More Tests

Chapter 9

9.1 CHI-SQUARE GOODNESS-OF-FIT TESTS

We now consider applications of the very important chi-square statistic, first proposed by Karl Pearson in 1900. As the reader will see, it is a very adaptable test statistic and can be used for many different types of tests. In particular, one application allows us to test the appropriateness of different probabilistic models.

So that the reader can get some idea as to why Pearson first proposed his chi-square statistic, we begin with the binomial case. That is, let Y_1 be $b(n, p_1)$, where $0 < p_1 < 1$. According to the central limit theorem,

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

has a distribution that is approximately $N(0,1)$ for large n, particularly when $np_1 \geq 5$ and $n(1-p_1) \geq 5$. Thus, it is not surprising that $Q_1 = Z^2$ is approximately $\chi^2(1)$. If we let $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$, we see that Q_1 may be written as

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)}.$$

Since

$$(Y_1 - np_1)^2 = (n - Y_1 - n[1-p_1])^2 = (Y_2 - np_2)^2,$$

we have

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}.$$

Let us now carefully consider each term in this last expression for Q_1. Of course, Y_1 is the number of "successes," and np_1 is the expected number of "successes"; that is, $E(Y_1) = np_1$. Likewise, Y_2 and np_2 are, respectively, the number and the expected

number of "failures." So each numerator consists of the square of the difference of an observed number and an expected number. Note that Q_1 can be written as

$$Q_1 = \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i}, \tag{9.1-1}$$

and we have seen intuitively that it has an approximate chi-square distribution with one degree of freedom. In a sense, Q_1 measures the "closeness" of the observed numbers to the corresponding expected numbers. For example, if the observed values of Y_1 and Y_2 equal their expected values, then the computed Q_1 is equal to $q_1 = 0$; but if they differ much from them, then the computed $Q_1 = q_1$ is relatively large.

To generalize, we let an experiment have k (instead of only two) mutually exclusive and exhaustive outcomes, say, $A_1, A_2, \ldots, A_k$. Let $p_i = P(A_i)$, and thus $\sum_{i=1}^{k} p_i = 1$. The experiment is repeated n independent times, and we let Y_i represent the number of times the experiment results in A_i, $i = 1, 2, \ldots, k$. This joint distribution of $Y_1, Y_2, \ldots, Y_{k-1}$ is a straightforward generalization of the binomial distribution, as follows.

In considering the joint pmf, we see that

$$f(y_1, y_2, \ldots, y_{k-1}) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_{k-1} = y_{k-1}),$$

where $y_1, y_2, \ldots, y_{k-1}$ are nonnegative integers such that $y_1 + y_2 + \cdots + y_{k-1} \leq n$. Note that we do not need to consider Y_k, since, once the other $k-1$ random variables are observed to equal $y_1, y_2, \ldots, y_{k-1}$, respectively, we know that

$$Y_k = n - y_1 - y_2 - \cdots - y_{k-1} = y_k, \text{ say.}$$

From the independence of the trials, the probability of each particular arrangement of y_1 A_1s, y_2 A_2s, $\ldots$, y_k A_ks is

$$p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}.$$

The number of such arrangements is the multinomial coefficient

$$\binom{n}{y_1, y_2, \ldots, y_k} = \frac{n!}{y_1! y_2! \cdots y_k!}.$$

Hence, the product of these two expressions gives the joint pmf of $Y_1, Y_2, \ldots, Y_{k-1}$:

$$f(y_1, y_2, \ldots, y_{k-1}) = \frac{n!}{y_1! y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}.$$

(Recall that $y_k = n - y_1 - y_2 - \cdots - y_{k-1}$.)

Pearson then constructed an expression similar to Q_1 (Equation 9.1-1), which involves Y_1 and $Y_2 = n - Y_1$, that we denote by Q_{k-1}, which involves $Y_1, Y_2, \ldots, Y_{k-1}$, and $Y_k = n - Y_1 - Y_2 - \cdots - Y_{k-1}$, namely,

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}.$$

He argued that Q_{k-1} has an approximate chi-square distribution with $k - 1$ degrees of freedom in much the same way we argued that Q_1 is approximately $\chi^2(1)$. We accept Pearson's conclusion, as the proof is beyond the level of this text.

Some writers suggest that n should be large enough so that $np_i \geq 5$, $i = 1, 2, \ldots, k$, to be certain that the approximating distribution is adequate. This is

probably good advice for the beginner to follow, although we have seen the approximation work very well when $np_i \geq 1$, $i = 1, 2, \ldots, k$. The important thing to guard against is allowing some particular np_i to become so small that the corresponding term in Q_{k-1}, namely, $(Y_i - np_i)^2/np_i$, tends to dominate the others because of its small denominator. In any case, it is important to realize that Q_{k-1} has only an approximate chi-square distribution.

We shall now show how we can use the fact that Q_{k-1} is approximately $\chi^2(k-1)$ to test hypotheses about probabilities of various outcomes. Let an experiment have k mutually exclusive and exhaustive outcomes, $A_1, A_2, \ldots, A_k$. We would like to test whether $p_i = P(A_i)$ is equal to a known number p_{i0}, $i = 1, 2, \ldots, k$. That is, we shall test the hypothesis

$$H_0\colon\ p_i = p_{i0}, \qquad i = 1, 2, \ldots, k.$$

In order to test such a hypothesis, we shall take a sample of size n; that is, we repeat the experiment n independent times. We tend to favor H_0 if the observed number of times that A_i occurred, say, y_i, and the number of times A_i was expected to occur if H_0 were true, namely, np_{i0}, are approximately equal. That is, if

$$q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}}$$

is "small," we tend to favor H_0. Since the distribution of Q_{k-1} is approximately $\chi^2(k-1)$, we shall reject H_0 if $q_{k-1} \geq \chi^2_\alpha(k-1)$, where α is the desired significance level of the test.

Example 9.1-1

If persons are asked to record a string of random digits, such as

$$3 \quad 7 \quad 2 \quad 4 \quad 1 \quad 9 \quad 7 \quad 2 \quad 1 \quad 5 \quad 0 \quad 8 \ldots,$$

we usually find that they are reluctant to record the same or even the two closest numbers in adjacent positions. And yet, in true random-digit generation, the probability of the next digit being the same as the preceding one is $p_{10} = 1/10$, the probability of the next being only one away from the preceding (assuming that 0 is one away from 9) is $p_{20} = 2/10$, and the probability of all other possibilities is $p_{30} = 7/10$. We shall test one person's concept of a random sequence by asking her to record a string of 51 digits that seems to represent a random-digit generation. Thus, we shall test

$$H_0\colon\ p_1 = p_{10} = \frac{1}{10},\ p_2 = p_{20} = \frac{2}{10},\ p_3 = p_{30} = \frac{7}{10}.$$

The critical region for an $\alpha = 0.05$ significance level is $q_2 \geq \chi^2_{0.05}(2) = 5.991$. The sequence of digits was as follows:

5	8	3	1	9	4	6	7	9	2	6	3	0
8	7	5	1	3	6	2	1	9	5	4	8	0
3	7	1	4	6	0	4	3	8	2	7	3	9
8	5	6	1	8	7	0	3	5	2	5	2	

We went through this listing and observed how many times the next digit was the same as or was one away from the preceding one:

	Frequency	Expected Number
Same	0	50(1/10) = 5
One away	8	50(2/10) = 10
Other	42	50(7/10) = 35
Total	50	50

