are $100(1 - \alpha)$ percent lower and upper confidence limits, respectively, for the ratio σ_τ^2/σ^2 . Therefore, a $100(1 - \alpha)$ percent confidence interval for $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$ is

$$\frac{L}{1+L} \leq \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2} \leq \frac{U}{1+U} \quad (3.60)$$

To illustrate this procedure, we find a 95 percent confidence interval on $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$ for the strength data in Example 3.11. Recall that $MS_{\text{Treatments}} = 29.73$, $MS_E = 1.90$, $a = 4$, $n = 4$, $F_{0.025, 3, 12} = 4.47$, and $F_{0.975, 3, 12} = 1/F_{0.025, 12, 3} = 1/14.34 = 0.070$. Therefore, from Equation 3.59a and b,

$$L = \frac{1}{4} \left[\left(\frac{29.73}{1.90} \right) \left(\frac{1}{4.47} \right) - 1 \right] = 0.625$$

$$U = \frac{1}{4} \left[\left(\frac{29.73}{1.90} \right) \left(\frac{1}{0.070} \right) - 1 \right] = 55.633$$

and from Equation 3.60, the 95 percent confidence interval on $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$ is

$$\frac{0.625}{1.625} \leq \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2} \leq \frac{55.633}{56.633}$$

or

$$0.38 \leq \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2} \leq 0.98$$

We conclude that variability between looms accounts for between 38 and 98 percent of the variability in the observed strength of the fabric produced. This confidence interval is relatively wide because of the small number of looms used in the experiment. Clearly, however, the variability between looms (σ_τ^2) is not negligible.

Estimation of the Overall Mean μ . In many random effects experiments the experimenter is interested in estimating the overall mean μ . From the basic model assumptions it is easy to see that the expected value of any observation is just the overall mean. Consequently, an unbiased estimator of the overall mean is

$$\hat{\mu} = \bar{y}_{..}$$

So for Example 3.11 the estimate of the overall mean strength is

$$\hat{\mu} = \bar{y}_{..} = \frac{y_{..}}{N} = \frac{1527}{16} = 95.44$$

It is also possible to find a $100(1 - \alpha)\%$ confidence interval on the overall mean. The variance of \bar{y} is

$$V(\bar{y}_{..}) = V\left(\frac{\sum_{i=1}^l \sum_{j=1}^n y_{ij}}{an}\right) = \frac{n\sigma_\tau^2 + \sigma^2}{an}$$

The numerator of this ratio is estimated by the treatment mean square, so an unbiased estimator of $V(\bar{y})$ is

$$\hat{V}(\bar{y}_{..}) = \frac{MS_{\text{Treatments}}}{an}$$

Therefore, the $100(1 - \alpha)\%$ CI on the overall mean is

$$\bar{y}_{..} - t_{\alpha/2, a(n-1)} \sqrt{\frac{MS_{\text{Treatments}}}{an}} \leq \mu \leq \bar{y}_{..} + t_{\alpha/2, a(n-1)} \sqrt{\frac{MS_{\text{Treatments}}}{an}} \quad (3.61)$$

To find a 95% CI on the overall mean in the fabric strength experiment from Example 3.11, we need $MS_{\text{Treatments}} = 29.73$ and $t_{0.025, 12} = 2.18$. The CI is computed from Equation 3.61 as follows:

$$\begin{aligned} \bar{y}_{..} - t_{\alpha/2, a(n-1)} \sqrt{\frac{MS_{\text{Treatments}}}{an}} &\leq \mu \leq \bar{y}_{..} + t_{\alpha/2, a(n-1)} \sqrt{\frac{MS_{\text{Treatments}}}{an}} \\ 95.44 - 2.18 \sqrt{\frac{29.73}{20}} &\leq \mu \leq 95.44 + 2.18 \sqrt{\frac{29.73}{20}} \\ 92.78 &\leq \mu \leq 98.10 \end{aligned}$$

So, at 95 percent confidence the mean strength of the fabric produced by the looms in this facility is between 92.78 and 98.10. This is a relatively wide confidence interval because a small number of looms were sampled and there is a large difference between looms as reflected by the large portion of total variability that is accounted for by the differences between looms.

Maximum Likelihood Estimation of the Variance Components. Earlier in this section we presented the analysis of variance method of variance component estimation. This method is relatively straightforward to apply and makes use of familiar quantities—the mean squares in the analysis of variance table. However, the method has some disadvantages. As we pointed out previously, it is a **method of moments estimator**, a technique that mathematical statisticians generally do not prefer to use for parameter estimation because it often results in parameter estimates that do not have good statistical properties. One obvious problem is that it does not always lead to an easy way to construct confidence intervals on the variance components of interest. For example, in the single-factor random model there is not a simple way to construct confidence intervals on σ_τ^2 , which is certainly a parameter of primary interest to the experimenter. The preferred parameter estimation technique is called the **method of maximum likelihood**. The implementation of this method can be somewhat involved, particularly for an experimental design model, but it has been incorporated in some modern computer software packages that support designed experiments, including JMP.

A complete presentation of the method of maximum likelihood is beyond the scope of this book, but the general idea can be illustrated very easily. Suppose that x is a random variable with probability distribution $f(x, \theta)$, where θ is an unknown parameter. Let x_1, x_2, \dots, x_n be a random sample of n observations. The joint probability distribution of the sample is $\prod_{i=1}^n f(x_i, \theta)$. The **likelihood function** is just this joint probability distribution with the sample observations considered fixed and the parameter θ unknown. Note that the likelihood function, say

$$L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i, \theta)$$

is now a function of only the unknown parameter θ . The **maximum likelihood estimator** of θ is the value of θ that maximizes the likelihood function $L(x_1, x_2, \dots, x_n; \theta)$. To illustrate how this applies to an experimental design model with random effects, let \mathbf{y} be the $an \times 1$ vector of observations for a single-factor random effects model with a treatments and n replicates and let Σ be the $an \times an$ covariance matrix of the observations. Refer to Section 3.9.1 where we developed this covariance matrix for the special case where $a = 3$ and $n = 2$. The likelihood function is

$$L(x_{11}, x_{12}, \dots, x_{a,n}, \mu, \sigma_\tau^2, \sigma^2) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{y} - \mathbf{j}_N \mu)' \Sigma^{-1} (\mathbf{y} - \mathbf{j}_N \mu) \right]$$

where $N = an$ is the total number of observations, \mathbf{j}_N is an $N \times 1$ vector of 1s, and μ is the overall mean in the model. The maximum likelihood estimates of the parameters μ , σ_τ^2 , and σ^2 are the values of these quantities that maximize the likelihood function.

Maximum likelihood estimators (MLEs) have some very useful properties. For large samples, they are unbiased, and they have a normal distribution. Furthermore, the inverse of the matrix of second derivatives of the likelihood function (multiplied by -1) is the covariance matrix of the MLEs. This makes it relatively easy to obtain approximate confidence intervals on the MLEs.

