CHAPTER 12

Comparing Several Means: Single-Factor and Randomized Block Experiments

IN THIS CHAPTER

Multiple comparisons (Bonferroni method)
One-way analysis of variance
Kruskal–Wallis test
Randomized block experiments
Friedman's test

Are there any differences in average fruit yield among apple trees treated with four different fertilizer supplements? Do four thermometers all give the same average reading of the melting point of a substance? What are the relative biological effects, on average, of three different anesthetics? Each of these questions concerns a comparison of three or more measures of central tendency.

We have already discussed ways to compare two measures of central tendency, considering both two-sample and paired-sample experimental designs. Now we consider inferences about more than two measures of central tendency, in single-factor experiments and randomized block experiments. The single-factor experiment is an extension to several populations of the two-sample experimental design, whereas the randomized block experiment is an extension of the paired-sample experimental design.

In a single-factor experiment, we have several independent random samples, one from each of several populations of interest. We want to use the samples to compare the means of the populations. The single-factor experiment is an extension of the two- (independent) sample experimental design we discussed in Chapter 11. In Section 12-2 we discuss the classical, parametric, way to test whether several means are equal: one-way analysis of variance. We consider a nonparametric analysis, the Kruskal–Wallis test, in Section 12-3.

A randomized block experimental design is an extension of the paired-sample design. In Section 12-4 we discuss the classical, parametric, analysis of a randomized block experiment. We consider a nonparametric analysis, Friedman's test, in Section 12-5.

We begin in Section 12-1 with a fairly crude, but useful, approach to comparing several means: the Bonferroni method of comparing means (or medians) two at a time.

12-1 Comparing Measures of Central Tendency Two at a Time Using the Bonferroni Method

Why do we need special procedures for comparing more than two location parameters? Why can we not just compare them two at a time? There is a problem with that strategy. To illustrate, suppose we have four independent random samples and we want to make inferences about the medians of the populations sampled.

For simplicity, let's consider just two possible comparisons. Suppose we test whether the medians of the first two populations are equal:

$$H_0\colon\ M_1 = M_2 \qquad \text{versus} \qquad H_a\colon\ M_1 \neq M_2$$

and we also test whether the medians of the other two populations are equal:

$$H_0\colon\ M_3 = M_4 \qquad \text{versus} \qquad H_a\colon\ M_3 \neq M_4$$

Since the samples are independent, we can say these two tests are independent.

For each of these two sets of hypotheses, we use a test statistic to measure how far the observed results are from what we would expect under the appı priate null hypothesis. To test the first set of hypotheses, we select a decisi rule for deciding whether the two population medians M_1 and M_2 are diffe ent—say, with significance level .05. Similarly, we choose a decision rule f deciding whether the two population medians M_3 and M_4 are different, al. with significance level .05.

Consider the two tests together. We can think of a combined null h pothesis that the first two population medians are equal *and* the other tw population medians are equal:

$$H_0: \quad M_1 = M_2 \text{ and } M_3 = M_4$$

The combined alternative states that at least one of these equalities does nc hold:

$$H_a: \quad M_1 \neq M_2 \text{ or } M_3 \neq M_4$$

We reject the combined null hypothesis if either test statistic is in the corresponding rejection region. What is the significance level for this combinec criterion?

The probability that the results are consistent with the null hypothesis H_0: $M_1 = M_2$ when these two medians really are equal is $1 - .05 = .95$ (because our significance level is .05). Likewise, the probability that the results are consistent with the null hypothesis H_0: $M_3 = M_4$ when these two medians really are equal is $1 - .05 = .95$.

Suppose the combined null hypothesis is true, so $M_1 = M_2$ and $M_3 = M_4$. The chance the results are consistent with this combined null hypothesis is $.95 \times .95 = .9025$, because the probability of two independent events occurring together is the product of the separate probabilities (Chapter 6). Therefore, the significance level for the combined test is $1 - .9025 = .0975$, almost twice the significance level for either of the separate tests! We have nearly a 10% chance of rejecting the combined null hypothesis when that combined null hypothesis is really true (that is, when $M_1 = M_2$ and $M_3 = M_4$). To test the overall null hypothesis H_0: $M_1 = M_2 = M_3 = M_4$ is even worse, since more pairwise comparisons are necessary and they are not all independent.

One way around this difficulty is to use the *Bonferroni method* for controlling the overall significance level (Rice, 1988, page 384). Suppose m pairwise comparisons are necessary to test a combined null hypothesis. We select a decision rule for each pairwise comparison. Denote the m significance levels associated with these m decision rules by α_1 through α_m. If α is the significance level for the overall test, then α is less than or equal to the sum of α_1 through α_m (Exercise 12-16).

The **Bonferroni method** is a technique for obtaining an upper bound on an overall significance level.

Suppose we have m separate tests of hypotheses. Using the significance level approach, we have a decision rule and associated significance level for each test. For a combined test of hypotheses, we say the results are inconsistent with the combined null hypothesis if any one of the m separate test statistics is in its associated rejection region. The significance level for this combined test is less than or equal to the sum of the significance levels of the separate tests.

By making the pairwise significance levels α_1 through α_m small, we can control the size of the overall significance level α. We will use this idea for the multiple comparisons procedures we discuss in Sections 12-2 and 12-3, when we want to calculate interval estimates for the differences between measures of central tendency for three or more populations.

Multiple comparisons of means refers to the process of comparing several means. For our purposes, multiple comparisons refers to the process of comparing several means, two at a time.

There is another way around the difficulty of comparing several measures of central tendency in a single-factor experiment. We can use one-way analysis of variance (Section 12-2) or the Kruskal–Wallis test (Section 12-3) to test the null hypothesis that the population means (or medians) are all equal. One-way analysis of variance is an extension of the two-sample t test (Section 11-3) to several independent samples. The Kruskal–Wallis test is an extension to several independent samples of the Wilcoxon–Mann–Whitney test for two independent samples (Section 11-4).

12-2 Inferences About Several Means in a Single-Factor Experiment: One-Way Analysis of Variance

In a single-factor experiment, we have several independent random samples and we want to use these observations to make inferences about the populations sampled.

In a **single-factor experiment**, we have several independent random samples and we want to make inferences about the populations sampled.

We are here concerned with making inferences about the means of the populations sampled. Suppose we have k independent random samples, one from each of k populations. For a classical analysis, we assume that the values in each population follow a Gaussian distribution and that all k of these Gaussian distributions have the same variance, σ^2. Let μ_1 through μ_k denote the k population means. We want to test the null hypothesis that these means are all equal:

$$H_0: \quad \mu_1 \text{ through } \mu_k \text{ are all equal}$$

The alternative hypothesis is that these means are not all equal:

$$H_a: \quad \mu_1 \text{ through } \mu_k \text{ are not all equal}$$

We call the procedure for testing these hypotheses *one-way analysis c variance.*

> **One-way analysis of variance** is the classical, parametric, approach to testing the null hypothesis that the population means are all equal in a single-factor experiment.

Let's look at an example; then we will discuss one-way analysis of variance an apply it to this example.

EXAMPLE 12-1

Does a nitrogen supplement improve apple production? To address this ques tion, researchers divided Jonathan apple trees into four treatment groups. The applied no nitrogen to the trees in the control group. They provided a nitroger supplement to the trees in the other three groups: either urea, potassium nitrate plus calcium, or ammonia plus ammonium sulphate. The researchers stored fruits of 42 trees for 4 months. They then weighed samples of fruit from each tree. Fruit weight, in grams, for each tree is shown below (data contributed by D. A. Ratkowsky to a collection of problems in Andrews and Herzberg, 1985, pages 355–356; from D. A. Ratkowsky and D. Martin, 1974). The researchers wanted to use the sample weights to compare mean fruit production under the four treatments.

Control	Urea	Potassium nitrate and calcium	Ammonia and ammonium sulphate
85.3	117.5	127.1	77.4
113.8	98.9	108.5	91.3
92.9	108.5	99.9	91.3
48.9	104.4	124.8	81.7
99.4	96.8	94.5	89.2
79.1	94.5	99.4	69.6
70.0	90.6	117.5	69.0
86.9	100.8	135.0	73.7
87.7	96.0	85.6	75.1
67.3	99.9	102.5	87.0
	84.6	110.8	

A plot of the observations by treatment group is shown in Figure 12-1. What does this plot suggest about the relative effects of the four treatments upon apple production? What suggestions would you make for carrying out this experiment, in order to reduce the effects of extraneous factors and allow valid comparisons across treatment groups?

We will use one-way analysis of variance to test the null hypothesis that the mean fruit weight is the same for all four treatments in Example 12-1. Before we can discuss this procedure, we need some notation.

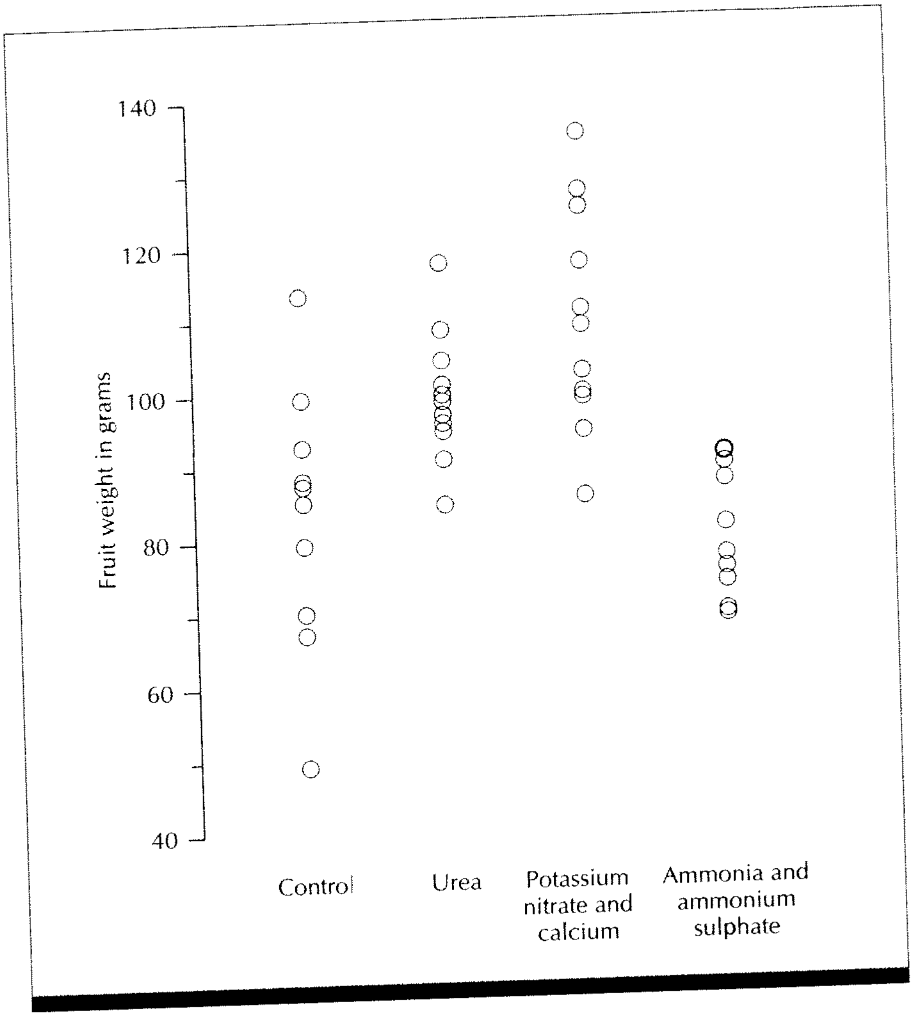

FIGURE 12-1 Fruit weight (in grams) from apple trees in four treatment groups, Example 12-1

One-Way Analysis of Variance for a Single-Factor Experiment

Suppose we have k samples. Let Y_{ij} denote the observation on experimental unit j in sample i. Let $\overline{Y}_i$ denote the sample mean, s_i^2 the sample variance, and n_i the size of sample i. As we did in the two-sample case, we can combine the separate sample variances into a pooled estimate of σ^2. This *pooled estimate of σ^2* is a weighted average of s_1^2 through s_k^2, calculated as follows:

The pooled estimate of the common population variance σ^2 in a single-factor experiment is

$$s_r^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}{N - k}$$

Here, N is the total sample size, the sum of n_1 through n_k. The double summation notation means that for each group i, we sum over the observations in that group, then sum over the groups.

Common names for this variance estimator s_r^2 are **residual mean square, within-groups mean square**, and **mean square within.**

We use the notation s_r^2 because it is an estimate of σ^2 based on residuals. Recall that a residual is the difference between an observation and a summary or estimate for the mean of the observation. In the case of a single-factor experiment, a residual is the difference between an observation and its group mean. (For more on residuals, see Tukey, 1977.)

A **residual** is the difference between an observation and an estimate of its expected value.

In a single-factor experiment, a residual is the difference $Y_{ij} - \bar{Y}_i$ between an observation and the average of all the observations in the same group.

In a single-factor experiment, the group mean is a summary value, estimating the mean of the population sampled. We see from the definition that s_r^2 is an average of the squares of residuals $Y_{ij} - \bar{Y}_i$; hence the name residual mean square.

The *between-groups variance estimate* is another measure of variation we need for our analysis:

The between-groups variance estimate in a single-factor experiment is

$$s_B^2 = \frac{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2}{k - 1}$$

where $\bar{Y}$ is the average of all N observations in the combined samples.

s_B^2 is sometimes called the **between-groups mean square** or **mean square between.**

We want to test the null hypothesis that the means of the populations sampled are all equal. The test statistic is the ratio of the between-groups variance estimate and the within-groups variance estimate:

The test statistic for testing the null hypothesis that the population means are all equal, in a single-factor experiment, is

$$\text{Test statistic} = \frac{s_B^2}{s_r^2}$$

To find the probability distribution of the test statistic under the null hypothesis, assume that we have k independent random samples from Gaussian populations with the same variance. Then under the null hypothesis the test statistic has the F distribution with $k - 1$ numerator degrees of freedom and $N - k$ denominator degrees of freedom.

The numerator degrees of freedom and denominator degrees of freedom are two constants (or parameters) that define an F distribution. The test statistic defined above has an F distribution under the null hypothesis that all the population means are equal. The numerator degrees of freedom of this F distribution equal $k - 1$, used in calculating the ***numerator*** of the test statistic, s_B^2. (We call $k - 1$ the degrees of freedom associated with the between-groups mean square s_B^2.) The denominator degrees of freedom of this F distribution equal $N - k$, used in calculating the ***denominator*** of the test statistic, s_r^2. (We call $N - k$ the degrees of freedom associated with the residual mean square s_r^2.)