The computed chi-square statistic is

$$\frac{(0-5)^2}{5}+\frac{(8-10)^2}{10}+\frac{(42-35)^2}{35}=6.8>5.991=\chi^2_{0.05}(2).$$

Thus, we would say that this string of 51 digits does not seem to be random. ■

One major disadvantage in the use of the chi-square test is that it is a many-sided test. That is, the alternative hypothesis is very general, and it would be difficult to restrict alternatives to situations such as H_1: $p_1 > p_{10}, p_2 > p_{20}, p_3 < p_{30}$ (with $k = 3$). As a matter of fact, some statisticians would probably test H_0 against this particular alternative H_1 by using a linear function of Y_1, Y_2, and Y_3. However, that sort of discussion is beyond the scope of the book because it involves knowing more about the distributions of linear functions of the dependent random variables Y_1, Y_2, and Y_3. In any case, the student who truly recognizes that this chi-square statistic tests H_0: $p_i = p_{i0}$, $i = 1, 2, \ldots, k$, against all alternatives can usually appreciate the fact that it is more difficult to reject H_0 at a given significance level α when the chi-square statistic is used than it would be if some appropriate "one-sided" test statistic were available.

Many experiments yield a set of data, say, $x_1, x_2, \ldots, x_n$, and the experimenter is often interested in determining whether these data can be treated as the observed values of a random sample $X_1, X_2, \ldots, X_n$ from a given distribution. That is, would this proposed distribution be a reasonable probabilistic model for these sample items? To see how the chi-square test can help us answer questions of this sort, consider a very simple example.

Example 9.1-2

Let X denote the number of heads that occur when four coins are tossed at random. Under the assumption that the four coins are independent and the probability of heads on each coin is 1/2, X is $b(4, 1/2)$. One hundred repetitions of this experiment resulted in 0, 1, 2, 3, and 4 heads being observed on 7, 18, 40, 31, and 4 trials, respectively. Do these results support the assumptions? That is, is $b(4, 1/2)$ a reasonable model for the distribution of X? To answer this, we begin by letting $A_1 = \{0\}$, $A_2 = \{1\}$, $A_3 = \{2\}$, $A_4 = \{3\}$, and $A_5 = \{4\}$. If $p_{i0} = P(X \in A_i)$ when X is $b(4, 1/2)$, then

$$p_{10} = p_{50} = \binom{4}{0}\left(\frac{1}{2}\right)^4 = \frac{1}{16} = 0.0625,$$

$$p_{20} = p_{40} = \binom{4}{1}\left(\frac{1}{2}\right)^4 = \frac{4}{16} = 0.25,$$

$$p_{30} = \binom{4}{2}\left(\frac{1}{2}\right)^4 = \frac{6}{16} = 0.375.$$

At an approximate $\alpha = 0.05$ significance level, the null hypothesis

$$H_0\colon\ p_i = p_{i0}, \qquad i = 1, 2, \ldots, 5,$$

is rejected if the observed value of Q_4 is greater than $\chi^2_{0.05}(4) = 9.488$. If we use the 100 repetitions of this experiment that resulted in the observed values $y_1 = 7$, $y_2 = 18$, $y_3 = 40$, $y_4 = 31$, and $y_5 = 4$, of $Y_1, Y_2, \ldots, Y_5$, respectively, then the computed value of Q_4 is

$$q_4 = \frac{(7-6.25)^2}{6.25} + \frac{(18-25)^2}{25} + \frac{(40-37.5)^2}{37.5} + \frac{(31-25)^2}{25} + \frac{(4-6.25)^2}{6.25}$$
$$= 4.47.$$

Since $4.47 < 9.488$, the hypothesis is not rejected. That is, the data support the hypothesis that $b(4, 1/2)$ is a reasonable probabilistic model for X. Recall that the mean of a chi-square random variable is its number of degrees of freedom. In this example, the mean is 4 and the observed value of Q_4 is 4.47, just a little greater than the mean. ■

Thus far, all the hypotheses H_0 tested with the chi-square statistic Q_{k-1} have been simple ones (i.e., completely specified—namely, in H_0: $p_i = p_{i0}$, $i = 1, 2, \ldots, k$, each p_{i0} has been known). This is not always the case, and it frequently happens that $p_{10}, p_{20}, \ldots, p_{k0}$ are functions of one or more unknown parameters. For example, suppose that the hypothesized model for X in Example 9.1-2 was H_0: X is $b(4, p)$, $0 < p < 1$. Then

$$p_{i0} = P(X \in A_i) = \frac{4!}{(i-1)!(5-i)!} p^{i-1}(1-p)^{5-i}, \qquad i = 1, 2, \ldots, 5,$$

which is a function of the unknown parameter p. Of course, if H_0: $p_i = p_{i0}$, $i = 1, 2, \ldots, 5$, is true, then, for large n,

$$Q_4 = \sum_{i=1}^{5} \frac{(Y_i - np_{i0})^2}{np_{i0}}$$

still has an approximate chi-square distribution with four degrees of freedom. The difficulty is that when $Y_1, Y_2, \ldots, Y_5$ are observed to be equal to $y_1, y_2, \ldots, y_5$, Q_4 cannot be computed, since $p_{10}, p_{20}, \ldots, p_{50}$ (and hence Q_4) are functions of the unknown parameter p.

One way out of the difficulty would be to estimate p from the data and then carry out the computations with the use of this estimate. It is interesting to note the following: Say the estimation of p is carried out by minimizing Q_4 with respect to p, yielding $\widetilde{p}$. This $\widetilde{p}$ is sometimes called a **minimum chi-square estimator** of p. If, then, this $\widetilde{p}$ is used in Q_4, the statistic Q_4 still has an approximate chi-square

distribution, but with only $4 - 1 = 3$ degrees of freedom. That is, the number of degrees of freedom of the approximating chi-square distribution is reduced by one for each parameter estimated by the minimum chi-square technique. We accept this result without proof (as it is a rather difficult one). Although we have considered it when p_{i0}, $i = 1, 2, \ldots, k$, is a function of only one parameter, it holds when there is more than one unknown parameter, say, d. Hence, in a more general situation, the test would be completed by computing Q_{k-1}, using Y_i and the estimated p_{i0}, $i = 1, 2, \ldots, k$, to obtain q_{k-1} (i.e., q_{k-1} is the minimized chi-square). This value q_{k-1} would then be compared with a critical value $\chi^2_\alpha(k-1-d)$. In our special case, the computed (minimized) chi-square q_4 would be compared with $\chi^2_\alpha(3)$.

There is still one trouble with all of this: It is usually very difficult to find minimum chi-square estimators. Hence, most statisticians usually use some reasonable method of estimating the parameters. (Maximum likelihood is satisfactory.) They then compute q_{k-1}, recognizing that it is somewhat larger than the minimized chi-square, and compare it with $\chi^2_\alpha(k-1-d)$. Note that this approach provides a slightly larger probability of rejecting H_0 than would the scheme in which the minimized chi-square were used because the computed q_{k-1} is larger than the minimum q_{k-1}.