The standard variant of maximum likelihood estimation that is used for estimating variance components is known as the **residual maximum likelihood (REML) method**. It is popular because it produces unbiased estimators and like all MLEs, it is easy to find CIs. The basic

characteristic of REML is that it takes the location parameters in the model into account when estimating the random effects. As a simple example, suppose that we want to estimate the mean and variance of a normal distribution using the method of maximum likelihood. It is easy to show that the MLEs are

$$\hat{\mu} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

Notice that the MLE $\hat{\sigma}^2$ is not the familiar sample standard deviation. It does not take the estimation of the location parameter μ into account. The REML estimator would be

$$S^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$$

The REML estimator is unbiased.

To illustrate the REML method, Table 3.19 presents the JMP output for the loom experiment in Example 3.11. The REML estimates of the model parameters μ , σ_r^2 , and σ^2 are shown in the output. Note that the REML estimates of the variance components are identical to those found earlier by the ANOVA method. These two methods will agree for balanced designs. However, the REML output also contains the covariance matrix of the variance components. The square roots of the main diagonal elements of this matrix are the standard

■ TABLE 3.19
JMP Output for the Loom Experiment in Example 3.11

Response Y						
Summary of Fit						
RSquare		0.793521				
RSquare Adj		0.793521				
Root Mean Square Error		1.376893				
Mean of Response		95.4375				
Observations (or Sum Wgts)		16				
Parameter Estimates						
Term	Estimate	Std Error	DFDen	t Ratio	Prob> t	
Intercept	95.4375	1.363111	3	70.01	<.0001*	
REML Variance Component Estimates						
Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Pct of Total
X1	3.6703297	6.9583333	6.0715247	-4.941636	18.858303	78.588
Residual		1.8958333	0.7739707	0.9748608	5.1660065	21.412
Total		8.8541667				100.000
Covariance Matrix of Variance Component Estimates						
Random Effect	X1	Residual				
X1	36.863412	-0.149758				
Residual	-0.149758	0.5990307				

errors of the variance components. If $\hat{\theta}$ is the MLE of θ and $\hat{\sigma}(\hat{\theta})$ is its estimated standard error, then the approximate $100(1 - \alpha)\%$ CI on θ is

$$\hat{\theta} - Z_{\alpha/2}\hat{\sigma}(\hat{\theta}) \leq \theta \leq \hat{\theta} + Z_{\alpha/2}\hat{\sigma}(\hat{\theta})$$

JMP uses this approach to find the approximate CIs σ_τ^2 and σ^2 shown in the output. The 95 percent CI from REML for σ^2 is very similar to the chi-square based interval computed earlier in Section 3.9.

3.10 The Regression Approach to the Analysis of Variance

We have given an intuitive or heuristic development of the analysis of variance. However, it is possible to give a more formal development. The method will be useful later in understanding the basis for the statistical analysis of more complex designs. Called the **general regression significance test**, the procedure essentially consists of finding the reduction in the total sum of squares for fitting the model with all parameters included and the reduction in sum of squares when the model is restricted to the null hypotheses. The difference between these two sums of squares is the treatment sum of squares with which a test of the null hypothesis can be conducted. The procedure requires the least squares estimators of the parameters in the analysis of variance model. We have given these parameter estimates previously (in Section 3.3.3); however, we now give a formal development.

3.10.1 Least Squares Estimation of the Model Parameters

We now develop estimators for the parameter in the single-factor ANOVA fixed-effects model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

using the method of least squares. To find the least squares estimators of μ and τ_i , we first form the sum of squares of the errors

$$L = \sum_{i=1}^a \sum_{j=1}^n \epsilon_{ij}^2 = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \mu - \tau_i)^2 \tag{3.61}$$

and then choose values of μ and τ_i , say $\hat{\mu}$ and $\hat{\tau}_i$, that minimize L . The appropriate values would be the solutions to the $a + 1$ simultaneous equations

$$\begin{aligned} \left. \frac{\partial L}{\partial \mu} \right|_{\hat{\mu}, \hat{\tau}_i} &= 0 \\ \left. \frac{\partial L}{\partial \tau_i} \right|_{\hat{\mu}, \hat{\tau}_i} &= 0 \quad i = 1, 2, \dots, a \end{aligned}$$

Differentiating Equation 3.61 with respect to μ and τ_i and equating to zero, we obtain

$$-2 \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0$$

and

$$-2 \sum_{j=1}^n (y_{ij} + \hat{\mu} - \hat{\tau}_i) = 0 \quad i = 1, 2, \dots, a$$

which, after simplification, yield

$$\begin{aligned} N\hat{\mu} + n\hat{\tau}_1 + n\hat{\tau}_2 + \dots + n\hat{\tau}_a &= y_{..} \\ n\hat{\mu} + n\hat{\tau}_1 &= y_{i.} \end{aligned}$$

$$\begin{array}{rcl}
 n\hat{\mu} & + & n\hat{\tau}_2 & = & y_2. \\
 & & \vdots & & \vdots \\
 n\hat{\mu} & & & + & n\hat{\tau}_a & = & y_a.
 \end{array} \tag{3.62}$$

The $a + 1$ equations (Equation 3.62) in $a + 1$ unknowns are called the **least squares normal equations**. Notice that if we add the last a normal equations, we obtain the first normal equation. Therefore, the normal equations are not linearly independent, and no unique solution for $\mu, \tau_1, \dots, \tau_a$ exists. This has happened because the effects model is **overparameterized**. This difficulty can be overcome by several methods. Because we have defined the treatment effects as deviations from the overall mean, it seems reasonable to apply the **constraint**

$$\sum_{i=1}^a \hat{\tau}_i = 0 \tag{3.63}$$

Using this constraint, we obtain as the solution to the normal equations

$$\begin{array}{l}
 \hat{\mu} = \bar{y}_{..} \\
 \hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..} \quad i = 1, 2, \dots, a
 \end{array} \tag{3.64}$$

This solution is obviously not unique and depends on the constraint (Equation 3.63) that we have chosen. At first this may seem unfortunate because two different experimenters could analyze the same data and obtain different results if they apply different constraints. However, certain **functions** of the model parameters *are* uniquely estimated, regardless of the constraint. Some examples are $\tau_i - \tau_j$, which would be estimated by $\hat{\tau}_i - \hat{\tau}_j = \bar{y}_{i.} - \bar{y}_{j.}$, and the i th treatment mean $\mu_i = \mu + \tau_i$, which would be estimated by $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$

Because we are usually interested in differences among the treatment effects rather than their actual values, it causes no concern that the τ_i cannot be uniquely estimated. In general, any function of the model parameters that is a linear combination of the left-hand side of the normal equations (Equations 3.48) can be uniquely estimated. Functions that are uniquely estimated regardless of which constraint is used are called **estimable functions**. For more information, see the **supplemental material** for this chapter. We are now ready to use these parameter estimates in a general development of the analysis of variance.