In general, an F distribution has numerator degrees of freedom d_1 and denominator degrees of freedom d_2. We denote such an F distribution by $F(d_1, d_2)$. An F distribution is a continuous probability distribution that is skewed to the right. A random variable having an F distribution takes on positive values only. (Our test statistic is a ratio of two variance estimates, and so can have only positive values.) The shape of an F distribution is illustrated in Figure 12-2. Table D at the back of the book lists values of c for which

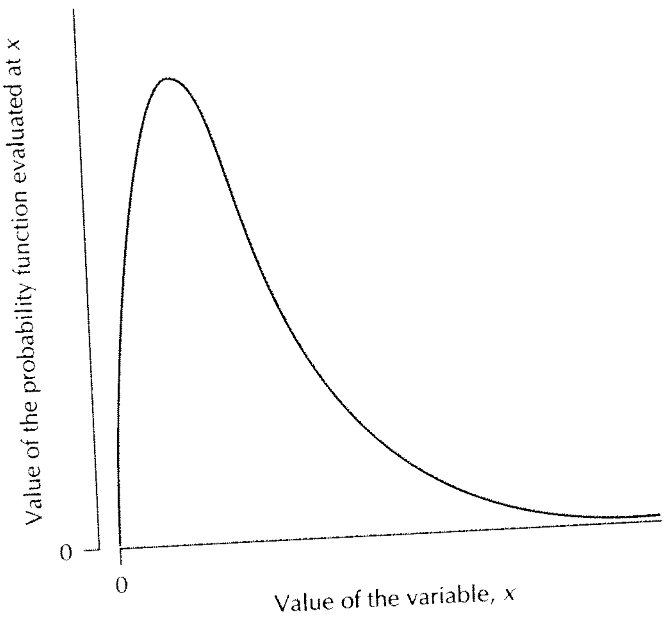

FIGURE 12-2 Illustration of the shape of an F distribution

$P(F \leq c)$ has specified values, where F is a random variable having selecte numerator degrees of freedom d_1 and denominator degrees of freedom d_2

An F distribution is also called a variance ratio distribution. Our test statistic is a ratio of two variance estimates, which has an F distribution if the nul hypothesis is true. Even though we are comparing hypotheses about population means, we call the procedure *analysis of variance* because our test statistic is a ratio of two variance estimates.

We know that s_r^2 is an estimate of σ^2, the variance in each population while s_B^2 is a measure of variation among the sample means. If the null hypothesis is true, the sample means $\bar{Y}_1$ through $\bar{Y}_k$ all estimate the common mean oi the k populations. Then s_B^2 is another estimate of σ^2. If the null hypothesis is not true, $\bar{Y}_1$ through $\bar{Y}_k$ do not all estimate the same mean. Then s_B^2 estimates the variation within populations plus the variation between the population means μ_1 through μ_k. Therefore, if the null hypothesis is not true, s_B^2 estimates something larger than σ^2.

With this in mind, we see that values of the test statistic near 1 are consistent with the null hypothesis. Values of the test statistic much larger than 1 are inconsistent with the null hypothesis. Using these ideas, we outline the significance level approach to one-way analysis of variance.

The significance level approach to comparing several means in a single-factor experiment, using one-way analysis of variance

1. The hypotheses are H_0: μ_1 through μ_k are all equal, and H_a: μ_1 through μ_k are not all equal, where μ_1 through μ_k represent the population means.
2. The test statistic is s_B^2/s_r^2, as defined above.
3. Assume that we have independent random samples from k Gaussian distributions with equal variances. Let N denote the sum of the k individual sample sizes. Then under the null hypothesis, the test statistic has the F distribution with $k - 1$ numerator degrees of freedom and $N - k$ denominator degrees of freedom.
4. Select significance level α.
5. Let F denote a random variable having the $F(k - 1, N - k)$ distribution. Find c from Table D at the back of the book such that $P(F \leq c) = 1 - \alpha$. Then the acceptance region is the interval $[0, c)$; the rejection region is the interval $[c, \infty)$.
6. The decision rule is:

   If test statistic $< c$, say the results are consistent with the null hypothesis.
   If test statistic $\geq c$, say the results are inconsistent with the null hypothesis.
7. Carry out an experiment that satisfies the conditions in step 3. Calculate the test statistic in step 2. Use the decision rule in step 6 to decide whether the results are consistent with the null hypothesis.

EXAMPLE 12-1
(continued)

In Example 12-1, we want to compare the effects of several nitrogen supplements upon fruit production in Jonathan apple trees. The null hypothesis states that the mean fruit weight is the same for all four treatments. The alternative states that the mean fruit weight is not the same for all four treatments. We can rewrite these hypotheses as

$$H_0\colon\ \mu_c = \mu_u = \mu_p = \mu_a$$
$$H_a\colon\ \text{The four means are not all equal}$$

where each subscript denotes a treatment group.

Assume that we have four independent random samples from Gaussian distributions with equal variances. Then under the null hypothesis, our test statistic would have the F distribution with $4 - 1 = 3$ numerator degrees of freedom and $42 - 4 = 38$ denominator degrees of freedom (since the total sample size N equals 42 and there are $k = 4$ treatment groups).

To verify the independence assumption, we would have to know more about how the experiment was conducted. What suggestions do you have for ensuring independence?

Figure 12-1 gives us no reason to doubt that each sample comes from a Gaussian distribution, since each of the four sample distributions is fairly symmetric. (One-way analysis of variance tends to be robust to deviations from the Gaussian assumption.)

The variation in fruit weights is somewhat larger in the control group and the potassium nitrate plus calcium group than in the other two groups, the least variation being in the ammonia plus ammonium sulphate group. However, these differences in variation are not extreme enough to make one-way analysis of variance seem inappropriate. (As with the two-sample t test, one-way analysis of variance is fairly robust to deviations from the equal-variance assumption. As long as the variances are not too different, actual significance levels and confidence levels are close to the levels we choose.)

Let's use significance level $\alpha = .01$. Since $1 - \alpha = .99$, we use the last page of Table D. For our test, there are 3 numerator degrees of freedom and 38 denominator degrees of freedom. Table D shows 3 numerator degrees of freedom but not 38 denominator degrees of freedom. We must choose either 30 or 40 for denominator degrees of freedom in the table. To be conservative (less likely to reject the null hypothesis), we will use the smaller value, 30. Then looking in the column for $d_1 = 3$ and the row for $d_2 = 30$, we find $c = 4.51$. The acceptance region is $[0, 4.51)$, the rejection region is $[4.51, \infty)$, and the decision rule is:

If test statistic < 4.51, say the results are consistent with the null hypothesis that there is no difference in mean fruit weight among the four treatments.

If test statistic ≥ 4.51, say the results are inconsistent with the null hypothesis, suggesting there is a difference in mean fruit weight among the four treatments.

The calculations we need for our analysis are outlined in Table 12-1.

Since the test statistic equals 11.28, we say the results are inconsistent with the null hypothesis, at the .01 significance level. The p-value, P(test statistic $\geq$ 11.28 when H_0 is true), is less than .01. This experiment suggests that mean

TABLE 12-1 Steps in calculating the one-way analysis of variance test statistic

Control	Urea	Potassium nitrate and calcium	Ammonia and ammonium sulphate
$\bar{Y}_c = 83.130$	$\bar{Y}_u = 99.318$	$\bar{Y}_p = 109.600$	$\bar{Y}_a = 80.530$
$s_1^2 = 327.949$	$s_2^2 = 77.662$	$s_3^2 = 230.006$	$s_4^2 = 76.525$
$n_1 = 10$	$n_2 = 11$	$n_3 = 11$	$n_4 = 10$
$N = 42$ $\bar{Y} = 93.683$			

$$s_r^2 = \frac{(10 - 1)(327.949) + (11 - 1)(77.662) + (11 - 1)(230.006) + (10 - 1)(76.525)}{42 - 4} = 176.76$$

$$s_B^2 = \frac{10(83.130 - 93.683)^2 + 11(99.318 - 93.683)^2 + 11(109.600 - 93.683)^2 + 10(80.530 - 93.683)^2}{4 - 1}$$

$$= 1{,}993.27$$

$$\text{Test statistic} = \frac{1{,}993.27}{176.76} = 11.28$$

Numerator degrees of freedom = 4 − 1 = 3
Denominator degrees of freedom = 42 − 4 = 38

TABLE 12-2 Analysis of variance table for a single-factor experiment

Source of variation	Sum of squares	Degrees of freedom	Mean square	Test statistic
Treatments (between groups)	$\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$	$k - 1$	s_B^2	$\frac{s_B^2}{s_r^2}$
Residual (within groups)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$N - k$	s_r^2	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	$N - 1$		

fruit weight is not the same for all four treatments. Looking at Figure 12-1, can you make a decision as to which group means are nearly the same and which are very different?

The Analysis of Variance Table for One-Way Analysis of Variance

We often summarize the calculations of one-way analysis of variance in a table called an *analysis of variance table*. Computer output for one-way analysis of variance is displayed in such a table. A general form of analysis of variance table for a single-factor experiment is shown in Table 12-2.

Computer output often has another column at the right of the table, showing the *p*-value associated with the test statistic. The last row in the table, the total row, is the sum of the previous two rows for sum of squares and degrees of freedom. We include this total row for completeness; we do not

TABLE 12-3 Analysis of variance table for the single-factor experiment in Example 12-1

Source of variation	Sum of squares	Degrees of freedom	Mean square	Test statistic	*p*-value
Between groups	5,979.8	3	1,993.27	11.28	0.0000
Residual (within groups)	6,716.9	38	176.76		
Total	12,696.7	41			

use it in our test of hypotheses. The analysis of variance table for Example 12-1 is shown in Table 12-3. The *p*-value is listed as 0.0000. This means that the *p*-value was less than .0001.

If our results are inconsistent with the null hypothesis, we would like to see where the differences are. Which population means seem to be similar and which seem to be different? To address this question formally, we can use a *multiple comparisons procedure*. There are many ways to make multiple comparisons. We will use a procedure based on the *Bonferroni method*.

We calculate a confidence interval for the difference $\mu_i - \mu_j$ between the means for populations i and j as

$$\bar{Y}_i - \bar{Y}_j \pm c\sqrt{\frac{s_r^2}{n_i} + \frac{s_r^2}{n_j}}$$

The subscripts i and j refer to samples i and j, respectively. The residual mean square s_r^2 is the pooled (within-groups) variance estimate based on the k samples. The number c comes from the t distribution with $N - k$ degrees of freedom, where N is the total sample size and k is the number of samples.

Suppose we calculate m such confidence intervals, making m pairwise comparisons of population means. Denote the confidence levels associated with these intervals by A_1 through A_m. Then the confidence level associated with the m intervals taken together is greater than or equal to $1 - \Sigma_{i=1}^{m} (1 - A_i)$.

The **Bonferroni method** is a technique for obtaining a lower bound on an overall confidence level.

Suppose we make m pairwise comparisons of means, with confidence levels A_1 through A_m. Using the Bonferroni method, we say the confidence level for the m intervals taken together is greater than or equal to

$1 - (1 - A_1) - (1 - A_2) - \cdots - (1 - A_m)$.

EXAMPLE 12-1 *(continued)*

Let's make multiple comparisons of mean fruit weights in Example 12-1, using the Bonferroni method. Because there are four groups, there are $\binom{4}{2} = 6$ possible pairwise comparisons, and we will let $m = 6$. Each separate interval will have confidence level .99. The total sample size is 42 and there are four

TABLE 12-4 Multiple comparisons for Example 12-1

Confidence interval for	Confidence interval
$\mu_c - \mu_u$	$83.1 - 99.3 \pm 2.750\sqrt{176.76\left(\frac{1}{10} + \frac{1}{11}\right)} = (-32.2, -.2)$
$\mu_c - \mu_p$	$83.1 - 109.6 \pm 2.750\sqrt{176.76\left(\frac{1}{10} + \frac{1}{11}\right)} = (-42.5, -10.5)$
$\mu_c - \mu_a$	$83.1 - 80.5 \pm 2.750\sqrt{176.76\left(\frac{1}{10} + \frac{1}{10}\right)} = (-13.8, 19.0)$
$\mu_u - \mu_p$	$99.3 - 109.6 \pm 2.750\sqrt{176.76\left(\frac{1}{11} + \frac{1}{11}\right)} = (-25.9, 5.3)$
$\mu_u - \mu_a$	$99.3 - 80.5 \pm 2.750\sqrt{176.76\left(\frac{1}{11} + \frac{1}{10}\right)} = (2.8, 34.8)$
$\mu_p - \mu_a$	$109.6 - 80.5 \pm 2.750\sqrt{176.76\left(\frac{1}{11} + \frac{1}{10}\right)} = (13.1, 45.1)$

The confidence level for each interval is .99. Therefore, the overall confidence level is greater than or equal to 1 − (.01 + .01 + .01 + .01 + .01 + .01) = .94.

Note: c = control, u = urea, p = potassium nitrate plus calcium, a = ammonia plus ammonium sulphate.

groups, so we get c from the t distribution with $42 - 4 = 38$ degrees of freedom. In Table C, we have a choice between 30 and 40 degrees of freedom. We will be more conservative (getting wider intervals) and use 30. Then $c = 2.750$. The calculations for the six confidence intervals are outlined in Table 12-4.

Zero is in the confidence interval for $\mu_c - \mu_a$ and the confidence interval for $\mu_u - \mu_p$, but not in the other intervals. Taken together, the intervals suggest that mean fruit weight is the same for the control group and the ammonia plus ammonium sulphate group. Mean fruit weight also seems to be the same for the urea group and the potassium nitrate plus calcium group; these two treatments appear to have mean fruit weights greater than the control and ammonia plus ammonium sulphate groups. This agrees with the visual comparisons we can make by examining the four distributions in Figure 12-1.

In Section 12-3, we discuss nonparametric tests of hypotheses and multiple comparisons for a single-factor experiment.

12-3 Nonparametric Analysis of a Single-Factor Experiment: The Kruskal–Wallis Test

The Kruskal–Wallis test is a nonparametric procedure used to check for equality of several distributions in a single-factor experiment. We will start with an example, then outline the significance level approach to the test of hypotheses, and apply it to the example. Finally, we discuss nonparametric multiple comparisons based on the Bonferroni method.

EXAMPLE 12-2 Maximal oxygen uptake is a measure of physical working capacity or aerobic power. In a survey of aerobic power in world-class athletes, Wilmore lists maximal oxygen uptake for nine young women athletes: three basketball players, four cross-country skiers, and two speed skaters (Wilmore, 1984). The results are shown below:

Sport	Maximal oxygen uptake $\left(\frac{\text{ml}}{\text{kg} \cdot \text{min}}\right)$
Basketball	42.3, 42.9, 49.6
Cross country skiing	56.9, 58.1, 61.5, 68.2
Speed skating	46.1, 52.0

Are there differences in aerobic power among women athletes in these three sports? How would you design an experiment to answer this question? What would you do to reduce the effects of extraneous factors? How should the experiment be conducted to ensure valid comparisons among the three sports?