Example 9.1-3

Let X denote the number of alpha particles emitted by barium-133 in one tenth of a second. The following 50 observations of X were taken with a Geiger counter in a fixed position:

7	4	3	6	4	4	5	3	5	3
5	5	3	2	5	4	3	3	7	6
6	4	3	11	9	6	7	4	5	4
7	3	2	8	6	7	4	1	9	8
4	8	9	3	9	7	7	9	3	10

The experimenter is interested in determining whether X has a Poisson distribution. To test H_0: X is Poisson, we first estimate the mean of X—say, λ—with the sample mean, $\bar{x} = 5.4$, of these 50 observations. We then partition the set of outcomes for this experiment into the sets $A_1 = \{0, 1, 2, 3\}$, $A_2 = \{4\}$, $A_3 = \{5\}$, $A_4 = \{6\}$, $A_5 = \{7\}$, and $A_6 = \{8, 9, 10, \ldots\}$. (Note that we combined {0, 1, 2, 3} into one set A_1 and $\{8, 9, 10, \ldots\}$ into another A_6 so that the expected number of outcomes for each set would be at least five when H_0 is true.) In Table 9.1-1, the data are grouped and the estimated probabilities specified by the hypothesis that X has a Poisson distribution

Table 9.1-1 Grouped Geiger counter data

	Outcome					
	A_1	A_2	A_3	A_4	A_5	A_6
Frequency	13	9	6	5	7	10
Probability	0.213	0.160	0.173	0.156	0.120	0.178
Expected ($50p_i$)	10.65	8.00	8.65	7.80	6.00	8.90

with an estimated $\widehat{\lambda} = \bar{x} = 5.4$ are given. Since one parameter was estimated, Q_{6-1} has an approximate chi-square distribution with $r = 6-1-1 = 4$ degrees of freedom. Also, since

$$q_5 = \frac{[13 - 50(0.213)]^2}{50(0.213)} + \cdots + \frac{[10 - 50(0.178)]^2}{50(0.178)}$$
$$= 2.763 < 9.488 = \chi^2_{0.05}(4),$$

H_0 is not rejected at the 5% significance level. That is, with only these data, we are quite willing to accept the model that X has a Poisson distribution. ■

Let us now consider the problem of testing a model for the distribution of a random variable W of the continuous type. That is, if $F(w)$ is the distribution function of W, we wish to test

$$H_0\colon\ F(w) = F_0(w),$$

where $F_0(w)$ is some known distribution function of the continuous type. Recall that we have considered problems of this type in which we used q–q plots. In order to use the chi-square statistic, we must partition the set of possible values of W into k sets. One way this can be done is as follows: Partition the interval $[0, 1]$ into k sets with the points $b_0, b_1, b_2, \ldots, b_k$, where

$$0 = b_0 < b_1 < b_2 < \cdots < b_k = 1.$$

Let $a_i = F_0^{-1}(b_i)$, $i = 1, 2, \ldots, k-1$; $A_1 = (-\infty, a_1]$, $A_i = (a_{i-1}, a_i]$ for $i = 2, 3, \ldots, k-1$, and $A_k = (a_{k-1}, \infty)$; and $p_i = P(W \in A_i)$, $i = 1, 2, \ldots, k$. Let Y_i denote the number of times the observed value of W belongs to A_i, $i = 1, 2, \ldots, k$, in n independent repetitions of the experiment. Then $Y_1, Y_2, \ldots, Y_k$ have a multinomial distribution with parameters $n, p_1, p_2, \ldots, p_{k-1}$. Also, let $p_{i0} = P(W \in A_i)$ when the distribution function of W is $F_0(w)$. The hypothesis that we actually test is a modification of H_0, namely,

$$H_0'\colon\ p_i = p_{i0}, \qquad i = 1, 2, \ldots, k.$$

This hypothesis is rejected if the observed value of the chi-square statistic

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_{i0})^2}{np_{i0}}$$

is at least as great as $\chi^2_\alpha(k-1)$. If the hypothesis H_0': $p_i = p_{i0}$, $i = 1, 2, \ldots, k$, is not rejected, we do not reject the hypothesis H_0: $F(w) = F_0(w)$.

Example 9.1-4

Example 6.1-5 gives 105 observations of the times in minutes between calls to 911. Also given is a histogram of these data, with the exponential pdf with $\theta = 20$ superimposed. We shall now use a chi-square goodness-of-fit test to see whether or not this is an appropriate model for the data. That is, if X is equal to the time between calls to 911, we shall test the null hypothesis that the distribution of X is exponential with a mean of $\theta = 20$. Table 9.1-2 groups the data into nine classes and gives the probabilities and expected values of these classes. Using the frequencies and expected values, the chi-square goodness-of-fit statistic is

$$q_8 = \frac{(41 - 38.0520)^2}{38.0520} + \frac{(22 - 24.2655)^2}{24.2655} + \cdots + \frac{(2 - 2.8665)^2}{2.8665} = 4.6861.$$

Table 9.1-2 Summary of times between calls to 911

Class	Frequency	Probability	Expected
$A_1 = [0, 9]$	41	0.3624	38.0520
$A_2 = (9, 18]$	22	0.2311	24.2655
$A_3 = (18, 27]$	11	0.1473	15.4665
$A_4 = (27, 36]$	10	0.0939	9.8595
$A_5 = (36, 45]$	9	0.0599	6.2895
$A_6 = (45, 54]$	5	0.0382	4.0110
$A_7 = (54, 63]$	2	0.0244	2.5620
$A_8 = (63, 72]$	3	0.0155	1.6275
$A_9 = (72, \infty)$	2	0.0273	2.8665

The p-value associated with this test is 0.7905, which means that it is an extremely good fit.

Note that we assumed that we knew $\theta = 20$. We could also have run this test letting $\theta = \bar{x}$, remembering that we then lose one degree of freedom. For this example, the outcome would be about the same. ■

It is also true, in dealing with models of random variables of the continuous type, that we must frequently estimate unknown parameters. For example, let H_0 be that W is $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. With a random sample $W_1, W_2, \ldots, W_n$, we first can estimate μ and σ^2, possibly with $\overline{w}$ and s_w^2. We partition the space $\{w : -\infty < w < \infty\}$ into k mutually disjoint sets $A_1, A_2, \ldots, A_k$. We then use the estimates of μ and σ^2—say, $\overline{w}$ and $s^2 = s_w^2$, respectively, to estimate

$$\widehat{p}_{i0} = \int_{A_i} \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{(w - \overline{w})^2}{2s^2}\right] dw,$$

$i = 1, 2, \ldots, k$. Using the observed frequencies $y_1, y_2, \ldots, y_k$ of $A_1, A_2, \ldots, A_k$, respectively, from the observed random sample $w_1, w_2, \ldots, w_n$, and $\widehat{p}_{10}, \widehat{p}_{20}, \ldots, \widehat{p}_{k0}$ estimated with $\overline{w}$ and $s^2 = s_w^2$, we compare the computed

$$q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - n\widehat{p}_{i0})^2}{n\widehat{p}_{i0}}$$

with $\chi_\alpha^2(k-1-2)$. This value q_{k-1} will again be somewhat larger than that which would be found using minimum chi-square estimation, and certain caution should be observed. Several exercises illustrate the procedure in which one or more parameters must be estimated. Finally, note that the methods given in this section frequently are classified under the more general title of goodness-of-fit tests. In particular, then, the tests in this section would be **chi-square goodness-of-fit tests**.

Exercises

9.1-1. A 1-pound bag of candy-coated chocolate-covered peanuts contained 224 pieces of candy, each colored brown, orange, green, or yellow. Test the null hypothesis that the machine filling these bags treats the four colors of candy equally likely; that is, test

$$H_0\colon\ p_B = p_O = p_G = p_Y = \frac{1}{4}.$$

The observed values were 42 brown, 64 orange, 53 green, and 65 yellow candies. You may select the significance level or give an approximate p-value.