3.10.2 The General Regression Significance Test

A fundamental part of this procedure is writing the normal equations for the model. These equations may always be obtained by forming the least squares function and differentiating it with respect to each unknown parameter, as we did in Section 3.9.1. However, an easier method is available. The following rules allow the normal equations for *any* experimental design model to be written directly:

RULE 1. There is one normal equation for each parameter in the model to be estimated.

RULE 2. The right-hand side of any normal equation is just the sum of all observations that contain the parameter associated with that particular normal equation.

To illustrate this rule, consider the single-factor model. The first normal equation is for the parameter μ ; therefore, the right-hand side is $y_{..}$ because *all* observations contain μ .

RULE 3. The left-hand side of any normal equation is the sum of all model parameters, where each parameter is multiplied by the number of times it appears in the total on the right-hand side. The parameters are written with a circumflex ($\hat{}$) to indicate that they are **estimators** and not the true parameter values.

For example, consider the first normal equation in a single-factor experiment. According to the above rules, it would be

$$N\hat{\mu} + n\hat{\tau}_1 + n\hat{\tau}_2 + \dots + n\hat{\tau}_a = y_{..}$$

because μ appears in all N observations, τ_1 appears only in the n observations taken under the first treatment, τ_2 appears only in the n observations taken under the second treatment, and so on. From Equation 3.62, we verify that the equation shown above is correct. The second normal equation would correspond to τ_1 and is

$$n\hat{\mu} + n\hat{\tau}_1 = y_{1.}$$

because only the observations in the first treatment contain τ_1 (this gives $y_{1.}$ as the right-hand side), μ and τ_1 appear exactly n times in $y_{1.}$, and all other τ_i appear zero times. In general, the left-hand side of any normal equation is the expected value of the right-hand side.

Now, consider finding the reduction in the sum of squares by fitting a particular model to the data. By fitting a model to the data, we “explain” some of the variability; that is, we reduce the unexplained variability by some amount. The reduction in the unexplained variability is always the sum of the parameter estimates, each multiplied by the right-hand side of the normal equation that corresponds to that parameter. For example, in a single-factor experiment, the reduction due to fitting the **full model** $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ is

$$\begin{aligned} R(\mu, \tau) &= \hat{\mu}y_{..} + \hat{\tau}_1y_{1.} + \hat{\tau}_2y_{2.} + \cdots + \hat{\tau}_ay_{a.} \\ &= \hat{\mu}y_{..} + \sum_{i=1}^a \hat{\tau}_iy_{i.} \end{aligned} \tag{3.65}$$

The notation $R(\mu, \tau)$ means that reduction in the sum of squares from fitting the model containing μ and $\{\tau_i\}$. $R(\mu, \tau)$ is also sometimes called the “regression” sum of squares for the full model $y_{ij} = \mu + \tau_i + \epsilon_{ij}$. The number of degrees of freedom associated with a reduction in the sum of squares, such as $R(\mu, \tau)$, is always equal to the number of linearly independent normal equations. The remaining variability unaccounted for by the model is found from

$$SS_E = \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - R(\mu, \tau) \tag{3.66}$$

This quantity is used in the denominator of the test statistic for $H_0: \tau_1 = \tau_2 = \dots = \tau_a = 0$.

We now illustrate the general regression significance test for a single-factor experiment and show that it yields the usual one-way analysis of variance. The model is $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, and the normal equations are found from the above rules as

$$\begin{array}{rcl} N\hat{\mu} + n\hat{\tau}_1 + n\hat{\tau}_2 + \cdots + n\hat{\tau}_a & = & y_{..} \\ n\hat{\mu} + n\hat{\tau}_1 & = & y_{1.} \\ n\hat{\mu} & + & n\hat{\tau}_2 & = & y_{2.} \\ & & \vdots & & \vdots \\ n\hat{\mu} & & & + & n\hat{\tau}_a & = & y_{a.} \end{array}$$

Compare these normal equations with those obtained in Equation 3.62.

Applying the constraint $\sum_{i=1}^a \hat{\tau}_i = 0$, we find that the estimators for μ and τ_i are

$$\hat{\mu} = \bar{y}_{..} \quad \hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..} \quad i = 1, 2, \dots, a$$

The reduction in the sum of squares due to fitting this full model is found from Equation 3.51 as

$$\begin{aligned} R(\mu, \tau) &= \hat{\mu}y_{..} + \sum_{i=1}^a \hat{\tau}_iy_{i.} \\ &= (\bar{y}_{..})y_{..} + \sum_{i=1}^a (\bar{y}_{i.} - \bar{y}_{..})y_{i.} \\ &= \frac{y_{..}^2}{N} + \sum_{i=1}^a \bar{y}_{i.}y_{i.} - \bar{y}_{..} \sum_{i=1}^a y_{i.} \\ &= \sum_{i=1}^a \frac{y_{i.}^2}{n} \end{aligned}$$

which has a degrees of freedom because there are a linearly independent normal equations. The error sum of squares is, from Equation 3.66,

$$\begin{aligned} SS_E &= \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - R(\mu, \tau) \\ &= \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - \sum_{i=1}^a \frac{y_{i.}^2}{n} \end{aligned}$$

and has $N - a$ degrees of freedom.

To find the sum of squares resulting from the treatment effects (the $\{\tau_i\}$), we consider a **reduced model**; that is, the model to be restricted to the null hypothesis ($\tau_i = 0$ for all i). The reduced model is $y_{ij} = \mu + \epsilon_{ij}$. There is only one normal equation for this model:

$$N\hat{\mu} = y_{..}$$

and the estimator of μ is $\hat{\mu} = \bar{y}_{..}$. Thus, the reduction in the sum of squares that results from fitting the reduced model containing only μ is

$$R(\mu) = (\bar{y}_{..})(y_{..}) = \frac{y_{..}^2}{N}$$

Because there is only one normal equation for this reduced model, $R(\mu)$ has one degree of freedom. The sum of squares due to the $\{\tau_i\}$, given that μ is already in the model, is the difference between $R(\mu, \tau)$ and $R(\mu)$, which is

$$\begin{aligned} R(\tau|\mu) &= R(\mu, \tau) - R(\mu) \\ &= R(\text{Full Model}) - R(\text{Reduced Model}) \\ &= \frac{1}{n} \sum_{i=1}^a y_{i.}^2 - \frac{y_{..}^2}{N} \end{aligned}$$

with $a - 1$ degrees of freedom, which we recognize from Equation 3.9 as $SS_{\text{Treatments}}$. Making the usual normality assumption, we obtain appropriate statistic for testing $H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$

$$F_0 = \frac{R(\tau|\mu)/(a - 1)}{\left[\sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - R(\mu, \tau) \right] / (N - a)}$$

which is distributed as $F_{a-1, N-a}$ under the null hypothesis. This is, of course, the test statistic for the single-factor analysis of variance.