A plot of the observations is shown in Figure 12-3. What does this plot

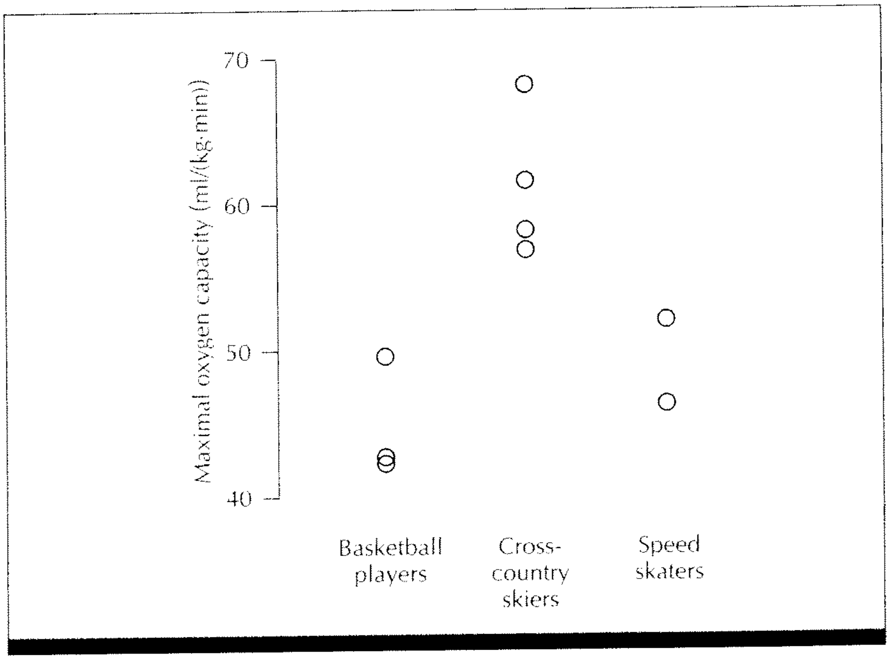

FIGURE 12-3 Plots of maximal oxygen uptake for female basketball players, cross-country skiers, and speed skaters in Example 12-2

suggest about the relative levels of aerobic power in women basketball players cross-country skiers, and speed skaters? Which groups of athletes seem similar and which seem very different, with respect to aerobic power?

We will use the Kruskal–Wallis test to compare the distributions of maximal oxygen uptake for women in the three sports. Before applying the test to this example, let's first describe it in general.

The Kruskal–Wallis Test for a Single-Factor Experiment

Suppose we have *k* independent random samples, one from each of *k* populations. The *k* distributions are continuous, with the same shape and variation, but possibly different locations (they may be shifted away from each other). Three continuous distributions with the same shape and variation are illustrated in Figure 12-4. A special case is one in which the distributions are all Gaussian with the same variance, the situation discussed in Section 12-2.

If the *k* distributions have the same shape and variation, then differences or shifts in location are described by differences between the population means (or by differences between the population medians). Our null hypothesis states that the *k* populations have the same location, and therefore the same distribution. This is the same as saying that the *k* populations have the same mean (and the same median).

Under the null hypothesis, the exact distribution of the Kruskal–Wallis test statistic described below is the Kruskal–Wallis distribution corresponding to the sample sizes in the experiment. A Kruskal–Wallis probability distribution is derived from the probability model for an experiment in which ranks 1 through *n* are randomly divided into three or more groups; see the Appendix on the Kruskal–Wallis distributions at the end of the text.

Table H at the back of the book lists probabilities of the form $P(KW \geq c)$, where *KW* denotes a random variable having a Kruskal–Wallis distribution. Table H covers only three groups ($k = 3$) and sample sizes from 2 to 5. There are many possible values for *k* and the sample sizes; it is not possible to table probabilities for many Kruskal–Wallis distributions. For situations not covered by Table H, we use an approximation to the distribution of the Kruskal–Wallis test statistic under the null hypothesis, comparing the test statistic with the chi-square distribution for $k - 1$ degrees of freedom. Degrees of freedom

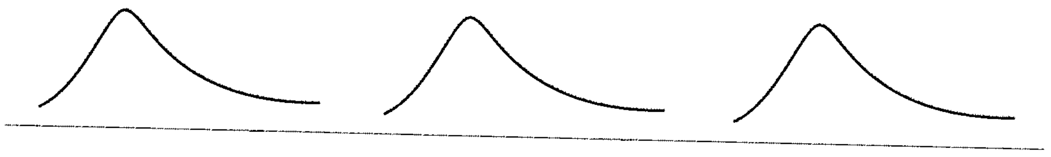

FIGURE 12-4 Illustration of three distributions having the same shape and variation, but different locations. The differences between the means μ_1, μ_2, and μ_3 of the distributions describe the differences in location.

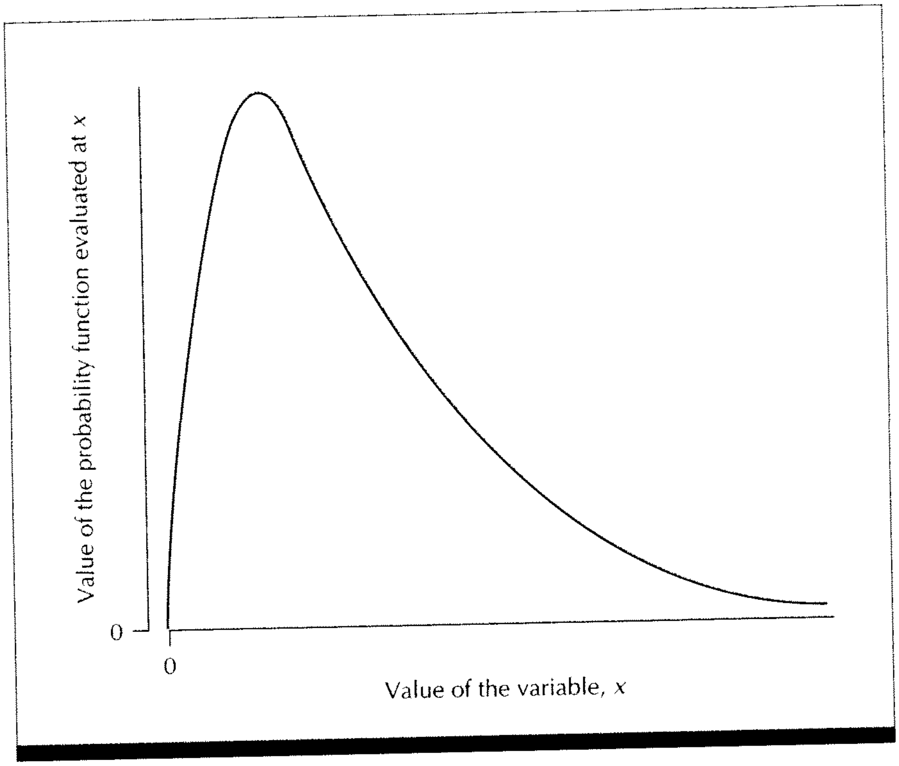

FIGURE 12-5 The shape of a chi-square distribution

here refers to the constant or parameter that defines a particular chi-square distribution.

A chi-square distribution is a continuous probability distribution that is skewed to the right. A random variable with such a distribution takes on only positive values. The general shape of a chi-square distribution is illustrated in Figure 12-5.

A chi-square distribution is characterized by a number called its degrees of freedom. We often denote the chi-square distribution with *d* degrees of freedom by χ^2_d. Some probabilities associated with several chi-square distributions are listed in Table E at the back of the book.

The Kruskal–Wallis procedure for comparing several distributions is outlined below.

The significance level approach to comparing several distributions in a single-factor experiment, using the Kruskal–Wallis test

1. The null hypothesis states that the *k* populations all have the same probability distribution. The alternative hypothesis states that the *k* distributions have the same shape and variation, but are shifted away from each other (they do not all have the same location).

2. Rank the observations in the combined samples from smallest to largest. If two or more observations have the same value, assign each the average of the ranks they share. Let R_i denote the sum of the ranks, n_i the number of observations, and $\bar{R}_i = R_i/n_i$ the average rank in sample i. Calculate the test statistic as

$$\text{Test statistic} = \frac{12}{N(N+1)} \sum_{i=1}^{k} \left(\bar{R}_i - \frac{N+1}{2}\right)^2$$

or, equivalently, as

$$\text{Test statistic} = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{(R_i)^2}{n_i} - 3(N+1)$$

where N is the total sample size, the sum of n_1 through n_k. This test statistic measures how far the individual sample rank averages $\bar{R}_i$ differ from the overall average rank, $(N + 1)/2$. If each $\bar{R}_i$ is close to $(N + 1)/2$, then the test statistic is small, consistent with the null hypothesis. If the $\bar{R}_i$'s are not all close to $(N + 1)/2$, then the test statistic is large, inconsistent with the null hypothesis.

3. Assume that the samples are independent random samples of continuous-type observations. Then under the null hypothesis, the test statistic has the Kruskal–Wallis distribution for sample sizes n_1 through n_k. An approximation to the distribution of the test statistic under the null hypothesis is given by the chi-square distribution with $k - 1$ degrees of freedom.
4. Select a significance level α.
5. Using Table H, find the number c such that $P(KW \geq c) \doteq \alpha$, where KW has the Kruskal–Wallis distribution corresponding to the experimental sample sizes. Alternatively, use Table E to find the number c such that $P(X \leq c) = 1 - \alpha$, where X has the chi-square distribution with $k - 1$ degrees of freedom. In either case, the acceptance region is the interval $[0, c)$; the rejection region is the interval $[c, \infty)$.
6. The decision rule is:

   If test statistic $< c$, say the results are consistent with the null hypothesis that the k population distributions are the same.

   If test statistic $\geq c$, say the results are inconsistent with the null hypothesis, suggesting that the k distributions do not all have the same location.
7. Carry out an experiment satisfying the assumptions in step 3. Calculate the test statistic in step 2. Use the decision rule in step 6 to decide whether the observations are consistent with the null hypothesis.

EXAMPLE 12-2
(continued)

Let's apply the Kruskal–Wallis test procedure to Example 12-2. We want to test the null hypothesis that the distribution of maximal oxygen uptake is the same for female world-class athletes in the three sports. The alternative hypothesis says these three distributions are not all the same; some are shifted away from each other. We assume that we have independent observations and that the women tested are representative of female world-class athletes in their respective sports. We cannot check these assumptions without additional information about the experiment. We also assume that the three distributions are the same except possibly for differences in location. This assumption does not seem unreasonable from the plots in Figure 12-3.

Implicit assumptions are that we can compare the measurements of maximal oxygen uptake across sports and that these measures truly reflect aerobic power in these women. (A treadmill test may not provide a good measure of aerobic power in swimmers, for example.) We have no way of checking these assumptions from the information provided. We will proceed in our analysis with caution.

We will use significance level .05. Let *KW* denote a random variable having the Kruskal–Wallis distribution for sample sizes 2, 3, and 4. From Table H we see that $P(KW \geq 5.4) = .051$, close to .05. The acceptance region is $[0, 5.4)$, the rejection region is $[5.4, \infty)$, and the decision rule is:

If test statistic < 5.4, say the results are consistent with the null hypothesis that the three distributions are the same.

If test statistic ≥ 5.4, say the results are inconsistent with the null hypothesis, suggesting that the three distributions have different locations (and different medians).

We calculate the test statistic as shown in Table 12-5. We use the second of the two formulas, because it is easier for hand calculations. The test statistic equals 6.444, inconsistent with the null hypothesis, and the *p*-value, $P(KW \geq 6.444$ when H_0 is true), is between .005 and .011. There appear to be differences in maximal oxygen uptake (as measured in this experiment) among female world-class athletes across the three sports.

Suppose we had used the chi-square approximation. Looking in Table E for $3 - 1 = 2$ degrees of freedom, we see that $P(X \leq 5.99) = .95$. Therefore, the cutoff for our acceptance and rejection regions is 5.99 (compared with 5.4

TABLE 12-5 Steps in calculating the Kruskal–Wallis statistic for Example 12-2

Basketball players		Cross-country skiers		Speed skaters	
Value	Rank	Value	Rank	Value	Rank
42.3	1	56.9	6	46.1	3
42.9	2	58.1	7	52.0	5
49.6	4	61.5	8		
		68.2	9		
$n_1 = 3$		$n_2 = 4$		$n_3 = 2$	
$R_1 = 7$		$R_2 = 30$		$R_3 = 8$	
$\bar{R}_1 = 2.333$		$\bar{R}_2 = 7.5$		$\bar{R}_3 = 4$	

$$\text{Test statistic} = \frac{12}{9(9+1)}\left(\frac{7^2}{3} + \frac{30^2}{4} + \frac{8^2}{2}\right) - 3(9+1) = 6.444$$

using the exact Kruskal–Wallis distribution). We still say our results are inconsistent with the null hypothesis; the approximate p-value is between .025 and .05.

To get a feel for the differences among the three sports, we can calculate a confidence interval for the difference between each pair of medians. We will use the Wilcoxon–Mann–Whitney procedure to calculate each confidence interval (see Section 11-4). The Bonferroni method gives a lower bound for the overall confidence level of these intervals taken together. There are $\binom{3}{2} = 3$ ways to compare our three groups two at a time, so we will calculate three separate confidence intervals.

First let's calculate a confidence interval for the difference between medians of maximal oxygen uptake for cross-country skiers and basketball players. We find the $4 \times 3 = 12$ differences between values for skiers and for basketball players. The smallest difference is $56.9 - 49.6 = 7.3$ and the largest difference is $68.2 - 42.3 = 25.9$. From Table G we know that $P(W \leq 0) = .029$, where W is a random variable having the Wilcoxon–Mann–Whitney distribution for sample sizes 3 and 4. The interval (7.3, 25.9) has confidence level $1 - 2(.029) = .942$.

Similarly, (4.9, 22.1) is an interval estimate for the difference between medians of maximal oxygen uptake for female cross-country skiers and speed skaters, with confidence level .866. An interval estimate for the difference between medians of maximal oxygen uptake for female speed skaters and basketball players is $(-3.5, 9.7)$, with confidence level .800.

Using the Bonferroni method, we see that the overall confidence level for these three intervals taken together is greater than or equal to $1 - (1 - .942) - (1 - .866) - (1 - .800) = .608$, or about 61%. This is not very large, but it is the best we can do with such small sample sizes.