9.1-2. A particular brand of candy-coated chocolate comes in five different colors that we shall denote as $A_1 = \{\text{brown}\}$, $A_2 = \{\text{yellow}\}$, $A_3 = \{\text{orange}\}$, $A_4 = \{\text{green}\}$, and $A_5 = \{\text{coffee}\}$. Let p_i equal the probability that the color of a piece of candy selected at random belongs to A_i, $i = 1, 2, \ldots, 5$. Test the null hypothesis

$$H_0\colon\ p_1 = 0.4,\ p_2 = 0.2,\ p_3 = 0.2,\ p_4 = 0.1,\ p_5 = 0.1,$$

using a random sample of $n = 580$ pieces of candy whose colors yielded the respective frequencies 224, 119, 130, 48, and 59. You may select the significance level or give an approximate p-value.

9.1-3. In the Michigan Lottery Daily3 Game, twice a day a three-digit integer is generated one digit at a time. Let p_i denote the probability of generating digit i, $i = 0, 1, \ldots, 9$. Let $\alpha = 0.05$, and use the following 50 digits to test $H_0\colon p_0 = p_1 = \cdots = p_9 = 1/10$:

1	6	9	9	3	8	5	0	6	7
4	7	5	9	4	6	5	6	4	4
4	8	0	9	3	2	1	5	4	5
7	3	2	1	4	6	7	1	3	4
4	8	8	6	1	6	1	2	8	8

9.1-4. In a biology laboratory, students use corn to test the Mendelian theory of inheritance. The theory claims that frequencies of the four categories "smooth and yellow," "wrinkled and yellow," "smooth and purple," and "wrinkled and purple" will occur in the ratio 9:3:3:1. If a student counted 124, 30, 43, and 11, respectively, for these four categories, would these data support the Mendelian theory? Let $\alpha = 0.05$.

9.1-5. Let X equal the number of female children in a three-child family. We shall use a chi-square goodness-of-fit statistic to test the null hypothesis that the distribution of X is $b(3, 0.5)$.

(a) Define the test statistic and critical region, using an $\alpha = 0.05$ significance level.

(b) Among students who were taking statistics, 52 came from families with three children. For these families, $x = 0, 1, 2$, and 3 for 5, 17, 24, and 6 families, respectively. Calculate the value of the test statistic and state your conclusion, considering how the sample was selected.

9.1-6. It has been claimed that, for a penny minted in 1999 or earlier, the probability of observing heads upon spinning the penny is $p = 0.30$. Three students got together, and they would each spin a penny and record the number X of heads out of the three spins. They repeated this experiment $n = 200$ times, observing 0, 1, 2, and 3 heads 57, 95, 38, and 10 times, respectively. Use these data to test the hypotheses that X is $b(3, 0.30)$. Give limits for the p-value of this test. In addition, out of the 600 spins, calculate the number of heads occurring and then a 95% confidence interval for p.

9.1-7. A rare type of heredity change causes the bacterium in *E. coli* to become resistant to the drug streptomycin. This type of change, called *mutation*, can be detected by plating many bacteria on petri dishes containing an antibiotic medium. Any colonies that grow on this medium result from a single mutant cell. A sample of $n = 150$ petri dishes of streptomycin agar were each plated with 10^6 bacteria, and the numbers of colonies were counted on each dish. The observed results were that 92 dishes had 0 colonies, 46 had 1, 8 had 2, 3 had 3, and 1 dish had 4 colonies. Let X equal the number of colonies per dish. Test the hypothesis that X has a Poisson distribution. Use $\bar{x} = 0.5$ as an estimate of λ. Let $\alpha = 0.01$.

9.1-8. For determining the half-lives of radioactive isotopes, it is important to know what the background radiation is for a given detector over a certain period. A γ-ray detection experiment over 300 one-second intervals yielded the following data:

0	2	4	6	6	1	7	4	6	1	1	2	3	6	4	2	7	4	4	2
2	5	4	4	4	1	2	4	3	2	2	5	0	3	1	1	0	0	5	2
7	1	3	3	3	2	3	1	4	1	3	5	3	5	1	3	3	0	3	2
6	1	1	4	6	3	6	4	4	2	2	4	3	3	6	1	6	2	5	0
6	3	4	3	1	1	4	6	1	5	1	1	4	1	4	1	1	1	3	3
4	3	3	2	5	2	1	3	5	3	2	7	0	4	2	3	3	5	6	1
4	2	6	4	2	0	4	4	7	3	5	2	2	3	1	3	1	3	6	5
4	8	2	2	4	2	2	1	4	7	5	2	1	1	4	1	4	3	6	2
1	1	2	2	2	2	3	5	4	3	2	2	3	3	2	4	4	3	2	2
3	6	1	1	3	3	2	1	4	5	5	1	2	3	3	1	3	7	2	5
4	2	0	6	2	3	2	3	0	4	4	5	2	5	3	0	4	6	2	2
2	2	2	5	2	2	3	4	2	3	7	1	1	7	1	3	6	0	5	3
0	0	3	3	0	2	4	3	1	2	3	3	3	4	3	2	2	7	5	3
5	1	1	2	2	6	1	3	1	4	4	2	3	4	5	1	3	4	3	1
0	3	7	4	0	5	2	5	4	4	2	2	3	2	4	6	5	5	3	4

Do these look like observations of a Poisson random variable with mean $\lambda = 3$? To answer this question, do the following:

(a) Find the frequencies of $0, 1, 2, \ldots, 8$.

(b) Calculate the sample mean and sample variance. Are they approximately equal to each other?

(c) Construct a probability histogram with $\lambda = 3$ and a relative frequency histogram on the same graph.

(d) Use $\alpha = 0.05$ and a chi-square goodness-of-fit test to answer this question.

9.1-9. Let X equal the amount of butterfat (in pounds) produced by 90 cows during a 305-day milk production period following the birth of their first calf. Test the hypothesis that the distribution of X is $N(\mu, \sigma^2)$, using $k = 10$ classes of equal probability. You may take $\bar{x} = 511.633$ and $s_x = 87.576$ as estimates of μ and σ, respectively. The data are as follows:

486	537	513	583	453	510	570	500	458	555
618	327	350	643	500	497	421	505	637	599
392	574	492	635	460	696	593	422	499	524
539	339	472	427	532	470	417	437	388	481
537	489	418	434	466	464	544	475	608	444
573	611	586	613	645	540	494	532	691	478
513	583	457	612	628	516	452	501	453	643
541	439	627	619	617	394	607	502	395	470
531	526	496	561	491	380	345	274	672	509

9.1-10. A biologist is studying the life cycle of the avian schistosome that causes swimmer's itch. His study uses Menganser ducks for the adult parasites and aquatic snails as intermediate hosts for the larval stages. The life history is cyclic. (For more information, see http://swimmersitch.org/.) As a part of this study, the biologist and his students used snails from a natural population to measure the distances (in cm) that snails travel per day. The conjecture is that snails that had a patent infection would not travel as far as those without without such an infection.

Here are the measurements in cm that snails traveled per day. There are 39 in the infected group and 31 in the control group.

Distances for Infected Snail Group (ordered):

263	238	226	220	170	155	139	123	119	107	107	97	90
90	90	79	75	74	71	66	60	55	47	47	47	45
43	41	40	39	38	38	35	32	32	28	19	10	10

Distances for Control Snail Group (ordered):

314	300	274	246	190	186	185	182	180	141	132
129	110	100	95	95	93	83	55	52	50	48
48	44	40	32	30	25	24	18	7		

(a) Find the sample means and sample standard deviations for the two groups of snails.