3.11 Nonparametric Methods in the Analysis of Variance

3.11.1 The Kruskal–Wallis Test

In situations where the normality assumption is unjustified, the experimenter may wish to use an alternative procedure to the F test analysis of variance that does not depend on this assumption. Such a procedure has been developed by Kruskal and Wallis (1952). The Kruskal–Wallis test is used to test the null hypothesis that the a treatments are identical against the alternative hypothesis that some of the treatments generate observations that are larger than others. Because the procedure is designed to be sensitive for testing differences in means, it is sometimes convenient to think of the Kruskal–Wallis test as a test for equality of treatment means. The Kruskal–Wallis test is a **nonparametric alternative** to the usual analysis of variance.

To perform a Kruskal–Wallis test, first rank the observations y_{ij} in ascending order and replace each observation by its rank, say R_{ij} , with the smallest observation having rank 1. In

the case of ties (observations having the same value), assign the average rank to each of the tied observations. Let R_i be the sum of the ranks in the i th treatment. The test statistic is

$$H = \frac{1}{S^2} \left[\sum_{i=1}^a \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4} \right] \tag{3.67}$$

where n_i is the number of observations in the i th treatment, N is the total number of observations, and

$$S^2 = \frac{1}{N-1} \left[\sum_{i=1}^a \sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4} \right] \tag{3.68}$$

Note that S^2 is just the variance of the ranks. If there are no ties, $S^2 = N(N+1)/12$ and the test statistic simplifies to

$$H = \frac{12}{N(N+1)} \sum_{i=1}^a \frac{R_i^2}{n_i} - 3(N+1) \tag{3.69}$$

When the number of ties is moderate, there will be little difference between Equations 3.68 and 3.69, and the simpler form (Equation 3.69) may be used. If the n_i are reasonably large, say $n_i \geq 5$, H is distributed approximately as χ^2_{a-1} under the null hypothesis. Therefore, if

$$H > \chi^2_{\alpha, a-1}$$

the null hypothesis is rejected. The P -value approach could also be used.

EXAMPLE 3.12

The data from Example 3.1 and their corresponding ranks are shown in Table 3.20. There are ties, so we use Equation 3.67 as the test statistic. From Equation 3.67

$$S^2 = \frac{1}{19} \left[2869.50 - \frac{20(21)^2}{4} \right] = 34.97$$

$$\begin{aligned} H &= \frac{1}{S^2} \left[\sum_{i=1}^a \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4} \right] \\ &= \frac{1}{34.97} [2796.30 - 2205] \\ &= 16.91 \end{aligned}$$

TABLE 3.20
Data and Ranks for the Plasma Etching Experiment in Example 3.1

Power							
160		180		200		220	
y_{1j}	R_{1j}	y_{2j}	R_{2j}	y_{3j}	R_{3j}	y_{4j}	R_{4j}
575	6	565	4	600	10	725	20
542	3	593	9	651	15	700	17
530	1	590	8	610	11.5	715	19
539	2	579	7	637	14	685	16
570	5	610	11.5	629	13	710	18
R_i	17		39.5		63.5		90

Because $H > \chi^2_{0.01,3} = 11.34$, we would reject the null hypothesis and conclude that the treatments differ. (The P -

value for $H = 16.91$ is $P = 7.38 \times 10^{-4}$.) This is the same conclusion as given by the usual analysis of variance F test.

3.11.2 General Comments on the Rank Transformation

The procedure used in the previous section of replacing the observations by their ranks is called the **rank transformation**. It is a very powerful and widely useful technique. If we were to apply the ordinary F test to the ranks rather than to the original data, we would obtain

$$F_0 = \frac{H(a-1)}{(N-1-H)/(N-a)} \quad (3.70)$$

as the test statistic [see Conover (1980), p. 337]. Note that as the Kruskal–Wallis statistic H increases or decreases, F_0 also increases or decreases, so the Kruskal–Wallis test is equivalent to applying the usual analysis of variance to the ranks.

The rank transformation has wide applicability in experimental design problems for which no nonparametric alternative to the analysis of variance exists. This includes many of the designs in subsequent chapters of this book. If the data are ranked and the ordinary F test is applied, an approximate procedure that has good statistical properties results [see Conover and Iman (1976, 1981)]. When we are concerned about the normality assumption or the effect of outliers or “wild” values, we recommend that the usual analysis of variance be performed on both the original data and the ranks. When both procedures give similar results, the analysis of variance assumptions are probably satisfied reasonably well, and the standard analysis is satisfactory. When the two procedures differ, the rank transformation should be preferred because it is less likely to be distorted by nonnormality and unusual observations. In such cases, the experimenter may want to investigate the use of transformations for nonnormality and examine the data and the experimental procedure to determine whether outliers are present and why they have occurred.

3.12 Problems

3.1. An experimenter has conducted a single-factor experiment with four levels of the factor, and each factor level has been replicated six times. The computed value of the F -statistic is $F_0 = 3.26$. Find bounds on the P -value.

3.2. An experimenter has conducted a single-factor experiment with six levels of the factor, and each factor level has been replicated three times. The computed value of the F -statistic is $F_0 = 5.81$. Find bounds on the P -value.

3.3. A computer ANOVA output is shown below. Fill in the blanks. You may give bounds on the P -value.

One-way ANOVA					
Source	DF	SS	MS	F	P
Factor	3	36.15	?	?	?
Error	?	?	?		
Total	19	196.04			

3.4. A computer ANOVA output is shown below. Fill in the blanks. You may give bounds on the P -value.

One-way ANOVA					
Source	DF	SS	MS	F	P
Factor	?	?	246.93	?	?
Error	25	186.53	?		
Total	29	1174.24			

3.5. An article appeared in *The Wall Street Journal* on Tuesday, April 27, 2010, with the title “Eating Chocolate Is Linked to Depression.” The article reported on a study funded by the National Heart, Lung and Blood Institute (part of the National Institutes of Health) and conducted by faculty at the University of California, San Diego, and the University of California, Davis. The research was also published in the *Archives of Internal Medicine* (2010, pp. 699–703). The study examined 931 adults who were not taking antidepressants and did not have known cardiovascular disease or diabetes. The group was about 70% men and the average age of the group was reported to be about 58. The participants were asked about chocolate consumption and then screened for depression using a questionnaire. People who score less than 16 on the questionnaire are not considered depressed, while those

with scores above 16 and less than or equal to 22 are considered possibly depressed, while those with scores above 22 are considered likely to be depressed. The survey found that people who were not depressed ate an average 5.4 servings of chocolate per month, possibly depressed individuals ate an average of 8.4 servings of chocolate per month, while those individuals who scored above 22 and were likely to be depressed ate the most chocolate, an average of 11.8 servings per month. No differentiation was made between dark and milk chocolate. Other foods were also examined, but no pattern emerged between other foods and depression. Is this study really a designed experiment? Does it establish a cause-and-effect link between chocolate consumption and depression? How would the study have to be conducted to establish such a cause-and effect link?