The results of our multiple comparisons are summarized in Table 12-6. Zero is not in the first two intervals; median maximal oxygen uptake seems to be greater for the cross-country skiers than for the other two groups of athletes. Zero is in the third interval, so based on these observations we cannot say there is any difference between basketball players and speed skaters with respect to maximal oxygen capacity. This agrees with what we observe in

TABLE 12-6 Nonparametric multiple comparisons for Example 12-2

Confidence interval for	Confidence interval	Individual confidence level
$M_C - M_B$	$(D_1, D_{12}) = (7.3, 25.9)$	.942
$M_C - M_S$	$(D_1, D_8) = (4.9, 22.1)$	.866
$M_S - M_B$	$(D_1, D_6) = (-3.5, 9.7)$	.800

The overall confidence level is greater than or equal to $1 - (.058 + .134 + .200) = .608$.

Note: The subscripts B, C, and S refer to the basketball players, cross-country skiers, and speed skaters, respectively.

Figure 12-3. The distributions for the basketball players and the speed skaters overlap. The distribution for the cross-country skiers is shifted toward larger values, not overlapping the other two sample distributions at all.

In Section 12-4, we discuss the classical, parametric, analysis of randomized block experiments.

12-4 Parametric Analysis of a Randomized Block Experiment

A randomized block design is an extension to several treatments of the paired-sample design we discussed in Section 11-6. We will consider the simplest randomized block design: the number of experimental units in a block equals the number of treatments. Within each block, experimental units are similar with respect to factors that could affect the outcome of the experiment. We randomly assign treatments to experimental units within each block, one unit per treatment. If there are no differences among treatment effects, we expect similar responses from experimental units within a block. If there are differences among treatment effects, we hope the randomized block design will help us see those differences.

> In a **randomized block experiment,** experimental units within a block are similar with respect to factors that could affect the response. In the simplest design, the number of experimental units in a block equals the number of treatments. The treatments are randomly assigned to experimental units within a block.

Let's consider an example.

EXAMPLE 12-3 Are there differences among thermometers in determining melting points? To address this question, three technicians used each of four thermometers to measure the melting point of Hydroquinone (Duncan, 1974, page 632; from Wernimont, 1947, page 8). The recorded melting points (in °C) are shown below.

	Technician			Thermometer
Thermometer	1	2	3	average
1	174.0	173.0	173.5	173.500
2	173.0	172.0	173.0	172.667
3	171.5	171.0	173.0	171.833
4	173.5	171.0	172.5	172.333
Technician average	173.000	171.750	173.000	

The investigators wanted to assess differences among thermometers. However, they were aware that different technicians might obtain different

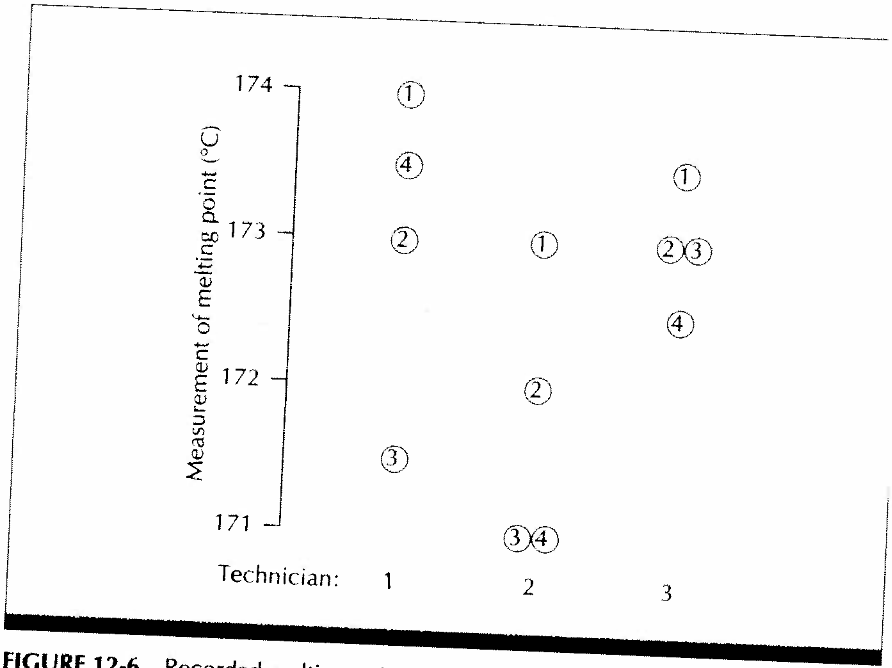

FIGURE 12-6 Recorded melting point of Hydroquinone plotted by technician. Thermometer numbers are shown in the dots.

results, even when using the same thermometer. Therefore, it made sense to have the same technician use each thermometer. Since three technicians were available, the researchers had each of them use each thermometer to determine the melting point of Hydroquinone.

In this experiment, each technician is a block. The idea of blocking in this way is that if there really are no differences among thermometers in measuring melting points, then a single technician should get similar results with each thermometer. The investigators want to control extraneous variation caused by differences among technicians. What suggestions do you have for carrying out this experiment? Should each technician use the thermometers in the same order? Should the first technician make all of his or her measurements, then the second technician, and then the third? Make suggestions regarding these design considerations and any others you can think of, in order to control extraneous sources of variation and make comparisons of thermometers valid.

Plots of the melting point measurements are shown in Figures 12-6 and 12-7. Figure 12-6 shows plots of the readings by technician. The thermometer numbers are shown in the dots. Readings are plotted by thermometer in Figure 12-7. The technician numbers are shown within the dots. What do these plots suggest about differences among thermometers and differences among technicians in determining the melting point of Hydroquinone?

We want to test for differences among thermometers. We will use the classical analysis of a randomized block experiment to assess these differences.

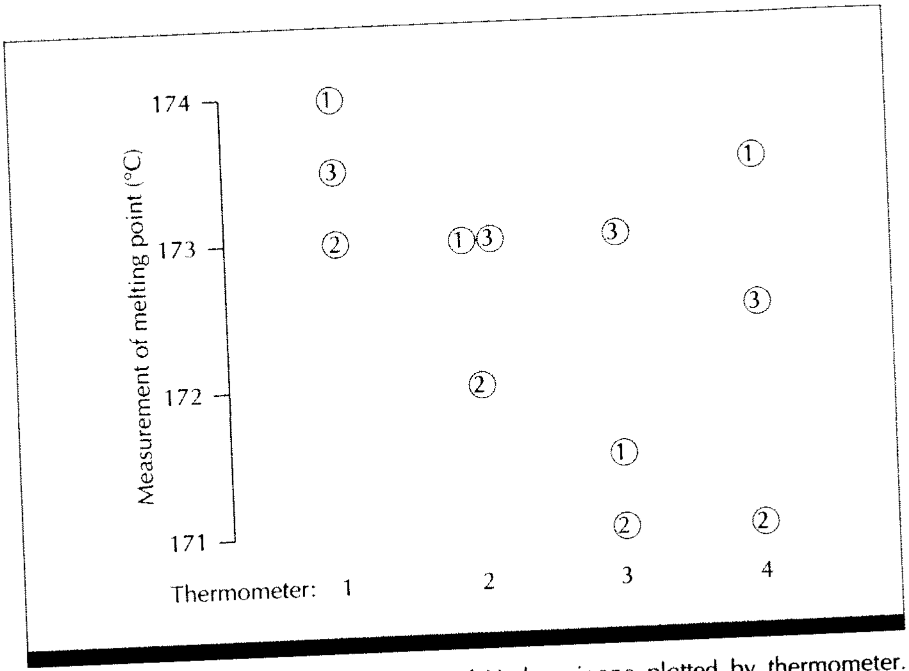

FIGURE 12-7 Recorded melting point of Hydroquinone plotted by thermometer. Technician numbers are shown in the dots.

This classical analysis allows us to test for differences among technicians (blocks) as well. Let's outline the classical, parametric, approach to analyzing a randomized block experiment, and then apply it to this example.

Classical Analysis of a Randomized Block Experiment

To outline the parametric analysis of a randomized block experiment, we need the notation in Table 12-7. Suppose there are b blocks, with k experimental units per block. The number of treatments equals k. Y_{ij} denotes the response of the experimental unit in block j receiving treatment i, $\overline{T}_i$ denotes the average response of the b experimental units receiving treatment i, and $\overline{B}_j$ represents the average response of the k experimental units in block j. The average of all $k \times b$ observations is denoted by $\overline{Y}$. The treatment mean square s_T^2 defined in Table 12-7 is a measure of random variation, plus differences among the means for the k treatments. The block mean square s_B^2 measures random variation, plus differences among the mean responses for the b blocks. The residual mean square s_r^2 is a measure of random variation among observations or responses in the experiment.

To assess differences among treatments on mean response, we compare s_T^2 and s_r^2. If there are really no differences among treatments, then s_T^2 and s_r^2 each estimate random variation among observations in the experiment, so these two variance estimates should be similar in magnitude. If there are dif-

TABLE 12-7 Notation for parametric analysis of a randomized block experiment. Y_{ij} denotes the response of the experimental unit in block j receiving treatment i.

Treatment	Block 1	2	⋯	b	Treatment average
1	Y_{11}	Y_{12}	⋯	Y_{1b}	$\bar{T}_1$
2	Y_{21}	Y_{22}	⋯	Y_{2b}	$\bar{T}_2$
.	.	.	⋯	.	.
.	.	.	⋯	.	.
.	.	.	⋯	.	.
k	Y_{k1}	Y_{k2}	⋯	Y_{kb}	$\bar{T}_k$
Block average	$\bar{B}_1$	$\bar{B}_2$	⋯	$\bar{B}_k$	$\bar{Y}$

$$s_T^2 = \text{Treatment mean square} = \frac{b}{k-1}\sum_{i=1}^{k}(\bar{T}_i - \bar{Y})^2$$

$$s_B^2 = \text{Block mean square} = \frac{k}{b-1}\sum_{j=1}^{b}(\bar{B}_j - \bar{Y})^2$$

$$s_r^2 = \text{Residual mean square} = \frac{1}{(k-1)(b-1)}\sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \bar{T}_i - \bar{B}_j + \bar{Y})^2$$

ferences among treatments, then s_r^2 still estimates random variation, but s_T^2 estimates random variation plus a measure of differences among the treatment means. Therefore, in the case of treatment differences, we expect s_T^2 to be larger than s_r^2.

Similarly, to assess differences among blocks, we compare s_B^2 and s_r^2. If there really are no differences in average responses among blocks, s_B^2 and s_r^2 should be similar in magnitude, since both then estimate random variation among observations in the experiment. If there are differences among blocks, we expect s_B^2 to be larger than s_r^2, since s_B^2 then estimates random variation plus differences among blocks.

With these ideas in mind, we can outline the significance level approach to the classical analysis of a randomized block experiment.

The significance level approach to classical analysis of a randomized block experiment

1. The hypotheses about treatment differences are:

   H_0: The k treatments all have the same average effect on response.
   H_a: The average effect on response is not the same for all k treatments.

   The hypotheses about block differences can be stated as:

   H_0^*: The average response is the same for all b blocks.
   H_a^*: The average response is not the same for all b blocks.

2. To test the hypotheses about treatment effects, we use the test statistic

$$\text{Test statistic(T)} = \frac{s_T^2}{s_r^2}$$

To test the hypotheses about block effects, we use the test statistic

$$\text{Test statistic(B)} = \frac{s_B^2}{s_r^2}$$

3. Assume that the $k \times b$ observations are all independent, from Gaussian distributions. These distributions have the same variance σ^2. The means may differ, depending on treatment and block. We also assume that the relative treatment effects are the same for each block.

   Under the null hypothesis of no treatment differences, test statistic(T) has the F distribution with $k - 1$ numerator degrees of freedom and $(k - 1)(b - 1)$ denominator degrees of freedom. Small values of test statistic(T), near 1, are consistent with the null hypothesis of no differences among treatments on average. Large values of test statistic(T) are inconsistent with this null hypothesis.

   Under the null hypothesis of no block differences, test statistic(B) has the F distribution with $b - 1$ numerator degrees of freedom and $(k - 1)(b - 1)$ denominator degrees of freedom. Small values of test statistic(B), near 1, are consistent with the null hypothesis of no differences in average response among blocks. Large values of test statistic(B) are inconsistent with this null hypothesis.
4. Select significance level α_1 for the first test of hypotheses, α_2 for the second test.
5. For the test of treatment effects, find the number c_1 from Table D such that $P(F_1 \leq c_1) = 1 - \alpha_1$. Here, F_1 denotes a random variable having the $F(k - 1, (k - 1)(b - 1))$ distribution. The acceptance region is the interval $[0, c_1)$; the rejection region is the interval $[c_1, \infty)$.

   For the test of block effects, find the number c_2 from Table D such that $P(F_2 \leq c_2) = 1 - \alpha_2$. Here, F_2 denotes a random variable having the $F(b - 1, (k - 1)(b - 1))$ distribution. The acceptance region is the interval $[0, c_2)$; the rejection region is the interval $[c_2, \infty)$.
6. To test for treatment differences, the decision rule is:

   If test statistic(T) $< c_1$, say the results are consistent with the null hypothesis of no treatment differences in average response.

   If test statistic(T) $\geq c_1$, say the results are inconsistent with this null hypothesis, suggesting there are treatment differences in average response.

   To test for block differences, the decision rule is:

   If test statistic(B) $< c_2$, say the results are consistent with the null hypothesis of no block differences in average response.

   If test statistic(B) $\geq c_2$, say the results are inconsistent with this null hypothesis, suggesting there are block differences in average response.
7. Carry out an experiment that satisfies the assumptions in step 3. Calculate the test statistics in step 2. Use the decision rules in step 6 to decide whether

there seem to be differences among treatments and differences among blocks. Draw conclusions based on the experimental results.

EXAMPLE 12-3
(continued)

Let's use this parametric approach to analyze the results of the experiment in Example 12-3. The four thermometers represent the treatments in this experiment, while the three technicians represent the blocks. A response is the melting point determination a technician makes with a particular thermometer. The hypotheses about thermometer (treatment) differences are:

H_0: On average, the four thermometers give the same reading for the melting point of Hydroquinone.

H_a: The four thermometers do not give the same reading on average.

The hypotheses about technician (block) differences are:

H_0^*: On average, the three technicians get the same reading for the melting point of Hydroquinone.

H_a^*: The three technicians do not get the same reading on average.

Note that we presented the results of the experiment in Example 12-3 in the format shown in Table 12-7. The only statistic not shown there is the average of all 12 observations, $\bar{Y} = 172.583$.

We assume that the 12 observations are all independent. We cannot check this assumption without more information on how the experiment was conducted. What suggestions would you make about the conduct of the experiment in order to ensure independence of observations?

We also assume that the observations come from Gaussian distributions with the same variance. One way to check this assumption is through plots of *residuals*. Recall that a residual is the difference between an observation and a summary, or predicted value, or estimate of the mean of the observation.