(b) Make box plots of the two groups of snails on the same graph.

(c) For the control snail group, test the hypothesis that the distances come from an exponential distribution. Use $\bar{x}$ as an estimate of θ. Group the data into 5 or 10 classes, with equal probabilities for each class. Thus, the expected value will be either 6.2 or 3.1, respectively.

(d) For the infected snail group, test the hypothesis that the distances come from a gamma distribution with $\alpha = 2$ and $\theta = 42$. Use 10 classes with equal probabilities so that the expected value of each class is 3.9. Use Minitab or some other computer program to calculate the boundaries of the classes.

9.1-11. In Exercise 6.1-4, data are given for the melting points for 50 metal alloy filaments. Here the data are repeated:

320	326	325	318	322	320	329	317	316	331
320	320	317	329	316	308	321	319	322	335
318	313	327	314	329	323	327	323	324	314
308	305	328	330	322	310	324	314	312	318
313	320	324	311	317	325	328	319	310	324

Test the hypothesis that these are observations of a normally distributed random variable. Note that you must estimate two parameters: μ and σ.

9.2 CONTINGENCY TABLES

In this section, we demonstrate the flexibility of the chi-square test. We first look at a method for testing whether two or more multinomial distributions are equal, sometimes called a *test for homogeneity*. Then we consider a *test for independence of attributes of classification*. Both of these lead to a similar test statistic.

Suppose that each of two independent experiments can end in one of the k mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_k$. Let

$$p_{ij} = P(A_i), \qquad i = 1, 2, \ldots, k, \qquad j = 1, 2.$$

That is, $p_{11}, p_{21}, \ldots, p_{k1}$ are the probabilities of the events in the first experiment, and $p_{12}, p_{22}, \ldots, p_{k2}$ are those associated with the second experiment. Let the experiments be repeated n_1 and n_2 independent times, respectively. Also, let $Y_{11}, Y_{21}, \ldots, Y_{k1}$ be the frequencies of $A_1, A_2, \ldots, A_k$ associated with the n_1 independent trials of the first experiment. Similarly, let $Y_{12}, Y_{22}, \ldots, Y_{k2}$ be the respective frequencies associated with the n_2 trials of the second experiment. Of course, $\sum_{i=1}^{k} Y_{ij} = n_j, j = 1, 2$. From the sampling distribution theory corresponding to the basic chi-square test, we know that each of

$$\sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}}, \qquad j = 1, 2,$$

has an approximate chi-square distribution with $k - 1$ degrees of freedom. Since the two experiments are independent (and thus the two chi-square statistics are independent), the sum

$$\sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}}$$

is approximately chi-square with $k - 1 + k - 1 = 2k - 2$ degrees of freedom.

Usually, the $p_{ij}, i = 1, 2, \ldots, k, j = 1, 2$, are unknown, and frequently we wish to test the hypothesis

$$H_0\colon\ p_{11} = p_{12},\ p_{21} = p_{22},\ \ldots,\ p_{k1} = p_{k2};$$

that is, H_0 is the hypothesis that the corresponding probabilities associated with the two independent experiments are equal. Under H_0, we can estimate the unknown

$$p_{i1} = p_{i2}, \qquad i = 1, 2, \ldots, k,$$

by using the relative frequency $(Y_{i1} + Y_{i2})/(n_1 + n_2), i = 1, 2, \ldots, k$. That is, if H_0 is true, we can say that the two experiments are actually parts of a larger one in which $Y_{i1} + Y_{i2}$ is the frequency of the event $A_i, i = 1, 2, \ldots, k$. Note that we have to estimate only the $k - 1$ probabilities $p_{i1} = p_{i2}$, using

$$\frac{Y_{i1} + Y_{i2}}{n_1 + n_2}, \qquad i = 1, 2, \ldots, k - 1,$$

since the sum of the k probabilities must equal 1. That is, the estimator of $p_{k1} = p_{k2}$ is

$$1 - \frac{Y_{11} + Y_{12}}{n_1 + n_2} - \cdots - \frac{Y_{k-1,1} + Y_{k-1,2}}{n_1 + n_2} = \frac{Y_{k1} + Y_{k2}}{n_1 + n_2}.$$

Substituting these estimators, we find that

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{[Y_{ij} - n_j(Y_{i1} + Y_{i2})/(n_1 + n_2)]^2}{n_j(Y_{i1} + Y_{i2})/(n_1 + n_2)}$$

has an approximate chi-square distribution with $2k-2-(k-1)=k-1$ degrees of freedom. Here $k-1$ is subtracted from $2k-2$, because that is the number of estimated parameters. The critical region for testing H_0 is of the form

$$q \geq \chi^2_\alpha(k-1).$$

Example 9.2-1

To test two methods of instruction, 50 students are selected at random from each of two groups. At the end of the instruction period, each student is assigned a grade (A, B, C, D, or F) by an evaluating team. The data are recorded as follows:

	Grade					
	A	B	C	D	F	Totals
Group I	8	13	16	10	3	50
Group II	4	9	14	16	7	50

Accordingly, if the hypothesis H_0 that the corresponding probabilities are equal is true, then the respective estimates of the probabilities are

$$\frac{8+4}{100} = 0.12,\ 0.22,\ 0.30,\ 0.26,\ \frac{3+7}{100} = 0.10.$$

Thus, the estimates of $n_1 p_{i1} = n_2 p_{i2}$ are 6, 11, 15, 13, and 5, respectively. Hence, the computed value of Q is

$$\begin{aligned} q &= \frac{(8-6)^2}{6} + \frac{(13-11)^2}{11} + \frac{(16-15)^2}{15} + \frac{(10-13)^2}{13} + \frac{(3-5)^2}{5} \\ &\quad + \frac{(4-6)^2}{6} + \frac{(9-11)^2}{11} + \frac{(14-15)^2}{15} + \frac{(16-13)^2}{13} + \frac{(7-5)^2}{5} \\ &= \frac{4}{6} + \frac{4}{11} + \frac{1}{15} + \frac{9}{13} + \frac{4}{5} + \frac{4}{6} + \frac{4}{11} + \frac{1}{15} + \frac{9}{13} + \frac{4}{5} = 5.18. \end{aligned}$$

Now, under H_0, Q has an approximate chi-square distribution with $k-1=4$ degrees of freedom, so the $\alpha = 0.05$ critical region is $q \geq 9.488 = \chi^2_{0.05}(4)$. Here $q = 5.18 < 9.488$, and hence H_0 is not rejected at the 5% significance level. Furthermore, the p-value for $q = 5.18$ is 0.269, which is greater than most significance levels. Thus, with these data, we cannot say that there is a difference between the two methods of instruction. ■

It is fairly obvious how this procedure can be extended to testing the equality of h independent multinomial distributions. That is, let

$$p_{ij} = P(A_i), \qquad i = 1,2,\ldots,k, \qquad j = 1,2,\ldots,h,$$

and test

$$H_0:\ p_{i1} = p_{i2} = \cdots = p_{ih} = p_i, \qquad i = 1,2,\ldots,k.$$

Repeat the jth experiment n_j independent times, and let $Y_{1j}, Y_{2j}, \ldots, Y_{kj}$ denote the frequencies of the respective events $A_1, A_2, \ldots, A_k$. Now,

$$Q = \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}}$$

has an approximate chi-square distribution with $h(k-1)$ degrees of freedom. Under H_0, we must estimate $k-1$ probabilities, using

$$\widehat{p}_i = \frac{\sum_{j=1}^{h} Y_{ij}}{\sum_{j=1}^{h} n_j}, \qquad i = 1, 2, \ldots, k-1,$$

because the estimate of p_k follows from $\widehat{p}_k = 1 - \widehat{p}_1 - \widehat{p}_2 - \cdots - \widehat{p}_{k-1}$. We use these estimates to obtain

$$Q = \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j \widehat{p}_i)^2}{n_j \widehat{p}_i},$$

which has an approximate chi-square distribution, with its degrees of freedom given by $h(k-1) - (k-1) = (h-1)(k-1)$.