3.6. An article in *Bioelectromagnetics* (“Electromagnetic Effects on Forearm Disuse Osteopenia: A Randomized, Double-Blind, Sham-Controlled Study,” Vol. 32, 2011, pp. 273–282) described a randomized, double-blind, sham-controlled, feasibility and dosing study to determine if a common pulsing electromagnetic field (PEMF) treatment could moderate the substantial osteopenia that occurs after forearm disuse. Subjects were randomized into four groups after a distal radius fracture, or carpal surgery requiring immobilization in a cast. Active or identical sham PEMF transducers were worn on the distal forearm for 1, 2, or 4h/day for 8 weeks starting after cast removal (“baseline”) when bone density continues to decline. Bone mineral density (BMD) and bone geometry were measured in the distal forearm by dual energy X-ray absorptiometry (DXA) and peripheral quantitative computed tomography (pQCT). The data below are the percent losses in BMD measurements on the radius after 16 weeks for patients wearing the active or sham PEMF transducers for 1, 2, or 4h/day (data were constructed to match the means and standard deviations read from a graph in the paper).

- (a) Is there evidence to support a claim that PEMF usage affects BMD loss? If so, analyze the data to determine which specific treatments produce the differences.
- (b) Analyze the residuals from this experiment and comment on the underlying assumptions and model adequacy.

Sham	PEMF 1 h/day	PEMF 2 h/day	PEMF 4 h/day
4.51	5.32	4.73	7.03
7.95	6.00	5.81	4.65
4.97	5.12	5.69	6.65
3.00	7.08	3.86	5.49
7.97	5.48	4.06	6.98
2.23	6.52	6.56	4.85
3.95	4.09	8.34	7.26
5.64	6.28	3.01	5.92

9.35	7.77	6.71	5.58
6.52	5.68	6.51	7.91
4.96	8.47	1.70	4.90
6.10	4.58	5.89	4.54
7.19	4.11	6.55	8.18
4.03	5.72	5.34	5.42
2.72	5.91	5.88	6.03
9.19	6.89	7.50	7.04
5.17	6.99	3.28	5.17
5.70	4.98	5.38	7.60
5.85	9.94	7.30	7.90
6.45	6.38	5.46	7.91

3.7. The tensile strength of Portland cement is being studied. Four different mixing techniques can be used economically. A completely randomized experiment was conducted and the following data were collected:

Mixing Technique	Tensile Strength (lb/in ²)			
1	3129	3000	2865	2890
2	3200	3300	2975	3150
3	2800	2900	2985	3050
4	2600	2700	2600	2765

- (a) Test the hypothesis that mixing techniques affect the strength of the cement. Use $\alpha = 0.05$.
 - (b) Construct a graphical display as described in Section 3.5.3 to compare the mean tensile strengths for the four mixing techniques. What are your conclusions?
 - (c) Use the Fisher LSD method with $\alpha = 0.05$ to make comparisons between pairs of means.
 - (d) Construct a normal probability plot of the residuals. What conclusion would you draw about the validity of the normality assumption?
 - (e) Plot the residuals versus the predicted tensile strength. Comment on the plot.
 - (f) Prepare a scatter plot of the results to aid the interpretation of the results of this experiment.
- 3.8(a)** Rework part (c) of Problem 3.7 using Tukey’s test with $\alpha = 0.05$. Do you get the same conclusions from Tukey’s test that you did from the graphical procedure and/or the Fisher LSD method?
- (b) Explain the difference between the Tukey and Fisher procedures.
- 3.9.** Reconsider the experiment in Problem 3.7. Find a 95 percent confidence interval on the mean tensile strength of the Portland cement produced by each of the four mixing techniques. Also find a 95 percent confidence interval on the difference in means for techniques 1 and 3. Does this aid you in interpreting the results of the experiment?

3.10. A product developer is investigating the tensile strength of a new synthetic fiber that will be used to make cloth for men's shirts. Strength is usually affected by the percentage of cotton used in the blend of materials for the fiber. The engineer conducts a completely randomized experiment with five levels of cotton content and replicates the experiment five times. The data are shown in the following table.

Cotton Weight Percent	Observations				
15	7	7	15	11	9
20	12	17	12	18	18
25	14	19	19	18	18
30	19	25	22	19	23
35	7	10	11	15	11

- (a) Is there evidence to support the claim that cotton content affects the mean tensile strength? Use $\alpha = 0.05$.
- (b) Use the Fisher LSD method to make comparisons between the pairs of means. What conclusions can you draw?
- (c) Analyze the residuals from this experiment and comment on model adequacy.

3.11. Reconsider the experiment described in Problem 3.10. Suppose that 30 percent cotton content is a control. Use Dunnett's test with $\alpha = 0.05$ to compare all of the other means with the control.

3.12. A pharmaceutical manufacturer wants to investigate the bioactivity of a new drug. A completely randomized single-factor experiment was conducted with three dosage levels, and the following results were obtained.

Dosage	Observations			
20 g	24	28	37	30
30 g	37	44	31	35
40 g	42	47	52	38

- (a) Is there evidence to indicate that dosage level affects bioactivity? Use $\alpha = 0.05$.
- (b) If it is appropriate to do so, make comparisons between the pairs of means. What conclusions can you draw?
- (c) Analyze the residuals from this experiment and comment on model adequacy.

3.13. A rental car company wants to investigate whether the type of car rented affects the length of the rental period. An experiment is run for one week at a particular location, and

10 rental contracts are selected at random for each car type. The results are shown in the following table.

Type of Car	Observations									
Subcompact	3	5	3	7	6	5	3	2	1	6
Compact	1	3	4	7	5	6	3	2	1	7
Midsize	4	1	3	5	7	1	2	4	2	7
Full size	3	5	7	5	10	3	4	7	2	7

- (a) Is there evidence to support a claim that the type of car rented affects the length of the rental contract? Use $\alpha = 0.05$. If so, which types of cars are responsible for the difference?
- (b) Analyze the residuals from this experiment and comment on model adequacy.
- (c) Notice that the response variable in this experiment is a count. Should this cause any potential concerns about the validity of the analysis of variance?

3.14. I belong to a golf club in my neighborhood. I divide the year into three golf seasons: summer (June–September), winter (November–March), and shoulder (October, April, and May). I believe that I play my best golf during the summer (because I have more time and the course isn't crowded) and shoulder (because the course isn't crowded) seasons, and my worst golf is during the winter (because when all of the part-year residents show up, the course is crowded, play is slow, and I get frustrated). Data from the last year are shown in the following table.

Season	Observations									
Summer	83	85	85	87	90	88	88	84	91	90
Shoulder	91	87	84	87	85	86	83			
Winter	94	91	87	85	87	91	92	86		

- (a) Do the data indicate that my opinion is correct? Use $\alpha = 0.05$.
- (b) Analyze the residuals from this experiment and comment on model adequacy.