For a randomized block design, the predicted value for observation Y_{ij} (treatment i, block j) is the estimated mean value $\bar{T}_i + \bar{B}_j - \bar{Y}$ based on our model assumptions. The residual for that observation is then $Y_{ij} - \bar{T}_i - \bar{B}_j + \bar{Y}$.

A **residual** is the difference between an observation and an estimate of its expected value. In the simplest randomized block design, a residual has the form $Y_{ij} - \bar{T}_i - \bar{B}_j + \bar{Y}$, where Y_{ij} denotes the observation corresponding to treatment i and block j, $\bar{T}_i$ is the average of all observations for treatment i, $\bar{B}_j$ is the average of observations in block j, and $\bar{Y}$ is the average of all the observations.

The residuals represent what is left over after we fit our randomized block probability model; they are like noise. If all the model assumptions hold, the residuals should (roughly) represent independent observations from the Gaussian distribution with mean 0 and variance σ^2. Note that in our definition of s_r^2 at the bottom of Table 12-7, we add up the squared residuals in order to calculate this variance estimate. This is why we often call s_r^2 the *residual mean square*. The residuals for Example 12-3 are shown in Table 12-8.

TABLE 12-8 Residuals for the randomized block probability model in Example 12-3

	Technician		
Thermometer	1	2	3
1	.08	.33	−.42
2	−.08	.17	−.08
3	−.75	.00	.75
4	.75	−.50	−.25

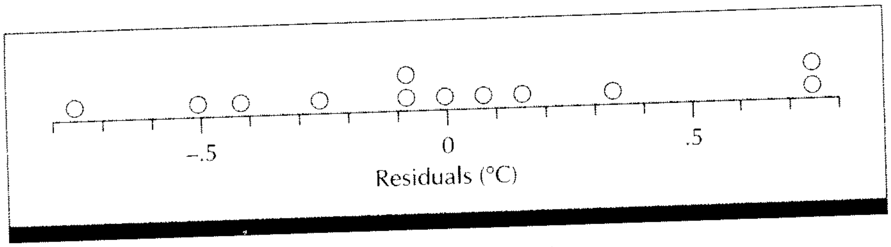

FIGURE 12-8 Dot plot of residuals in Example 12-3

A dot plot of the residuals is shown in Figure 12-8. The plot gives us no reason to doubt the Gaussian assumption. Figure 12-9 shows a plot of residuals by thermometer; we see that the variation among residuals for thermometer 2 is less than for the other three thermometers. Figure 12-10 is a plot of residuals by technician; the variation among residuals is somewhat smaller for technician 2 than for the other two technicians. These differences in variation are not extreme enough to make us avoid the classical analysis of this randomized block experiment. (The analysis is fairly robust to small deviations from the equal-variance assumption, meaning that as long as variances are not too different, significance levels and confidence levels are close to the levels we choose.)

For our analysis, we also assume that the relative thermometer effects are the same for each technician. Consider the dot plots in Figure 12-6. If the assumption held, we would expect the order of the thermometers to be the same for each technician. In fact, all three technicians obtained the highest readings with thermometer 1. But the order varies across technicians for the other three thermometers. We do not have strong evidence for or against the assumption that the relative thermometer effects are the same for all three technicians. We will proceed with our analysis, with our usual caution in interpretations.

If our model assumptions all hold, test statistic(T) has the $F(3, 6)$ distribution under the null hypothesis of no treatment differences. Test statistic(B) has the $F(2, 6)$ distribution under the null hypothesis of no technician differences.

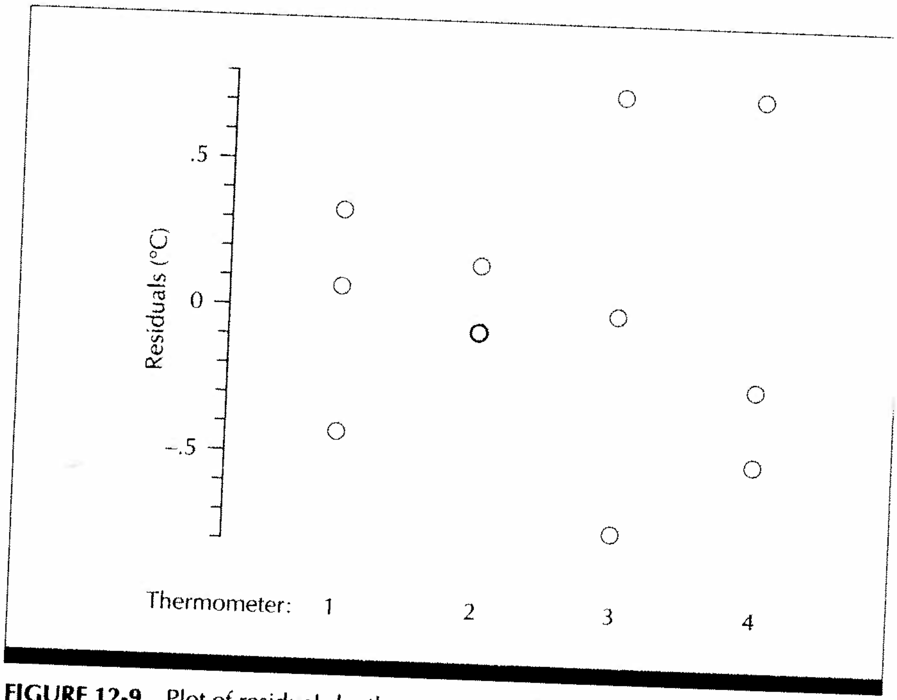

FIGURE 12-9 Plot of residuals by thermometer in Example 12-3

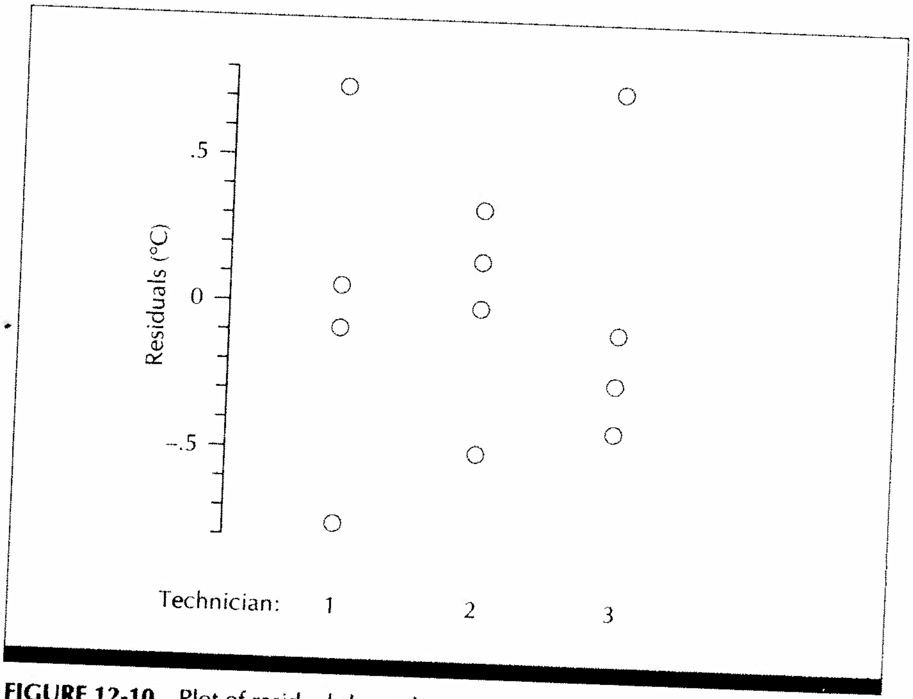

FIGURE 12-10 Plot of residuals by technician in Example 12-3

TABLE 12-9 Calculations for parametric analysis of the randomized block experiment in Example 12-3

k = Number of treatments = 4 $\quad\quad$ $\bar{Y}$ = Overall sample mean = 172.583

b = Number of blocks = 3

$$s_T^2 = \frac{3}{4-1}[(173.500 - 172.583)^2 + (172.667 - 172.583)^2 + (171.833 - 172.583)^2 + (172.333 - 172.583)^2]$$

$$= 1.47$$

$$s_B^2 = \frac{4}{3-1}[(173.000 - 172.583)^2 + (171.750 - 172.583)^2 + (173.000 - 172.583)^2]$$

$$= 2.08$$

$$s_r^2 = \frac{1}{(4-1)(3-1)}[(.08)^2 + (.33)^2 + (-.42)^2 + (-.08)^2 + (.17)^2 + (-.08)^2 + (-.75)^2 + (0)^2 + (.75)^2 + (.75)^2 + (-.50)^2 + (-.25)^2]$$

$$= .39$$

$$\text{Test statistic(T)} = \frac{1.47}{.39} = 3.8 \qquad \text{Test statistic(B)} = \frac{2.08}{.39} = 5.3$$

$k - 1 = 3 \qquad b - 1 = 2 \qquad (k-1)(b-1) = 6$

We will use significance level .10 for both tests of hypotheses. Looking in Table D, we see that if F_1 has the $F(3, 6)$ distribution, then $P(F_1 \leq 3.29) = .90$. The acceptance region for the test about thermometer effects is $[0, 3.29)$, the rejection region is $[3.29, \infty)$, and the decision rule is:

If test statistic(T) < 3.29, say the results are consistent with the null hypothesis that the thermometers give the same reading on average.

If test statistic(T) ≥ 3.29, say the results are inconsistent with this null hypothesis, suggesting that the thermometers do not give the same reading on average.

If F_2 has the $F(2, 6)$ distribution, then $P(F_2 \leq 3.46) = .90$. The acceptance region for the test about technician effects is $[0, 3.46)$, the rejection region is $[3.46, \infty)$, and the decision rule is:

If test statistic(B) < 3.46, say the results are consistent with the null hypothesis that the technicians get the same reading on average.

If test statistic(B) ≥ 3.46, say the results are inconsistent with this null hypothesis, suggesting that the technicians do not get the same reading on average.

The calculations for our analysis are outlined in Table 12-9.

Test statistic(T) equals 3.8, which is inconsistent with the null hypothesis that there are no differences among thermometers on average, at the .10 significance level. The p-value is between .05 and .10.

Test statistic(B) equals 5.3, which is inconsistent with the null hypothes that there are no differences among technicians on average, at the .10 signi cance level. The p-value is a little less than .05.

Our analysis suggests that the thermometers gave somewhat differe readings of the melting point of Hydroquinone on average. From Figures 12- and 12-7 we see that the highest readings were with thermometer 1. Thermon eters 3 and 4 gave lower readings on average and had greater spread in th readings. Thermometer 2 gave intermediate readings.

The analysis also suggests the the technicians got somewhat differen results on average. The most striking difference is that technician 2 got lowe readings than technicians 1 and 3, for each thermometer (this shows up clearl in Figure 12-7).

The p-value for the test of thermometer differences was between .05 an .10, which we might consider borderline statistical significance. Whether o not we consider the results of this experiment of practical importance depends on the accuracy (closeness to the correct value) and precision (lack of variation, or repeatability) required when melting points are determined in practical situations.

Suppose we had ignored the technicians in our analysis. If we had done a one-way analysis of variance, with three readings for each thermometer, we would have found no significant difference among thermometers, with p-value = .2 (Exercise 12-18). The randomized block design was useful in this experiment. Because there were differences among technicians, blocking helped us see differences among thermometers. Also, from a quality control point of view, it is useful to see that different technicians can get different readings on average.

A general form of analysis of variance table for the simplest randomized block design is shown in Table 12-10. The analysis of variance table for Ex-

TABLE 12-10 Analysis of variance table for a randomized block experiment (number of treatments equals the size of each block)

Source of variation	Sum of squares	Degrees of freedom	Mean square	Test statistic
Treatments	$b\sum_{i=1}^{k}(\bar{T}_i - \bar{Y})^2$	$k - 1$	s_T^2	$\frac{s_T^2}{s_r^2}$
Blocks	$k\sum_{j=1}^{b}(\bar{B}_j - \bar{Y})^2$	$b - 1$	s_B^2	$\frac{s_B^2}{s_r^2}$
Residuals	$\sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \bar{T}_i - \bar{B}_j + \bar{Y})^2$	$(k - 1)(b - 1)$	s_r^2	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \bar{Y})^2$	$kb - 1$		

TABLE 12-11 Analysis of variance table for the randomized block experiment in Example 12-3

Source of variation	Sum of squares	Degrees of freedom	Mean square	Test statistic	p-value
Thermometers	4.42	3	1.47	3.8	.08
Technicians	4.17	2	2.08	5.3	.05
Residuals	2.33	6	.39		
Total	10.92	11			

ample 12-3 is shown in Table 12-11, with an added column showing the p-value for each test statistic.

In Section 12-5, we consider a nonparametric method, called Friedman's test, for analyzing a randomized block experiment.

12-5 Nonparametric Analysis of a Randomized Block Experiment: Friedman's Test

Before considering an example, we will first outline a nonparametric procedure for analysis of a randomized block experiment, *Friedman's test*. This is a test of treatment differences (not block differences). Suppose we have b blocks, with k experimental units per block. There are k treatments, one treatment per experimental unit in each block.

The significance level approach to nonparametric analysis of a randomized block experiment, using Friedman's test

1. The hypotheses are:

   H_0: The treatments have the same average effect on response.

   H_a: The treatments do not all have the same average effect on response.

2. Rank the k observations within each block. The smallest observation gets rank 1 and the largest gets rank k. Tied observations get the average of the ranks they share. Let $\bar{R}_1$ denote the average of the ranks for treatment 1. Let $\bar{R}_2$ denote the average of the ranks for treatment 2, and so on. $\bar{R}_k$ is the average of the ranks for treatment k. The overall average rank is $(k + 1)/2$. The test statistic is

$$\text{Test statistic} = \frac{12b}{k(k + 1)} \sum_{i=1}^{k} \left(\bar{R}_i - \frac{k + 1}{2} \right)^2$$

3. We assume that the $k \times b$ observations are all independent, from distributions with similar shape and variation. We also assume that the relative treatment effects are the same for each block. Then under the null hypothesis of no treatment differences, the test statistic has approximately the chi-square distribution with $k - 1$ degrees of freedom. Small values of the test statistic

are consistent with the null hypothesis, while large values are inconsisten with the null hypothesis.

4. Select significance level α.
5. Find the number c in Table E such that $P(X \leq c) = 1 - \alpha$, where X has the chi-square distribution with $k - 1$ degrees of freedom. The acceptance region is the interval $[0, c)$. The rejection region is the interval $[c, \infty)$.
6. The decision rule is:

   If test statistic $< c$, say the results are consistent with the null hypothesis of no treatment differences.