Let us see how we can use the preceding procedures to test the equality of two or more independent distributions that are not necessarily multinomial. Suppose first that we are given random variables U and V with distribution functions $F(u)$ and $G(v)$, respectively. It is sometimes of interest to test the hypothesis H_0: $F(x) = G(x)$ for all x. Previously, we considered tests of $\mu_U = \mu_V$, $\sigma_U^2 = \sigma_V^2$. In Section 8.4, we will look at the two-sample Wilcoxon test. Now we shall assume only that the distributions are independent and of the continuous type.

We are interested in testing the hypothesis H_0: $F(x) = G(x)$ for all x. This hypothesis will be replaced by another one. Partition the real line into k mutually disjoint sets $A_1, A_2, \ldots, A_k$. Let

$$p_{i1} = P(U \in A_i), \qquad i = 1, 2, \ldots, k,$$

and

$$p_{i2} = P(V \in A_i), \qquad i = 1, 2, \ldots, k.$$

We observe that if $F(x) = G(x)$ for all x, then $p_{i1} = p_{i2}$, $i = 1, 2, \ldots, k$. We replace the hypothesis H_0: $F(x) = G(x)$ with the less restrictive hypothesis H_0': $p_{i1} = p_{i2}$, $i = 1, 2, \ldots, k$. That is, we are now essentially interested in testing the equality of two multinomial distributions.

Let n_1 and n_2 denote the number of independent observations of U and V, respectively. For $i = 1, 2, \ldots, k$, let Y_{ij} denote the number of these observations of U and V, $j = 1, 2$, respectively, that fall into a set A_i. At this point, we proceed to make the test of H_0' as described earlier. Of course, if H_0' is rejected at the (approximate) significance level α, then H_0 is rejected with the same probability. However, if H_0' is true, H_0 is not necessarily true. Thus, if H_0' is not rejected, then we do not reject H_0.

In applications, the question of how to select $A_1, A_2, \ldots, A_k$ is frequently raised. Obviously, there is no single choice for k or for the dividing marks of the partition. But it is interesting to observe that the combined sample can be used in this selection without upsetting the approximate distribution of Q. For example, suppose that $n_1 = n_2 = 20$. Then we could easily select the dividing marks of the partition so that $k = 4$, and one fourth of the combined sample falls into each of the four sets.

Example 9.2-2

Select, at random, 20 cars of each of two comparable major-brand models. All 40 cars are submitted to accelerated life testing; that is, they are driven many miles over very poor roads in a short time, and their failure times (in weeks) are recorded as follows:

Brand U:	25	31	20	42	39	19	35	36	44	26
	38	31	29	41	43	36	28	31	25	38
Brand V:	28	17	33	25	31	21	16	19	31	27
	23	19	25	22	29	32	24	20	34	26

If we use 23.5, 28.5, and 34.5 as dividing marks, we note that exactly one fourth of the 40 cars fall into each of the resulting four sets. Thus, the data can be summarized as follows:

	A_1	A_2	A_3	A_4	Totals
Brand U	2	4	4	10	20
Brand V	8	6	6	0	20

The estimate of each p_i is $10/40 = 1/4$, which, multiplied by $n_j = 20$, gives 5. Hence, the computed Q is

$$q = \frac{(2-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(10-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(0-5)^2}{5}$$

$$= \frac{72}{5} = 14.4 > 7.815 = \chi^2_{0.05}(3).$$

Also, the p-value is 0.0024. Thus, it seems that the two brands of cars have different distributions for the length of life under accelerated life testing. Brand U seems better than brand V. ■

Again, it should be clear how this approach can be extended to more than two distributions, and this extension will be illustrated in the exercises.

Now let us suppose that a random experiment results in an outcome that can be classified by two different attributes, such as height and weight. Assume that the first attribute is assigned to one and only one of k mutually exclusive and exhaustive event—say $A_1, A_2, \ldots, A_k$—and the second attribute falls into one and only one of h mutually exclusive and exhaustive events—say, $B_1, B_2, \ldots, B_h$. Let the probability of $A_i \cap B_j$ be defined by

$$p_{ij} = P(A_i \cap B_j), \qquad i = 1, 2, \ldots, k, \qquad j = 1, 2, \ldots, h.$$

The random experiment is to be repeated n independent times, and Y_{ij} will denote the frequency of the event $A_i \cap B_j$. Since there are kh such events as $A_i \cap B_j$, the random variable

$$Q_{kh-1} = \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_{ij} - np_{ij})^2}{np_{ij}}$$

has an approximate chi-square distribution with $kh-1$ degrees of freedom, provided that n is large.

Suppose that we wish to test the hypothesis of the independence of the A and B attributes, namely,

$$H_0\colon\ P(A_i \cap B_j) = P(A_i)P(B_j), \qquad i = 1,2,\ldots,k, \qquad j = 1,2,\ldots,h.$$

Let us denote $P(A_i)$ by $p_{i\cdot}$ and $P(B_j)$ by $p_{\cdot j}$; that is,

$$p_{i\cdot} = \sum_{j=1}^{h} p_{ij} = P(A_i) \qquad \text{and} \qquad p_{\cdot j} = \sum_{i=1}^{k} p_{ij} = P(B_j).$$

Of course,

$$1 = \sum_{j=1}^{h}\sum_{i=1}^{k} p_{ij} = \sum_{j=1}^{h} p_{\cdot j} = \sum_{i=1}^{k} p_{i\cdot}.$$

Then the hypothesis can be formulated as

$$H_0\colon\ p_{ij} = p_{i\cdot}p_{\cdot j}, \qquad i = 1,2,\ldots,k, \qquad j = 1,2,\ldots,h.$$

To test H_0, we can use Q_{kh-1} with p_{ij} replaced by $p_{i\cdot}p_{\cdot j}$. But if $p_{i\cdot}$, $i = 1,2,\ldots,k$, and $p_{\cdot j}$, $j = 1,2,\ldots,h$, are unknown, as they usually are in applications, we cannot compute Q_{kh-1} once the frequencies are observed. In such a case, we estimate these unknown parameters by

$$\widehat{p}_{i\cdot} = \frac{y_{i\cdot}}{n}, \qquad \text{where} \quad y_{i\cdot} = \sum_{j=1}^{h} y_{ij}$$

is the observed frequency of A_i, $i = 1,2,\ldots,k$; and

$$\widehat{p}_{\cdot j} = \frac{y_{\cdot j}}{n}, \qquad \text{where} \quad y_{\cdot j} = \sum_{i=1}^{k} y_{ij}$$

is the observed frequency of B_j, $j = 1,2,\ldots,h$. Since $\sum_{i=1}^{k} p_{i\cdot} = \sum_{j=1}^{h} p_{\cdot j} = 1$, we actually estimate only $k-1+h-1 = k+h-2$ parameters. So if these estimates are used in Q_{kh-1}, with $p_{ij} = p_{i\cdot}p_{\cdot j}$, then, according to the rule stated earlier, the random variable

$$Q = \sum_{j=1}^{h}\sum_{i=1}^{k} \frac{[Y_{ij} - n(Y_{i\cdot}/n)(Y_{\cdot j}/n)]^2}{n(Y_{i\cdot}/n)(Y_{\cdot j}/n)}$$

has an approximate chi-square distribution with $kh-1-(k+h-2) = (k-1)(h-1)$ degrees of freedom, provided that H_0 is true. The hypothesis H_0 is rejected if the computed value of this statistic exceeds $\chi^2_\alpha[(k-1)(h-1)]$.