3.15. A regional opera company has tried three approaches to solicit donations from 24 potential sponsors. The 24 potential sponsors were randomly divided into three groups of eight, and one approach was used for each group. The dollar amounts of the resulting contributions are shown in the following table.

Approach	Contributions (in \$)							
1	1000	1500	1200	1800	1600	1100	1000	1250
2	1500	1800	2000	1200	2000	1700	1800	1900
3	900	1000	1200	1500	1200	1550	1000	1100


- (a) Do the data indicate that there is a difference in results obtained from the three different approaches? Use $\alpha = 0.05$.
- (b) Analyze the residuals from this experiment and comment on model adequacy.

3.16. An experiment was run to determine whether four specific firing temperatures affect the density of a certain type of brick. A completely randomized experiment led to the following data:

Temperature	Density				
100	21.8	21.9	21.7	21.6	21.7
125	21.7	21.4	21.5	21.4	
150	21.9	21.8	21.8	21.6	21.5
175	21.9	21.7	21.8	21.4	

- (a) Does the firing temperature affect the density of the bricks? Use $\alpha = 0.05$.
- (b) Is it appropriate to compare the means using the Fisher LSD method (for example) in this experiment?
- (c) Analyze the residuals from this experiment. Are the analysis of variance assumptions satisfied?
- (d) Construct a graphical display of the treatment as described in Section 3.5.3. Does this graph adequately summarize the results of the analysis of variance in part (a)?


3.17. Rework part (d) of Problem 3.16 using the Tukey method. What conclusions can you draw? Explain carefully how you modified the technique to account for unequal sample sizes.

 **3.18.** A manufacturer of television sets is interested in the effect on tube conductivity of four different types of coating for color picture tubes. A completely randomized experiment is conducted and the following conductivity data are obtained:

Coating Type	Conductivity			
1	143	141	150	146
2	152	149	137	143
3	134	136	132	127
4	129	127	132	129

- (a) Is there a difference in conductivity due to coating type? Use $\alpha = 0.05$.
- (b) Estimate the overall mean and the treatment effects.
- (c) Compute a 95 percent confidence interval estimate of the mean of coating type 4. Compute a 99 percent confidence interval estimate of the mean difference between coating types 1 and 4.


- (d) Test all pairs of means using the Fisher LSD method with $\alpha = 0.05$.
- (e) Use the graphical method discussed in Section 3.5.3 to compare the means. Which coating type produces the highest conductivity?
- (f) Assuming that coating type 4 is currently in use, what are your recommendations to the manufacturer? We wish to minimize conductivity.

3.19. Reconsider the experiment from Problem 3.18. Analyze the residuals and draw conclusions about model adequacy. 

3.20. An article in the *ACI Materials Journal* (Vol. 84, 1987, pp. 213–216) describes several experiments investigating the rodding of concrete to remove entrapped air. A 3-inch \times 6-inch cylinder was used, and the number of times this rod was used is the design variable. The resulting compressive strength of the concrete specimen is the response. The data are shown in the following table:

Rodding Level	Compressive Strength		
	10	1530	1530
15	1610	1650	1500
20	1560	1730	1530
25	1500	1490	1510

- (a) Is there any difference in compressive strength due to the rodding level? Use $\alpha = 0.05$.
- (b) Find the P -value for the F statistic in part (a).
- (c) Analyze the residuals from this experiment. What conclusions can you draw about the underlying model assumptions?
- (d) Construct a graphical display to compare the treatment means as described in Section 3.5.3.

3.21. An article in *Environment International* (Vol. 18, No. 4, 1992) describes an experiment in which the amount of radon released in showers was investigated. Radon-enriched water was used in the experiment, and six different orifice diameters were tested in shower heads. The data from the experiment are shown in the following table: 

Orifice Diameter	Radon Released (%)			
0.37	80	83	83	85
0.51	75	75	79	79
0.71	74	73	76	77
1.02	67	72	74	74
1.40	62	62	67	69
1.99	60	61	64	66


- (a) Does the size of the orifice affect the mean percentage of radon released? Use $\alpha = 0.05$.

- (b) Find the P -value for the F statistic in part (a).
- (c) Analyze the residuals from this experiment.
- (d) Find a 95 percent confidence interval on the mean percent of radon released when the orifice diameter is 1.40.
- (e) Construct a graphical display to compare the treatment means as described in Section 3.5.3 What conclusions can you draw?

3.22. The response time in milliseconds was determined for three different types of circuits that could be used in an automatic valve shutoff mechanism. The results from a completely randomized experiment are shown in the following table:

Circuit Type	Response Time					
1	9	12	10	8	15	
2	20	21	23	17	30	
3	6	5	8	16	7	

- (a) Test the hypothesis that the three circuit types have the same response time. Use $\alpha = 0.01$.
- (b) Use Tukey's test to compare pairs of treatment means. Use $\alpha = 0.01$.
- (c) Use the graphical procedure in Section 3.5.3 to compare the treatment means. What conclusions can you draw? How do they compare with the conclusions from part (b)?
- (d) Construct a set of orthogonal contrasts, assuming that at the outset of the experiment you suspected the response time of circuit type 2 to be different from the other two.
- (e) If you were the design engineer and you wished to minimize the response time, which circuit type would you select?
- (f) Analyze the residuals from this experiment. Are the basic analysis of variance assumptions satisfied?

 **3.23.** The effective life of insulating fluids at an accelerated load of 35 kV is being studied. Test data have been obtained for four types of fluids. The results from a completely randomized experiment were as follows:

Fluid Type	Life (in h) at 35 kV Load					
1	17.6	18.9	16.3	17.4	20.1	21.6
2	16.9	15.3	18.6	17.1	19.5	20.3
3	21.4	23.6	19.4	18.5	20.5	22.3
4	19.3	21.1	16.9	17.5	18.3	19.8

- (a) Is there any indication that the fluids differ? Use $\alpha = 0.05$.
- (b) Which fluid would you select, given that the objective is long life?

- (c) Analyze the residuals from this experiment. Are the basic analysis of variance assumptions satisfied?

3.24. Four different designs for a digital computer circuit are being studied to compare the amount of noise present. The following data have been obtained:


Circuit Design	Noise Observed				
1	19	20	19	30	8
2	80	61	73	56	80
3	47	26	25	35	50
4	95	46	83	78	97

- (a) Is the same amount of noise present for all four designs? Use $\alpha = 0.05$.
- (b) Analyze the residuals from this experiment. Are the analysis of variance assumptions satisfied?
- (c) Which circuit design would you select for use? Low noise is best.