   If test statistic $\geq c$, say the results are inconsistent with the null hypothesis, suggesting that there are treatment differences.

7. Carry out an experiment satisfying the assumptions in step 3. Calculate the test statistic in step 2. Use the decision rule in step 6 to decide whether the results are consistent with the null hypothesis. Draw conclusions based on the experimental results.

Let's apply Friedman's test to the following example.

EXAMPLE 12-4 Investigators wanted to compare the effects of three anesthetics upon plasma epinephrine concentration in dogs. They measured plasma epinephrine concentration (in nanograms per milliliter) for ten dogs while under each of these three anesthetics: isofluorane, halothane, and cyclopropane. The measurements are listed below (Rice, 1988, page 431; from Perry, Van Dyke, and Theye, 1974).

	Dog									
Anesthetic	1	2	3	4	5	6	7	8	9	10
Isofluorane	.28	.51	1.00	.39	.29	.36	.32	.69	.17	.33
Halothane	.30	.39	.63	.68	.38	.21	.88	.39	.51	.32
Cyclopropane	1.07	1.35	.69	.28	1.24	1.53	.49	.56	1.02	.30

What suggestions would you make for the design of this experiment? How would you seek to reduce the effects of extraneous factors? Should each dog receive the anesthetics in the same order? Would you worry about carryover effects of anesthetics from one treatment period to the next? What other concerns would you have and how would your experimental design address those concerns?

Plots of the observations are shown in Figures 12-11 and 12-12. Figure 12-11 shows a plot of the ten measurements of plasma epinephrine concentration, for each of the three anesthetics. Responses under the three anesthetics are plotted for each dog in Figure 12-12. What do these plots suggest about differences among anesthetics and differences among dogs with respect to plasma epinephrine concentration?

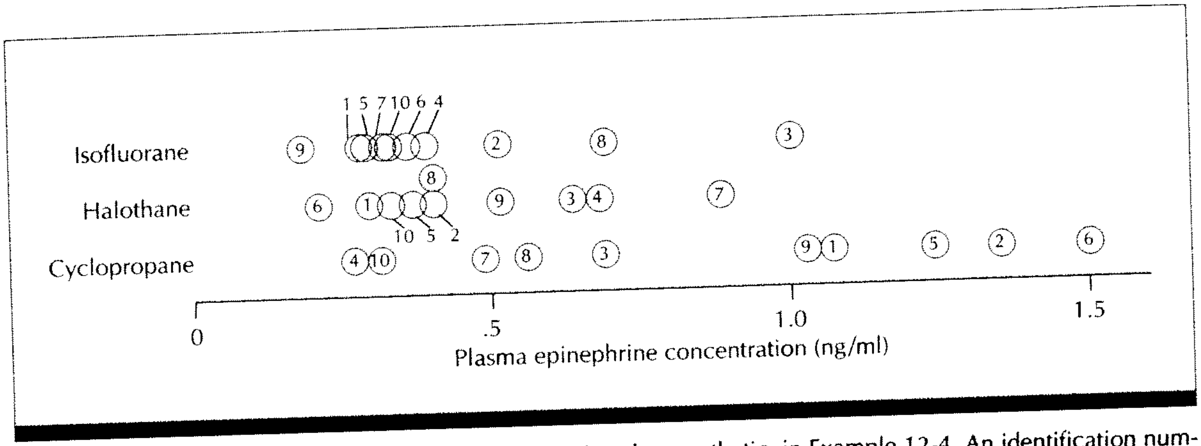

FIGURE 12-11 Plot of plasma epinephrine concentrations by anesthetic, in Example 12-4. An identification number denotes each dog's responses.

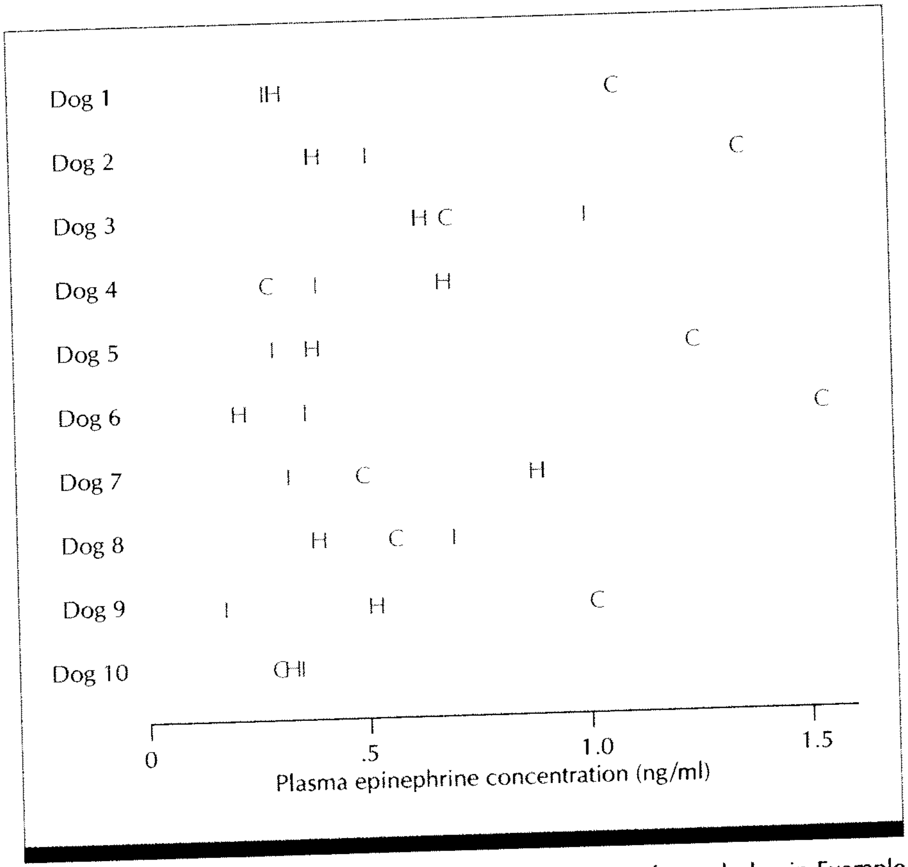

FIGURE 12-12 Plot of plasma epinephrine concentrations for each dog in Example 12-4. The letters I, H, and C denote the response for a dog while under isofluorane, halothane, and cyclopropane, respectively.

In this experiment, the anesthetics are the treatments and the dogs ar blocks. We want to test for differences in average effects of the anesthetics o plasma epinephrine concentrations, with hypotheses:

H_0: The three anesthetics have the same average effect on plasma epinephrine concentration.

H_a: The three anesthetics do not have the same average effect on plasma epinephrine concentration.

We must assume that the 30 observations are independent. We cannot check this assumption without more information on how the experiment was carried out. What suggestions would you make in order to ensure independence?

We also assume that the relative effects of the anesthetics are the same for the ten dogs. Looking at Figure 12-12, we see that this assumption is badly violated. For five dogs (dogs 1, 2, 5, 6, and 9), plasma epinephrine concentrations were much higher under cyclopropane than under the other two anesthetics. These are the five largest values plotted for cyclopropane in Figure 12-11. For the other five dogs, there are smaller differences among the anesthetics. Also, cyclopropane did not result in the largest values for these dogs. Clearly, the relative effects of the three anesthetics are not the same for all ten dogs.

For a valid analysis, we must assume that the relative differences among anesthetics (treatments) are the same for each dog (block). Our plots show us that this assumption is not reasonable. For a moment we will ignore this problem and go through the mechanics of the procedure; then we will discuss our results in terms of this violation of assumptions. [This example has appeared in a number of references as a randomized block experiment requiring a standard analysis. In fact, as our plots show, a major assumption of both parametric and nonparametric analysis of a standard (unreplicated) randomized block experiment is badly violated.]

If all assumptions for Friedman's test did hold, then under the null hypothesis the test statistic would have approximately the chi-square distribution with $3 - 1 = 2$ degrees of freedom. We will use significance level .10. From Table E, we find that $P(X \leq 4.61) = .90$, where X has the chi-square distribution with 2 degrees of freedom. The acceptance region is $[0, 4.61)$, the rejection region is $[4.61, \infty)$, and the decision rule is:

If test statistic < 4.61, say the results are consistent with the null hypothesis of no difference in mean plasma epinephrine concentration among the three anesthetics.

If test statistic ≥ 4.61, say the results are inconsistent with the null hypothesis, suggesting that there are differences in mean plasma epinephrine concentrations among the three anesthetics.

The calculations we need for Friedman's test are outlined in Table 12-12.

The test statistic equals 1.4, consistent with the null hypothesis of no difference among anesthetics, at the .10 significance level. The approximate p-value, based on the chi-square distribution with 2 degrees of freedom, is about .5.

TABLE 12-12 Calculations for Friedman's test in Example 12-4. Observations for each dog are ranked from 1 to 3.

	Dog										Sum of
Anesthetic	1	2	3	4	5	6	7	8	9	10	ranks
Isofluorane	1	2	3	2	1	2	1	3	1	3	19
Halothane	2	1	1	3	2	1	3	1	2	2	18
Cyclopropane	3	3	2	1	3	3	2	2	3	1	23

$\bar{R}_1 = \frac{19}{10} = 1.9 \qquad \bar{R}_2 = \frac{18}{10} = 1.8 \qquad \bar{R}_3 = \frac{23}{10} = 2.3$

Overall average rank $= \frac{3 + 1}{2} = 2$

k = Number of treatments = 3 $\qquad b$ = Number of blocks = 10

Test statistic $= \frac{12 \times 10}{3(3 + 1)} [(1.9 - 2)^2 + (1.8 - 2)^2 + (2.3 - 2)^2] = 1.4$

Degrees of freedom $= k - 1 = 2$

If we ignore the violation of our model assumptions, Friedman's test tells us there do not appear to be differences among the anesthetics. This seems to be true for five of the dogs (dogs 3, 4, 7, 8, and 10). However, as we saw in Figure 12-12, dogs 1, 2, 5, 6, and 9 had plasma epinephrine concentrations much higher under cyclopropane than under isofluorane or halothane. Our plots suggest that there may be differences among anesthetics. Also, the relative effects of the anesthetics vary with dogs. This is called an *interaction* effect of anesthetic and dog upon the response. There is no way to account for this interaction in our analysis of this randomized block experiment. In Example 12-4, the plots are much more useful and informative than the formal analysis, which is misleading because not all the assumptions for the analysis are satisfied.

Exercise 12-19 asks you to use Friedman's test to check for thermometer differences in Example 12-3, where the assumptions for the test seem more reasonable than in Example 12-4.

We say we have an *interaction effect* of treatment and block on response if the relative effects of treatments differ for different blocks. The only way to account for such interaction in our analysis is to have larger blocks. Then we assign each treatment to two or more experimental units within each block. We can analyze this larger experiment using two-way analysis of variance for a replicated randomized block experiment. See, for example, Kirk (1982, Chapter 6).

In Chapter 13 we discuss two-way analysis of variance, for a two-factor experiment.

Summary of Chapter 12

The Bonferroni method provides an upper bound on the overall significance level when we make several tests of hypotheses. The method provides a lower bound on the overall confidence level when several confidence intervals are used for multiple comparisons.

In a single-factor experiment (an extension to several samples of the two-sample design), we assume that we have k independent random samples, one from each of k populations. If we assume that the samples come from Gaussian distributions with the same variance, then we can use one-way analysis of variance to test the null hypothesis that the population means are all equal. If we assume that the samples come from distributions with the same shape and variation, but possibly different locations, we can use the Kruskal–Wallis test to test the null hypothesis that the distributions are equal.

In the simplest randomized block design (an extension of the paired-sample design), experimental units within a block are similar with respect to characteristics that might affect the response. The number of experimental units in each block equals the number of treatments. The treatments are randomly assigned to experimental units within a block.

For a parametric analysis of a randomized block experiment, we assume that the observations are all independent, from Gaussian distributions with the same variance; the means may vary depending on treatment and block. We also assume that the relative treatment effects are the same within each block. We can test the null hypothesis that the average effect is the same for all treatments, as well as the null hypothesis that the average response is the same for all blocks.

For a nonparametric analysis of a randomized block experiment, we assume that the observations are independent, from distributions having similar shape and variation. We also assume that the relative treatment effects are the same for each block. Friedman's test assesses the null hypothesis that the average treatment effect is the same for all treatments.

Residual plots are useful for checking whether model assumptions seem reasonable. A residual is the difference between an observation and a summary, predicted value, or estimate of the mean of the observation.

Minitab Appendix for Chapter 12

Finding Probabilities for *F* and Chi-Square Distributions

We introduced the F distributions and the chi-square distributions in Chapter 12. We can use the CDF, PDF, and RANDOM commands with the F and CHISQUARE subcommands, as we have discussed for other distributions. With the F subcommand, we specify numerator degrees of freedom and then denominator degrees of freedom. With the CHISQUARE subcommand, we specify the degrees of freedom. For instance,

```
MTB>   cdf 12.2;
SUBC>  f 4 6.
 12.2000  0.9952
MTB>   cdf 6.5;
SUBC>  chisquare 2.
 6.5000  0.9612
```

Minitab tells us that if the random variable X has the $F(4, 6)$ distribution, then $P(X \leq 12.2) = .9952$. If the random variable Y has the chi-square distribution with 2 degrees of freedom, then $P(Y \leq 6.5) = .9612$.

Performing One-Way Analysis of Variance with the ONEWAY Command

Suppose the data from Example 12-1 are in two columns on our worksheet. Column 1 (named GROUP) contains a code for group: 1 = control, 2 = urea, 3 = potassium nitrate and calcium, 4 = ammonia and ammonium sulphate. Fruit weights are in column 2 (named FRUITWT). We use the ONEWAY command for a parametric one-way analysis of variance:

```
MTB>  oneway 'fruitwt' 'group'
```

The results are shown in Figure M12-1.