Example 9.2-3

The 400 undergraduate students in a random sample at the University of Iowa were classified according to the college in which the students were enrolled and according to their gender. The results are recorded in Table 9.2-1, called a $k \times h$ **contingency table**, where, in this case, $k = 2$ and $h = 5$. (Do not be concerned about the numbers in parentheses at this point.) Incidentally, these data do actually reflect the composition of the undergraduate colleges at Iowa, but they were modified a little to make the computations easier in this example.

Table 9.2-1 Undergraduates at the University of Iowa

	College					
Gender	Business	Engineering	Liberal Arts	Nursing	Pharmacy	Totals
Male	21	16	145	2	6	190
	(16.625)	(9.5)	(152)	(7.125)	(4.75)	
Female	14	4	175	13	4	210
	(18.375)	(10.5)	(168)	(7.875)	(5.25)	
Totals	35	20	320	15	10	400

We desire to test the null hypothesis H_0: $p_{ij} = p_{i\cdot}p_{\cdot j}$, $i = 1,2$ and $j = 1,2,3,4,5$, that the college in which a student enrolls is independent of the gender of that student. Under H_0, estimates of the probabilities are

$$\widehat{p}_{1\cdot} = \frac{190}{400} = 0.475 \qquad \text{and} \qquad \widehat{p}_{2\cdot} = \frac{210}{400} = 0.525$$

and

$$\widehat{p}_{\cdot 1} = \frac{35}{400} = 0.0875,\ \widehat{p}_{\cdot 2} = 0.05,\ \widehat{p}_{\cdot 3} = 0.8,\ \widehat{p}_{\cdot 4} = 0.0375,\ \widehat{p}_{\cdot 5} = 0.025.$$

The expected numbers $n(y_{i\cdot}/n)(y_{\cdot j}/n)$ are computed as follows:

$$400(0.475)(0.0875) = 16.625,$$
$$400(0.525)(0.0875) = 18.375,$$
$$400(0.475)(0.05) = 9.5,$$

and so on. These are the values recorded in parentheses in Table 9.2-1. The computed chi-square statistic is

$$\begin{aligned} q &= \frac{(21-16.625)^2}{16.625} + \frac{(14-18.375)^2}{18.375} + \cdots + \frac{(4-5.25)^2}{5.25} \\ &= 1.15 + 1.04 + 4.45 + 4.02 + 0.32 + 0.29 + 3.69 \\ &\quad + 3.34 + 0.33 + 0.30 \ = \ 18.93. \end{aligned}$$

Since the number of degrees of freedom equals $(k-1)(h-1) = 4$, this $q = 18.93 > 13.28 = \chi^2_{0.01}(4)$, and we reject H_0 at the $\alpha = 0.01$ significance level. Moreover, since the first two terms of q come from the business college, the next two from engineering, and so on, it is clear that the enrollments in engineering and nursing are more highly dependent on gender than in the other colleges, because they have contributed the most to the value of the chi-square statistic. It is also interesting to note that one expected number is less than 5, namely, 4.75. However, as the associated term in q does not contribute an unusual amount to the chi-square value, it does not concern us. ■

It is fairly obvious how to extend the preceding testing procedure to more than two attributes. For example, if the third attribute falls into one and only one of m mutually exclusive and exhaustive events—say, $C_1, C_2, \ldots, C_m$—then we test the independence of the three attributes by using

$$Q = \sum_{r=1}^{m} \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{[Y_{ijr} - n(Y_{i\cdot\cdot}/n)(Y_{\cdot j\cdot}/n)(Y_{\cdot\cdot r}/n)]^2}{n(Y_{i\cdot\cdot}/n)(Y_{\cdot j\cdot}/n)(Y_{\cdot\cdot r}/n)},$$

where Y_{ijr}, $Y_{i\cdot\cdot}$, $Y_{\cdot j\cdot}$, and $Y_{\cdot\cdot r}$ are the respective observed frequencies of the events $A_i \cap B_j \cap C_r$, A_i, B_j, and C_r in n independent trials of the experiment. If n is large and if the three attributes are independent, then Q has an approximate chi-square distribution with $khm - 1 - (k-1) - (h-1) - (m-1) = khm - k - h - m + 2$ degrees of freedom.

Rather than explore this extension further, it is more instructive to note some interesting uses of contingency tables.

Example 9.2-4

Say we observed 30 values $x_1, x_2, \ldots, x_{30}$ that are claimed to be the values of a random sample. That is, the corresponding random variables $X_1, X_2, \ldots, X_{30}$ were supposed to be mutually independent and each of these random variables is supposed to have the same distribution. Say, however, by looking at the 30 values, we detect an upward trend which indicates that there might have been some dependence and/or the random variables did not actually have the same distribution. One simple way to test whether they could be thought of as being observed values of a random sample is the following: Mark each x high (H) or low (L), depending on whether it is above or below the sample median. Then divide the x values into three groups: $x_1, \ldots, x_{10}$; $x_{11}, \ldots, x_{20}$; and $x_{21}, \ldots, x_{30}$. Certainly, if the observations are those of a random sample, we would expect five H's and five L's in each group. That is, the attribute classified as H or L should be independent of the group number. The summary of these data provides a 3×2 contingency table. For example, say the 30 values are

5.6	8.2	7.8	4.8	5.5	8.1	6.7	7.7	9.3	6.9
8.2	10.1	7.5	6.9	11.1	9.2	8.7	10.3	10.7	10.0
9.2	11.6	10.3	11.7	9.9	10.6	10.0	11.4	10.9	11.1

The median can be taken to be the average of the two middle observations in magnitude, namely, 9.2 and 9.3. Marking each item H or L after comparing it with this median, we obtain the following 3×2 contingency table:

Group	L	H	Totals
1	9	1	10
2	5	5	10
3	1	9	10
Totals	15	15	30

Here each $n(y_{i.}/n)(y_{.j}/n) = 30(10/30)(15/30) = 5$, so that the computed value of Q is

$$q = \frac{(9-5)^2}{5} + \frac{(1-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(1-5)^2}{5} + \frac{(9-5)^2}{5}$$
$$= 12.8 > 5.991 = \chi^2_{0.05}(2),$$

since in this instance $(k-1)(h-1) = 2$ degrees of freedom. (The p-value is 0.0017.) Hence, we reject the conjecture that these 30 values could be the observations of a random sample. Obviously, modifications could be made to this scheme: dividing the sample into more (or fewer) than three groups and rating items differently, such as low (L), middle (M), and high (H). ■

It cannot be emphasized enough that the chi-square statistic can be used fairly effectively in almost any situation in which there should be independence. For example, suppose that we have a group of workers who have essentially the same qualifications (training, experience, etc.). Many believe that the salary and gender of the workers should be independent attributes, yet there have been several claims in special cases that there is a dependence—or discrimination—in attributes associated with such a problem.