3.25. Four chemists are asked to determine the percentage of methyl alcohol in a certain chemical compound. Each chemist makes three determinations, and the results are the following:

Chemist	Percentage of Methyl Alcohol		
1	84.99	84.04	84.38
2	85.15	85.13	84.88
3	84.72	84.48	85.16
4	84.20	84.10	84.55

- (a) Do chemists differ significantly? Use $\alpha = 0.05$.
- (b) Analyze the residuals from this experiment.
- (c) If chemist 2 is a new employee, construct a meaningful set of orthogonal contrasts that might have been useful at the start of the experiment.

3.26. Three brands of batteries are under study. It is suspected that the lives (in weeks) of the three brands are different. Five randomly selected batteries of each brand are tested with the following results: 

Weeks of Life		
Brand 1	Brand 2	Brand 3
100	76	108
96	80	100
92	75	96
96	84	98
92	82	100

- (a) Are the lives of these brands of batteries different?
- (b) Analyze the residuals from this experiment.
- (c) Construct a 95 percent confidence interval estimate on the mean life of battery brand 2. Construct a 99 percent confidence interval estimate on the mean difference between the lives of battery brands 2 and 3.
- (d) Which brand would you select for use? If the manufacturer will replace without charge any battery that fails in less than 85 weeks, what percentage would the company expect to replace?

3.27. Four catalysts that may affect the concentration of one component in a three-component liquid mixture are being investigated. The following concentrations are obtained from a completely randomized experiment:

Catalyst			
1	2	3	4
58.2	56.3	50.1	52.9
57.2	54.5	54.2	49.9
58.4	57.0	55.4	50.0
55.8	55.3		51.7
54.9			

- (a) Do the four catalysts have the same effect on the concentration?
- (b) Analyze the residuals from this experiment.
- (c) Construct a 99 percent confidence interval estimate of the mean response for catalyst 1.



3.28. An experiment was performed to investigate the effectiveness of five insulating materials. Four samples of each material were tested at an elevated voltage level to accelerate the time to failure. The failure times (in minutes) are shown below:

Material	Failure Time (minutes)			
1	110	157	194	178
2	1	2	4	18
3	880	1256	5276	4355
4	495	7040	5307	10,050
5	7	5	29	2

- (a) Do all five materials have the same effect on mean failure time?
- (b) Plot the residuals versus the predicted response. Construct a normal probability plot of the residuals. What information is conveyed by these plots?
- (c) Based on your answer to part (b) conduct another analysis of the failure time data and draw appropriate conclusions.

3.29. A semiconductor manufacturer has developed three different methods for reducing particle counts on wafers. All three methods are tested on five different wafers and the after treatment particle count obtained. The data are shown below:

Method	Count				
1	31	10	21	4	1
2	62	40	24	30	35
3	53	27	120	97	68

- (a) Do all methods have the same effect on mean particle count?
- (b) Plot the residuals versus the predicted response. Construct a normal probability plot of the residuals. Are there potential concerns about the validity of the assumptions?
- (c) Based on your answer to part (b) conduct another analysis of the particle count data and draw appropriate conclusions.

3.30. A manufacturer suspects that the batches of raw material furnished by his supplier differ significantly in calcium content. There are a large number of batches currently in the warehouse. Five of these are randomly selected for study. A chemist makes five determinations on each batch and obtains the following data:

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

- (a) Is there significant variation in calcium content from batch to batch? Use $\alpha = 0.05$.
- (b) Estimate the components of variance.
- (c) Find a 95 percent confidence interval for $\sigma_7^2 / (\sigma_7^2 + \sigma^2)$.
- (d) Analyze the residuals from this experiment. Are the analysis of variance assumptions satisfied?

3.31. Several ovens in a metal working shop are used to heat metal specimens. All the ovens are supposed to operate at the same temperature, although it is suspected that this may not be true. Three ovens are selected at random, and their temperatures on successive heats are noted. The data collected are as follows:

Oven	Temperature					
1	491.50	498.30	498.10	493.50	493.60	
2	488.50	484.65	479.90	477.35		
3	490.10	484.80	488.25	473.00	471.85	478.65

- (a) Is there significant variation in temperature between ovens? Use $\alpha = 0.05$.
- (b) Estimate the components of variance for this model.
- (c) Analyze the residuals from this experiment and draw conclusions about model adequacy.

3.32. An article in the *Journal of the Electrochemical Society* (Vol. 139, No. 2, 1992, pp. 524–532) describes an experiment to investigate the low-pressure vapor deposition of polysilicon. The experiment was carried out in a large-capacity reactor at Sematech in Austin, Texas. The reactor has several wafer positions, and four of these positions are selected at random. The response variable is film thickness uniformity. Three replicates of the experiment were run, and the data are as follows:

Wafer Position	Uniformity		
1	2.76	5.67	4.49
2	1.43	1.70	2.19
3	2.34	1.97	1.47
4	0.94	1.36	1.65

- (a) Is there a difference in the wafer positions? Use $\alpha = 0.05$.
- (b) Estimate the variability due to wafer positions.
- (c) Estimate the random error component.
- (d) Analyze the residuals from this experiment and comment on model adequacy.

3.33. Consider the vapor-deposition experiment described in Problem 3.32.

- (a) Estimate the total variability in the uniformity response.
- (b) How much of the total variability in the uniformity response is due to the difference between positions in the reactor?
- (c) To what level could the variability in the uniformity response be reduced if the position-to-position variability in the reactor could be eliminated? Do you believe this is a significant reduction?

3.34. An article in the *Journal of Quality Technology* (Vol. 13, No. 2, 1981, pp. 111–114) describes an experiment that investigates the effects of four bleaching chemicals on pulp brightness. These four chemicals were selected at random from a large population of potential bleaching agents. The data are as follows:

Oven	Temperature				
1	77.199	74.466	92.746	76.208	82.876
2	80.522	79.306	81.914	80.346	73.385
3	79.417	78.017	91.596	80.802	80.626
4	78.001	78.358	77.544	77.364	77.386

- (a) Is there a difference in the chemical types? Use $\alpha = 0.05$.
- (b) Estimate the variability due to chemical types.
- (c) Estimate the variability due to random error.
- (d) Analyze the residuals from this experimental and comment on model adequacy.

3.35. Consider the single-factor random effects model discussed in this chapter. Develop a procedure for finding a $100(1 - \alpha)\%$ confidence interval on the ratio $\sigma^2/(\sigma_\tau^2 + \sigma^2)$. Assume that the experiment is balanced.

3.36. Consider testing the equality of the means of two normal populations, where the variances are unknown but are assumed to be equal. The appropriate test procedure is the pooled t -test. Show that the pooled t -test is equivalent to the single-factor analysis of variance.

3.37. Show that the variance of the linear combination $\sum_{i=1}^a c_i y_i$ is $\sigma^2 \sum_{i=1}^a n_i c_i^2$.

3.38. In a fixed effects experiment, suppose that there are n observations for each of the four treatments. Let Q_1^2, Q_2^2, Q_3^2 be single-degree-of-freedom components for the orthogonal contrasts. Prove that $SS_{\text{Treatments}} = Q_1^2 + Q_2^2 + Q_3^2$.