If we specify two additional columns at the end of the ONEWAY command, we can save residuals and estimated (or predicted) values of observations. In Example 12-1, the command

```
MTB>  oneway 'fruitwt' 'group' c3 c4
MTB>  name c3 'resid' c4 'predict'
```

produces the same output as shown in Figure M12-1. In addition, Minitab stores the residuals from the fitted model in C3, which we name RESID. Mini-

```
ANALYSIS OF VARIANCE ON fruitwt
SOURCE      DF       SS        MS         F        p
group        3     5980      1993     11.28    0.000
ERROR       38     6717       177
TOTAL       41    12697
                                   INDIVIDUAL 95 PCT CI'S FOR MEAN
                                   BASED ON POOLED STDEV
LEVEL        N     MEAN     STDEV  --+---------+---------+---------+----
    1       10    83.13     18.11     (----*-----)
    2       11    99.32      8.81                (----*-----)
    3       11   109.60     15.17                       (----*----)
    4       10    80.53      8.75  (-----*----)
                                   --+---------+---------+---------+----
POOLED STDEV =    13.30            75        90       105       120
```

FIGURE M12-1 Output from the ONEWAY command in Example 12-1

```
resid   -
        -
        -          *
     25+
        -                                     *                *
        -          *                          *                *
        -     3    *                          *                *
        -     *    2                          *                *
       0+     *    *                          3
        -     3    *                          3                2
        -     2                               *                *
        -          2                          *                2
        -                                                      *
    -25+
        -                                                      *
        -          *
        -
         +---------+---------+---------+---------+---------+------predic
      78.0      84.0      90.0      96.0     102.0     108.0
```

FIGURE M12-2 Scatterplot of residuals versus predicted values in Example 12-1

```
LEVEL      NOBS     MEDIAN  AVE. RANK   Z VALUE
    1         3      42.90        2.3     -2.07
    2         4      59.80        7.5      2.45
    3         2      49.05        4.0     -0.59
OVERALL       9                   5.0

H = 6.444
* NOTE  * ONE OR MORE SMALL SAMPLES
```

FIGURE M12-3 Output from the KRUSKAL-WALLIS command for Example 12-2

tab stores the estimated (or predicted) observations in C4, named PREDICT. We can use these saved values, say in a histogram of residuals or a plot of residuals versus predicted values, to check model assumptions. For instance, the command

```
MTB> plot 'resid' 'predict'
```

produces the scatterplot in Figure M12-2.

We can use the TWOT command and the Bonferroni method to make multiple comparisons, if the results of ONEWAY indicate differences among means.

Carrying Out a Kruskal–Wallis Comparison of Means

The KRUSKAL-WALLIS command carries out the calculations for the Kruskal-Wallis test. Suppose our worksheet contains the data for Example 12-2. Column 1 (named SPORT) contains a code for sport: 1 = basketball, 2 = cross-country skiing, 3 = speed skating. Column 2 (named UPTAKE) contains maximal oxygen uptake. The command

```
MTB> krus 'uptake' 'sport'
```

results in the output in Figure M12-3.

The student edition of Minitab does not print a p-value for the Kruskal–Wallis test statistic. We can look it up in a table for the Kruskal–Wallis distribution. If we think the sample sizes are large enough, we can use the large-sample chi-square approximation. In our example, we might want to compare the test statistic with the chi-square distribution with 2 degrees of freedom (since there are three groups). The test statistic equals 6.444 and the approximate p-value is $P(X \geq 6.444)$, where X has the chi-square distribution with 2 degrees of freedom. We can find this approximate p-value using Minitab as follows:

```
MTB>   cdf 6.444 k1;
SUBC>  chisquare 2.
MTB>   let k2=1-k1
MTB>   print k1 k2
K1     0.960125
K2     0.0398753
```

Minitab prints the cumulative probability K1 = 0.960125 and the approximate p-value K2 = 1 − K1 = 0.0398753 corresponding to the observed value 6.444 of the test statistic. (Recall that the p-value based on the exact Kruskal–Wallis distribution was between .005 and .011.)

If we want to make multiple comparisons based on Mann–Whitney intervals, we must unstack the observations in 'UPTAKE' based on 'SPORT':

```
MTB>   unstack 'uptake' c3-c5;
SUBC>  subscripts 'sport'.
```

C3 contains uptake values for sport 1; C4, for sport 2; C5, for sport 3. Now we can use the MANN-WHITNEY command two columns at a time on C3 through C5. We apply the Bonferroni method to get the overall confidence level for our multiple comparisons.

Analyzing a Randomized Block Experiment with the TWOWAY Command

To analyze the results of a randomized block experiment, as discussed in Section 12-4, we use the TWOWAY command. Consider the experiment in Example 12-3. Suppose column 1 (named THERM) of our worksheet contains thermometer number, column 2 (named TECH) contains technician number, and column 3 (named MELT) contains measured melting points. We can get descriptive statistics for melting point by thermometer and by technician. The command

```
MTB>   describe 'melt';
SUBC>  by 'therm'.
```

gives the output in Figure M12-4, while the command

```
           therm        N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
melt           1        3    173.50    173.50    173.50      0.50      0.29
               2        3    172.67    173.00    172.67      0.58      0.33
               3        3    171.83    171.50    171.83      1.04      0.60
               4        3    172.33    172.50    172.33      1.26      0.73

           therm      MIN       MAX        Q1        Q3
melt           1   173.00    174.00    173.00    174.00
               2   172.00    173.00    172.00    173.00
               3   171.00    173.00    171.00    173.00
               4   171.00    173.50    171.00    173.50
```

FIGURE M12-4 Descriptive statistics for MELT by thermometer

```
            tech        N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
melt           1        4    173.00    173.25    173.00      1.08      0.54
               2        4    171.75    171.50    171.75      0.96      0.48
               3        4    173.00    173.00    173.00      0.41      0.20

            tech      MIN       MAX        Q1        Q3
melt           1   171.50    174.00    171.87    173.88
               2   171.00    173.00    171.00    172.75
               3   172.50    173.50    172.63    173.37
```

FIGURE M12-5 Descriptive statistics for MELT by technician

```
ANALYSIS OF VARIANCE  melt

SOURCE         DF         SS         MS
therm           3      4.417      1.472
tech            2      4.167      2.083
ERROR           6      2.333      0.389
TOTAL          11     10.917
```

FIGURE M12-6 Analysis of variance table for the randomized block experiment in Example 12-3

```
MTB>    describe 'melt';
SUBC>   by 'tech'.
```

gives the output in Figure M12-5.

The TWOWAY command

```
MTB>    twoway 'melt' 'therm' 'tech'
```

gives the output in Figure M12-6.

Notice that the Minitab output for TWOWAY does not include test statistics or *p*-values. To test for thermometer differences, we calculate the test statistic from the analysis of variance table and use Minitab to calculate the *p*-value:

```
MTB>    let k1=1.472/0.389
MTB>    cdf k1 k2;
SUBC>   f 3 6.
MTB>    let k2=1-k2
MTB>    print k1 k2
K1      3.78406
K2      0.0777537
```

Minitab prints the value of the test statistic K1 = 3.78406 and the p-value K2 = 0.0777537. We go through similar steps to test for technician differences:

```
MTB>    let k3=2.083/0.389
MTB>    cdf k3 k4;
SUBC>   f 2 6.
MTB>    let k4=1-k4
MTB>    print k3 k4
K3      5.35476
K4      0.0462980
```

Minitab prints the value of the test statistic K3 = 5.35476 and the p-value K4 = 0.0462980. Because of round-off error, these values do not exactly equal what we found in Example 12-3.

As with ONEWAY, we can specify two extra columns as part of the TWOWAY command to save residuals and predicted values. The command

```
MTB>    twoway 'melt' 'therm' 'tech' c4 c5
```

produces the same output as in Figure M12-6, and saves residuals in C4 and predicted values in C5. We can use these saved values in plots to check model assumptions.

Using Minitab to Carry Out Friedman's Test

The student edition of Minitab does not have a procedure for Friedman's test. We can calculate the test statistic using Minitab, however. Consider the data in Example 12-4. Suppose column 1 (named DOG) of our worksheet contains dog number. Column 2 (named ANES) contains a code for anesthetic: 1 = isofluorane, 2 = halothane, 3 = cyclopropane. Column 3 (named RESPONSE) contains plasma epinephrine concentration. We want to unstack the RESPONSE column into ten columns, one for each dog:

```
MTB>    unstack 'response' c11-c20;
SUBC>   subscript 'dog'.
```

C11 contains the three observations for dog 1, C12 contains the three observations for dog 2, and so on. If we print columns 11–20, we get the output in Figure M12-7.

```
ROW    C11    C12    C13    C14    C15    C16    C17    C18    C19    C20

  1   0.28   0.51   1.00   0.39   0.29   0.36   0.32   0.69   0.17   0.33
  2   0.30   0.39   0.63   0.68   0.38   0.21   0.88   0.39   0.51   0.32
  3   1.07   1.35   0.69   0.28   1.24   1.53   0.49   0.56   1.02   0.30
```

FIGURE M12-7 Contents of columns 11–20

```
ROW   C11   C12   C13   C14   C15   C16   C17   C18   C19   C20

  1     1     2     3     2     1     2     1     3     1     3
  2     2     1     1     3     2     1     3     1     2     2
  3     3     3     2     1     3     3     2     2     3     1
```

FIGURE M12-8 Columns 11–20, after replacing observations by ranks within columns

Now we want to rank the observations for each dog (block). We will replace the observations with the ranks in columns 11–20:

```
MTB>   rank c11 c11
        .
        .
        .
MTB>   rank c20 c20
```

where the dots indicate that we type this command for all ten columns. If we now print columns 11–20, we get the output in Figure M12-8.

We need the sum of the ranks for each treatment (anesthetic). We will put these sums in column 21:

```
MTB>   rsum c11-c20 c21
```

To find average ranks for the three rows, we divide column 21 by 10:

```
MTB>   let c21=c21/10
```

The overall average rank is (3 + 1)/2, where 3 is the number of treatments. We subtract this value from each element of column 21:

```
MTB>   let c21=c21-(3+1)/2
```

Then we square each element of column 21:

```
MTB>   let c21=c21**2
```

We sum the elements of column 21:

```
MTB>   sum c21 k1
```

Since the number of blocks is 10 and the number of treatments is 3, the test statistic is calculated as

```
MTB>   let k2=k1*12*10/(3*(3+1))
```

We compare this test statistic with the chi-square distribution with 2 degrees of freedom, since there are three treatments:

```
MTB>    cdf k2 k3;
SUBC>   chisquare 2.
MTB>    let k3=1-k3
MTB>    print k2 k3
K2      1.40000
K3      0.496585
```

Minitab prints the value K2 = 1.40000 of Friedman's test statistic and the large-sample approximate p-value K3 = 0.496585.

Exercises for Chapter 12

For each exercise, plot the observations in any ways that seem reasonable. Describe the population(s) sampled, whether real or hypothetical. For each statistical procedure, state appropriate hypotheses. Discuss the assumptions that make the analysis appropriate. Do these assumptions seem reasonable? What additional information would you like to have about the experiment? Discuss the results of each analysis.

EXERCISE 12-1

Does an insect electrocuting device reduce mosquito biting? Researchers equipped suburban yards with either an insect electrocuting device, a standard 6-volt CDC trap, or no device. People serving as bait captured mosquitos coming to bite in each yard. The investigators took steps to allow for differences among yards and differences in attractiveness of the volunteers as mosquito bait. Details of the experiment are given in Nasci, Harris, and Porter (1983). The response for each yard is the percentage of the highest total number of mosquitos collected in any yard that night. The results are shown below.

Device	Percentage of maximum mosquito count						
Electrocuting device	66	57	57	31	87	97	89
	100	85	100	61	58		
CDC trap	100	75	50	77	58	100	62
	82	88	86	100	44		
None	75	84	100	74	40	94	87
	55	91	63	83	87		

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that the mean mosquito response is the same for each device. Do the assumptions of the analysis seem reasonable?

c. Go through the steps for a nonparametric analysis. Do the assumptions of this analysis seem reasonable?

d. Compare your results in parts (b) and (c). Discuss your findings.

e. Why did the investigators choose as a response the percentage of highest total number of mosquitos collected in a yard in a night?

EXERCISE 12-2

In a study of a synthetic vaccine for malaria, scientists divided twelve 18–21-year-old male volunteers into four groups. They assigned three volunteers to a saline control group. They divided the other nine men among three different vaccine dose/treatment regimens. After vaccination, the researchers recorded a stimulation index for each volunteer, determined from proliferation assays of peripheral blood mononuclear cells. The results are shown below (Patarroyo et al., 1988).

Group	Stimulation index			
Saline control	1.4	1.0	4.0	
Regimen 1	1.5	5.6	12.4	
Regimen 2	6.6	9.1		
Regimen 3	35.1	13.4	0.8	3.3

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that the mean stimulation index is the same for each treatment group. Do the assumptions for this analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

c. Take the logarithm of each stimulation index. Use a parametric analysis to test the null hypothesis that the mean of the logarithm of stimulation index is the same for each treatment group. Do the assumptions for this analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

d. Go through the steps for a nonparametric analysis. Do the assumptions for this analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

e. Compare your results in parts (b), (c), and (d). Discuss your findings.

EXERCISE 12-3

Researchers measured the amount of nitrogen expired by people on four different diets (Devore, 1982, page 600; from "Production of Gaseous Nitrogen in Human Steady-State Conditions," *J. Applied Physiology*, 1972, pages 155–159). The results are shown below.

Diet	Expired nitrogen (liters)					
Fasting	4.079	4.859	3.540	5.047	3.298	4.679
	2.870	4.648	3.847			
23% protein	4.368	5.668	3.752	5.848	3.802	4.844
	3.578	5.393	4.374			
32% protein	4.169	5.709	4.416	5.666	4.123	5.059
	4.403	4.496	4.688			
67% protein	4.928	5.608	4.940	5.291	4.674	5.038
	4.905	5.208	4.806			

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that mean expired nitrogen is the same for all four diet groups. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

c. Go through the steps for a nonparametric analysis. Do the assumptions for this analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

d. Compare your answers to parts (b) and (c). Discuss your findings.

EXERCISE 12-4

Researchers measured skin potential (in millivolts) in each of eight volunteers after requesting each of four emotions: fear, happiness, depression, and calmness. The results are shown below (Devore, 1982, page 599; from "Physiological Effects During Hypnotically Requested Emotions," *Psychosomatic Med.*, 1963, pages 334–343).

	Volunteer							
Emotion	1	2	3	4	5	6	7	8
Fear	23.1	57.6	10.5	23.6	11.9	54.6	21.0	20.3
Happiness	22.7	53.2	9.7	19.6	13.8	47.1	13.6	23.6
Depression	22.5	53.7	10.8	21.1	13.7	39.2	13.7	16.3
Calmness	22.6	53.1	8.3	21.6	13.3	37.0	14.8	14.8

a. Plot the observations.

b. Are the relative differences among emotions similar for all eight volunteers?

c. Use a parametric analysis to analyze the results of this experiment. Use residual plots to check model assumptions.

d. Use a nonparametric analysis to test for differences in skin potential under the four emotions. Compare your results with what you found in part (c).

e. Discuss your findings.

EXERCISE 12-5

In a study of the effects of long-term freezing on bread dough, researchers used three types of flour. They made four batches of bread dough using each of the three types of flour. They then froze the dough. After the period of freezing, the researchers removed the dough from the freezer and recorded the volume increase in the bread dough 4 hours later. The results are shown below (from an example in Hocking, 1985, page 7).