3.39. Use Bartlett's test to determine if the assumption of equal variances is satisfied in Problem 3.24. Use $\alpha = 0.05$. Did you reach the same conclusion regarding equality of variances by examining residual plots?

3.40. Use the modified Levene test to determine if the assumption of equal variances is satisfied in Problem 3.26. Use $\alpha = 0.05$. Did you reach the same conclusion regarding the equality of variances by examining residual plots?

3.41. Refer to Problem 3.22. If we wish to detect a maximum difference in mean response times of 10 milliseconds with a probability of at least 0.90, what sample size should be used? How would you obtain a preliminary estimate of σ^2 ?

3.42. Refer to Problem 3.26.

- (a) If we wish to detect a maximum difference in battery life of 10 hours with a probability of at least 0.90, what sample size should be used? Discuss how you would obtain a preliminary estimate of σ^2 for answering this question.
- (b) If the difference between brands is great enough so that the standard deviation of an observation is increased by 25 percent, what sample size should be used if we wish to detect this with a probability of at least 0.90?


3.43. Consider the experiment in Problem 3.26. If we wish to construct a 95 percent confidence interval on the difference in two mean battery lives that has an accuracy of ± 2 weeks, how many batteries of each brand must be tested?

3.44. Suppose that four normal populations have means of $\mu_1 = 50, \mu_2 = 60, \mu_3 = 50,$ and $\mu_4 = 60$. How many observations should be taken from each population so that the probability of rejecting the null hypothesis of equal population means is at least 0.90? Assume that $\alpha = 0.05$ and that a reasonable estimate of the error variance is $\sigma^2 = 25$.



- 3.45.** Refer to Problem 3.44.
- (a) How would your answer change if a reasonable estimate of the experimental error variance were $\sigma^2 = 36$?
 - (b) How would your answer change if a reasonable estimate of the experimental error variance were $\sigma^2 = 49$?
 - (c) Can you draw any conclusions about the sensitivity of your answer in this particular situation about how your estimate of σ affects the decision about sample size?
 - (d) Can you make any recommendations about how we should use this general approach to choosing n in practice?

3.46. Refer to the aluminum smelting experiment described in Section 3.8.3. Verify that ratio control methods do not affect average cell voltage. Construct a normal probability plot of the residuals. Plot the residuals versus the predicted values. Is there an indication that any underlying assumptions are violated?

 **3.47.** Refer to the aluminum smelting experiment in Section 3.8.3. Verify the ANOVA for pot noise summarized in Table 3.16. Examine the usual residual plots and comment on the experimental validity.

3.48. Four different feed rates were investigated in an experiment on a CNC machine producing a component part used in an aircraft auxiliary power unit. The manufacturing engineer in charge of the experiment knows that a critical part dimension of interest may be affected by the feed rate. However, prior experience has indicated that only dispersion effects are likely to be present. That is, changing the feed rate does not affect the *average* dimension, but it could affect dimensional variability. The engineer makes five production runs at each feed rate and obtains the standard deviation of the critical dimension (in 10^{-3} mm). The data are shown below. Assume that all runs were made in random order.

Feed Rate (in/min)	Production Run				
	1	2	3	4	5
10	0.09	0.10	0.13	0.08	0.07
12	0.06	0.09	0.12	0.07	0.12
14	0.11	0.08	0.08	0.05	0.06
16	0.19	0.13	0.15	0.20	0.11


- (a) Does feed rate have any effect on the standard deviation of this critical dimension?
- (b) Use the residuals from this experiment to investigate model adequacy. Are there any problems with experimental validity?

3.49. Consider the data shown in Problem 3.22.

- (a) Write out the least squares normal equations for this problem and solve them for $\hat{\mu}$ and $\hat{\tau}_i$, using the usual constraint ($\sum_{i=1}^3 \hat{\tau}_i = 0$). Estimate $\tau_1 - \tau_2$.

- (b) Solve the equations in (a) using the constraint $\hat{\tau}_3 = 0$. Are the estimators $\hat{\tau}_i$ and $\hat{\mu}$ the same as you found in (a)? Why? Now estimate $\tau_1 - \tau_2$ and compare your answer with that for (a). What statement can you make about estimating contrasts in the τ_i ?
- (c) Estimate $\mu + \tau_1$, $2\tau_1 - \tau_2 - \tau_3$, and $\mu + \tau_1 + \tau_2$ using the two solutions to the normal equations. Compare the results obtained in each case.

3.50. Apply the general regression significance test to the experiment in Example 3.5. Show that the procedure yields the same results as the usual analysis of variance.

3.51. Use the Kruskal–Wallis test for the experiment in Problem 3.23. Compare the conclusions obtained with those from the usual analysis of variance. 

3.52. Use the Kruskal–Wallis test for the experiment in Problem 3.23. Are the results comparable to those found by the usual analysis of variance?

3.53. Consider the experiment in Example 3.5. Suppose that the largest observation on etch rate is incorrectly recorded as 250 Å/min. What effect does this have on the usual analysis of variance? What effect does it have on the Kruskal–Wallis test?

3.54. A textile mill has a large number of looms. Each loom is supposed to provide the same output of cloth per minute. To investigate this assumption, five looms are chosen at random, and their output is noted at different times. The following data are obtained:

Loom	Output (lb/min)				
1	14.0	14.1	14.2	14.0	14.1
2	13.9	13.8	13.9	14.0	14.0
3	14.1	14.2	14.1	14.0	13.9
4	13.6	13.8	14.0	13.9	13.7
5	13.8	13.6	13.9	13.8	14.0

- (a) Explain why this is a random effects experiment. Are the looms equal in output? Use $\alpha = 0.05$.
- (b) Estimate the variability between looms.
- (c) Estimate the experimental error variance.
- (d) Find a 95 percent confidence interval for $\sigma_\tau^2 / (\sigma_\tau^2 + \sigma^2)$.
- (e) Analyze the residuals from this experiment. Do you think that the analysis of variance assumptions are satisfied?
- (f) Use the REML method to analyze this data. Compare the 95 percent confidence interval on the error variance from REML with the exact chi-square confidence interval.

3.55. A manufacturer suspects that the batches of raw material furnished by his supplier differ significantly in calcium content. There are a large number of batches currently in the warehouse. Five of these are randomly selected for study.

A chemist makes five determinations on each batch and obtains the following data:

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

- Is there significant variation in calcium content from batch to batch? Use $\alpha = 0.05$.
- Estimate the components of variance.
- Find a 95 percent confidence interval for $\sigma_{\tau}^2 / (\sigma_{\tau}^2 + \sigma^2)$.
- Analyze the residuals from this experiment. Are the analysis of variance assumptions satisfied?
- Use the REML method to analyze this data. Compare the 95 percent confidence interval on the error variance from REML with the exact chi-square confidence interval.