Flour type	Volume increase			
1	1.1	1.8	1.0	1.2
2	2.7	2.9	3.3	2.8
3	3.1	3.2	3.3	3.2

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that the mean volume increase is the same for the three types of flour. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

c. Use a nonparametric analysis to test the null hypothesis that the median volume increase is the same for the three types of flour. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

d. Compare your results in parts (b) and (c). Discuss your findings.

EXERCISE 12-6

An investigator wanted to compare the working life of three types of stopwatch (Rice, 1988, page 432; from Natrella, 1963). He tested several of each type, using each stopwatch through repeated cycles (on, off, restart) until it no longer worked. Survival times (thousands of cycles until failure) are listed below.

Type 1	1.7	1.9	6.1	12.5	16.5	25.1	30.5
	42.1	82.5					
Type 2	13.6	19.8	25.2	46.2	46.2	61.1	
Type 3	13.4	20.9	25.1	29.7	46.9		

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that mean life is the same for the three types of stopwatch. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

c. Use a nonparametric analysis to test the null hypothesis that median life is the same for the three types of stopwatch. Do the assumptions for this analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

d. Compare your answers to parts (b) and (c). Discuss your findings.

EXERCISE 12-7

W. F. Woodward, a shortstop for the 1970 Cincinnati Reds, compared three methods of rounding first base. Twenty-two volunteers used each method to round first base. Woodward recorded the time it took a volunteer to run from a point between home and first base (35 feet from home plate) to a point between first and second (15 feet short of second base). The response is the average time of two runs (units not given). The results are shown below (Hollander and Wolfe, 1973, pages 140–141; from W. F. Woodward, "A Comparison of Base Running Methods in Baseball," M. Sc. thesis, Florida State University, 1970).

Runner	Round out method	Narrow angle method	Wide angle method
1	5.40	5.50	5.55
2	5.85	5.70	5.75
3	5.20	5.60	5.50
4	5.55	5.50	5.40
5	5.90	5.85	5.70
6	5.45	5.55	5.60
7	5.40	5.40	5.35
8	5.45	5.50	5.35
9	5.25	5.15	5.00
10	5.85	5.80	5.70
11	5.25	5.20	5.10
12	5.65	5.55	5.45
13	5.60	5.35	5.45
14	5.05	5.00	4.95
15	5.50	5.50	5.40
16	5.45	5.55	5.50
17	5.55	5.55	5.35
18	5.45	5.50	5.55
19	5.50	5.45	5.25
20	5.65	5.60	5.40
21	5.70	5.65	5.55
22	6.30	6.30	6.25

a. Plot the observations.

b. Are the relative differences among methods similar for all 22 runners?

c. Use a parametric analysis to analyze the results of this experiment. Use residual plots to check model assumptions.

d. Use a nonparametric analysis to test for differences among methods of rounding first base. Do the assumptions for the analysis seem reasonable? Compare your results with what you found in part (c).

e. Discuss your findings.

EXERCISE 12-8

Researchers scored smoothness of nine types of fabric dried five ways (Devore, 1987, page 447; from "Line-Dried vs. Machine-Dried Fabrics: Comparison of Appearance, Hand, and Consumer Acceptance," *Home Econ. Research J.*, 1984, pages 27–35). The results are listed here.

Fabric	Machine dry	Line dry	Line dry, then 15-minute tumble	Line dry with softener	Line dry with air movement
Crepe	3.3	2.5	2.8	2.5	1.9
Double knit	3.6	2.0	3.6	2.4	2.3
Twill	4.2	3.4	3.8	3.1	3.1
Twill mix	3.4	2.4	2.9	1.6	1.7

Fabric	Machine dry	Line dry	Line dry, then 15-minute tumble	Line dry with softener	Line dry with air movement
Terry	3.8	1.3	2.8	2.0	1.6
Broadcloth	2.2	1.5	2.7	1.5	1.9
Sheeting	3.5	2.1	2.8	2.1	2.2
Corduroy	3.6	1.3	2.8	1.7	1.8
Denim	2.6	1.4	2.4	1.3	1.6

a. Plot the observations.

b. Are the relative differences among drying methods similar for the nine types of fabric?

c. Use a parametric analysis to analyze the results of this experiment. Use residual plots to check model assumptions.

d. Use a nonparametric analysis to test for differences among drying methods. Do the assumptions for the analysis seem reasonable? Compare your results with what you found in part (c).

e. Discuss your findings.

EXERCISE 12-9

Researchers wanted to study the effects of four treatments on earthworm populations. They applied all treatments at concentrations of 1,000 liters/hectare. (A hectare, abbreviated ha, is a metric unit of area equal to 2.471 acres.) The researchers divided a large rectangular field into 40 square plots, separated by buffer areas. They divided the 40 plots into groups of 10. All the plots in a group received one treatment. After treatment, the researchers applied an irritant that caused the earthworms to rise to the surface. They recorded total biomass/m^2 in equal sized subplots of each of the 40 plots. The results are shown below (part of a data set contributed by R. P. Blackshaw and P. J. Diggle to a collection of problems in Andrews and Herzberg, 1985, pages 301–306).

Treatment	Biomass/meter^2					
Water only	17.61	21.19	19.34	33.11	26.63	24.49
	39.12	16.40	53.32	39.26		
.5 kg/ha Benlate	72.61	24.47	9.38	63.90	36.10	28.38
	18.91	36.77	10.65	49.58		
.6 kg/ha Bevistin	57.10	74.06	23.74	28.40	32.31	32.15
	78.15	23.20	21.63	68.21		
1.4 kg/ha Cercobin	32.34	22.17	26.20	59.82	26.90	70.68
	63.01	55.54	49.26	78.62		

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that the mean biomass/m^2 is the same for all four treatments. Do the assumptions of the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

c. Use a nonparametric analysis to test the null hypothesis that the median biomass/m^2 is the same for all four treatments. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

d. Compare your results in parts (b) and (c). Discuss your findings.

EXERCISE 12-10 Some researchers wanted to compare yield using four methods of manufacturing penicillin. One important ingredient in producing penicillin is corn steep liquor. Because this ingredient is extremely variable, the researchers decided on a randomized block design. They divided a single blend of corn steep liquor into four parts and randomly assigned the four parts to the four manufacturing methods. These four runs comprised a block. To further reduce the effects of extraneous factors, the researchers used a random process to determine the order of runs within a block. The yields of penicillin (units not given) under the four manufacturing methods are listed below for each of five blends of corn steep liquor. The order of the run within a block is shown in parentheses next to the yield (from an example in *Statistics for Experimenters*, by Box, Hunter, and Hunter, John Wiley and Sons, New York, 1978, page 209).

Manufacturing method	Blend of corn steep liquor				
	1	2	3	4	5
1	89 (1)	84 (4)	81 (2)	87 (1)	79 (3)
2	88 (3)	77 (2)	87 (1)	92 (3)	81 (4)
3	97 (2)	92 (3)	87 (4)	89 (2)	80 (1)
4	94 (4)	79 (1)	85 (3)	84 (4)	88 (2)

a. Plot the observations.

b. Are the relative differences among manufacturers similar within all five blends of corn steep liquor?

c. Use a parametric analysis to analyze the results of this experiment. Use residual plots to check model assumptions.

d. Within each block, plot residuals versus run order. Does there appear to be a trend? If there were a trend, what would it mean?

e. Use a nonparametric analysis to test for differences among manufacturing methods on penicillin yield. Compare your results with what you found using the parametric analysis in part (c).

f. Discuss your findings.

EXERCISE 12-11 Investigators wanted to compare aggressive behavior of three species of mice, labeled I, II, and III. Species III was a cross of species I and II. The experimenters placed a mouse in the center of a box that was 1 meter square. The floor of the box was divided into 49 equal squares. The researchers recorded the number of squares the mouse crossed in 5 minutes. The results are shown below (Rice, 1988, pages 431–432).

Species I	309	229	182	228	326	289	231	225	307
	281	316	290	318	273	328	325	191	219
	216	221	198	181	110	256	240	122	290
	253	164	211	215	211	152	178	194	144
	95	157	240	146	106	252	266	284	274
	285	366	360	237	270	114	176	224	
Species II	37	90	39	104	43	62	17	19	21
	9	16	65	187	17	79	77	60	8
	81	39	133	102	36	19	53	59	29
	47	22	140	41	122	10	41	61	19
	62	86	66	64	53	79	46	89	74
	44	39	59	29	13	11	23	40	
Species III	140	218	215	109	151	154	93	103	90
	184	7	46	9	41	241	118	15	156
	111	120	163	101	170	225	177	72	288
	129								

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that the mean number of squares crossed is the same for each species. Do the assumptions for this analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

c. Use a nonparametric analysis to test the null hypothesis that the median number of squares crossed is the same for each species. Do the assumptions for this analysis seem reasonable?

d. Compare your answers to parts (b) and (c). Discuss your findings.

EXERCISE 12-12

Researchers applied five types of electrode to the arms of 16 volunteers and measured resistance (Berry, 1987). They wanted to see if the different types of electrode gave similar measurements. The results (in k.ohms) are shown below.

	Type of electrode				
Volunteer	1	2	3	4	5
1	500	400	98	200	250
2	660	600	600	75	310
3	250	370	220	250	220
4	72	140	240	33	54
5	135	300	450	430	70
6	27	84	135	190	180
7	100	50	82	73	78
8	105	180	32	58	32
9	90	180	220	34	64
10	200	290	320	280	135
11	15	45	75	88	80
12	160	200	300	300	220
13	250	400	50	50	92
14	170	310	230	20	150
15	66	1,000	1,050	280	220
16	107	48	26	45	51

a. Plot the observations.

b. Are the relative differences among electrode types similar for all 16 volunteers?

c. Use a parametric analysis to analyze the results of this experiment. Use residual plots to check model assumptions.

d. Use a nonparametric analysis to test for differences among electrode types. Compare your results with what you found using the parametric analysis in part (c).

e. There are two very large readings for volunteer 15. The investigators speculated that this may have been due to a large amount of hair on this volunteer's arm. However, we have no information on the amount of arm hair for any of the volunteers. Exclude the observations for volunteer 15 and repeat parts (b), (c), and (d). Compare your results when the observations for volunteer 15 are included and excluded.

f. Discuss your findings.

EXERCISE 12-13 Does knowledge of output improve performance in repetitive work? In this experiment, investigators looked at performance in grinding a piece of metal to meet size and shape specifications (Hollander and Wolfe, 1973, page 121; from Hundal, 1969). They randomly divided 18 men into three groups. The investigators gave the six men in the first group no information on their output. They gave the six men in the second group rough estimates of their output. They gave the six men in the third group accurate and detailed information on their output. The response variable is the number of pieces finished by each worker during a fixed time interval. The results are shown below.

No information						Rough information						Detailed information					
40	35	38	43	44	41	38	40	47	44	40	42	48	40	45	43	46	44

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that the mean output is the same under the three conditions. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

c. Use a nonparametric analysis to test the null hypothesis that the median output is the same under the three conditions. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

d. Compare your answers to parts (b) and (c). Discuss your findings.

EXERCISE 12-14 Researchers treated 12 patients with cardiac arrhythmias with each of three active drugs in a double-blind experiment. The researchers treated a patient with one drug for 1 week. They then made a 24-hour ambulatory electrocardiograph recording. They repeated this regimen for each of the three drugs, with treatment periods widely separated by intervals with no drugs. The response is the number of premature ventricular contractions per hour. The results are shown below (Berry, 1987).

Patient	Drug A	Drug B	Drug C
1	170	7	0
2	19	1.4	6
3	187	205	18
4	10	.3	1
5	216	.2	22
6	49	33	30
7	7	37	3
8	474	9	5
9	.4	.6	0
10	1.4	63	36
11	27	145	26
12	29	0	0

a. Plot the observations.

b. Are the relative differences among drugs similar for all 12 patients?

c. Use a parametric analysis to analyze the results of this experiment. Use residual plots to check model assumptions.

d. Use a nonparametric analysis to test for differences among drugs. Do the assumptions for the analysis seem reasonable? Compare with your results in part (c).

e. Discuss your findings.

EXERCISE 12-15

For a middle-school science project to study possible effects of acid rain, a student planted 12 tomato seeds in loam, in separate containers (Foster, 1986). She randomly divided the 12 containers into three groups. The student watered the four seeds in group 1 every day with water having pH 4.0. She watered the four seeds in group 2 every day with water having pH 5.6. Finally, she watered the four seeds in group 3 every day with distilled water having pH 7.0. When plants came up, she watered the soil (not the leaves). Three of the four plants in group 1 came up; all of the plants in the other two groups came up. After 3 weeks, the student measured the height of each plant. The results are shown below.

Group	Height of tomato plants (centimeters)			
pH 4.0	1.8	1.5	1.9	
pH 5.6	2.1	2.1	2.0	1.8
pH 7.0	2.7	2.6	2.4	2.3

Base your analyses on the 11 plants that came up.

a. Plot the observations.

b. Use a parametric analysis to test the null hypothesis that the mean height is the same for the three levels of pH. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

c. Use a nonparametric analysis to test the null hypothesis that the median height is the same for the three levels of pH. Do the assumptions for the analysis seem reasonable? Use the Bonferroni method to make multiple comparisons.

d. Compare your answers to parts (b) and (c). Discuss your findings.

e. How does the fact that one seed in group 1 did not germinate contribute to your discussion of this experiment?

EXERCISE 12-16 Suppose we have m null hypotheses to test. For each test, we state a criterion for deciding whether the data are inconsistent with the null hypothesis. We say the data are inconsistent with the "combined" null hypothesis if at least one of the m criteria is satisfied. Let α denote the significance level associated with this "combined" criterion. Let α_1 through α_m denote the significance levels associated with the m separate criteria. Show that α is less than or equal to the sum of α_1 through α_m. (*Hint:* We know from Chapter 6 that if event E can be written as the union of events E_1 through E_m, then the probability of E is less than or equal to the sum of the probabilities of events E_1 through E_m.)

EXERCISE 12-17 Find the Kruskal–Wallis distribution for samples of size 1, 2, and 2.

EXERCISE 12-18 Consider the experiment in Example 12-3. Ignore the technicians and go through the steps for one-way analysis of variance to test for differences among thermometers. Compare your results with what we found in Example 12-3.

EXERCISE 12-19 Test for differences among thermometers in Example 12-3, using Friedman's test. Compare your results with the results of the parametric analysis in Section 12-3. Do the assumptions for Friedman's test seem reasonable?

EXERCISE 12-20 Discuss the sampling situations in which one-way analysis of variance and the Kruskal–Wallis test are appropriate. Which procedure is preferred in each situation?

EXERCISE 12-21 Discuss the sampling situations in which the classical analysis of a randomized block experiment and Friedman's test are appropriate. Which procedure is preferred in each situation?