Example 9.2-5 Two groups of workers have the same qualifications for a particular type of work. Their experience in salaries is summarized by the following 2×5 contingency table, in which the upper bound of each salary range is not included in that listing:

	Salary (Thousands of Dollars)					
Group	27–29	29–31	31–33	33–35	35 and over	Totals
1	6	11	16	14	13	60
2	5	9	8	6	2	30
Totals	11	20	24	20	15	90

To test whether the group assignment and the salaries seem to be independent with these data at the $\alpha = 0.05$ significance level, we compute

$$q = \frac{[6 - 90(60/90)(11/90)]^2}{90(60/90)(11/90)} + \cdots + \frac{[2 - 90(30/90)(15/90)]^2}{90(30/90)(15/90)}$$
$$= 4.752 < 9.488 = \chi^2_{0.05}(4).$$

Also, the p-value is 0.314. Hence, with these limited data, group assignment and salaries seem to be independent. ■

Before turning to the exercises, note that we could have thought of the last two examples in this section as testing the equality of two or more multinomial distributions. In Example 9.2-4, the three groups define three binomial distributions, and in Example 9.2-5, the two groups define two multinomial distributions. What would have happened if we had used the computations outlined earlier in the section? It is interesting to note that we obtain exactly the same value of chi-square and in each

case the number of degrees of freedom is equal to $(k-1)(h-1)$. Hence, it makes no difference whether we think of it as a test of independence or a test of the equality of several multinomial distributions. Our advice is to use the terminology that seems most natural for the particular situation.

Exercises

9.2-1. We wish to see if two groups of nurses distribute their time in six different categories about the same way. That is, the hypothesis under consideration is H_0: $p_{i1} = p_{i2}$, $i = 1,2,\ldots,6$. To test this hypothesis, nurses are observed at random throughout several days, each observation resulting in a mark in one of the six categories. A summary of the results is given by the following frequency table:

	Category						
	1	2	3	4	5	6	Totals
Group I	95	36	71	21	45	32	300
Group II	53	26	43	18	32	28	200

Use a chi-square test with $\alpha = 0.05$.

9.2-2. Suppose that a third group of nurses was observed along with groups I and II of Exercise 9.2-1, resulting in the respective frequencies 130, 75, 136, 33, 61, and 65. Test H_0: $p_{i1} = p_{i2} = p_{i3}$, $i = 1,2,\ldots,6$, at the $\alpha = 0.025$ significance level.

9.2-3. Each of two comparable classes of 15 students responded to two different methods of instructions, giving the following scores on a standardized test:

Class U:	91	42	39	62	55	82	67	44
	51	77	61	52	76	41	59	
Class V:	80	71	55	67	61	93	49	78
	57	88	79	81	63	51	75	

Use a chi-square test with $\alpha = 0.05$ to test the equality of the distributions of test scores by dividing the combined sample into three equal parts (low, middle, high).

9.2-4. Suppose that a third class (W) of 15 students was observed along with classes U and V of Exercise 9.2-3, resulting in scores of

91	73	67	83	59	98	87	69
78	80	65	94	82	74	85	

Again, use a chi-square test with $\alpha = 0.05$ to test the equality of the three distributions by dividing the combined sample into three equal parts.

9.2-5. In the following contingency table, 1015 individuals are classified by gender and by whether they favor, oppose, or have no opinion on a complete ban on smoking in public places:

	Smoking in Public Places			
Gender	Favor	Oppose	No Opinion	Totals
Male	262	231	10	503
Female	302	205	5	512
Totals	564	436	15	1015

Test the null hypothesis that gender and opinion on smoking in public places are independent. Give the approximate p-value of this test.

9.2-6. A random survey of 100 students asked each student to select the most preferred form of recreational activity from five choices. Following are the results of the survey:

	Recreational Choice					
Gender	Basketball	Baseball Softball	Swimming	Jogging Running	Tennis	Totals
Male	21	5	9	12	13	60
Female	9	3	1	15	12	40
Totals	30	8	10	27	25	100

Test whether the choice is independent of the gender of the respondent. Approximate the p-value of the test. Would we reject the null hypothesis at $\alpha = 0.05$?

9.2-7. One hundred music majors in a random sample were classified as follows by gender and by the kind of instrument (including voice) that they played:

	Instrument					
Gender	Piano	Woodwind	Brass	String	Vocal	Totals
Male	4	11	15	6	9	45
Female	7	18	6	6	18	55
Totals	11	29	21	12	27	100

Test whether the selection of instrument is independent of the gender of the respondent. Approximate the p-value of this test.

9.2-8. A student who uses a certain college's recreational facilities was interested in whether there is a difference between the facilities used by men and those used by women. Use $\alpha = 0.05$ and the following data to test the null hypothesis that facility and gender are independent attributes:

	Facility		
Gender	Racquetball Court	Track	Totals
Male	51	30	81
Female	43	48	91
Totals	94	78	172

9.2-9. A survey of high school girls classified them by two attributes: whether or not they participated in sports and whether or not they had one or more older brothers. Use the following data to test the null hypothesis that these two attributes of classification are independent:

	Participated in Sports		
Older Brother(s)	Yes	No	Totals
Yes	12	8	20
No	13	27	40
Totals	25	35	60

Approximate the p-value of this test. Do we reject the null hypothesis if $\alpha = 0.05$?

9.2-10. A random sample of 50 women who were tested for cholesterol was classified according to age and cholesterol level and grouped into the following contingency table.

	Cholesterol Level			
Age	<180	180–210	>210	Totals
<50	5	11	9	25
≥50	4	3	18	25
Totals	9	14	27	50

Test the null hypothesis H_0: Age and cholesterol level are independent attributes of classification. What is your conclusion if $\alpha = 0.01$?

9.2-11. Although high school grades and testing scores, such as SAT or ACT, can be used to predict first-year college grade-point average (GPA), many educators claim that a more important factor influencing GPA is the living conditions of students. In particular, it is claimed that the roommate of the student will have a great influence on his or her grades. To test this hypothesis, suppose we selected at random 200 students and classified each according to the following two attributes:

(a) Ranking of the student's roommate on a scale from 1 to 5, with 1 denoting a person who was difficult to live with and discouraged scholarship, and 5 signifying a person who was congenial and encouraged scholarship.

(b) The student's first-year GPA.

Say this classification gives the following 5×4 contingency table:

	Grade-Point Average				
Rank of Roommate	Under 2.00	2.00–2.69	2.70–3.19	3.20–4.00	Totals
1	8	9	10	4	31
2	5	11	15	11	42
3	6	7	20	14	47
4	3	5	22	23	53
5	1	3	11	12	27
Totals	23	35	78	64	200

Compute the chi-square statistic used to test the independence of the two attributes, and compare it with the critical value associated with $\alpha = 0.05$.

9.2-12. In a psychology experiment, 140 students were divided into majors emphasizing left-hemisphere brain skills (e.g., philosophy, physics, and mathematics) and majors emphasizing right-hemisphere skills (e.g., art, music, theater, and dance). They were also classified into

one of three groups on the basis of hand posture (right noninverted, left inverted, and left noninverted). The data are as follows:

	LH	RH
RN	89	29
LI	5	4
LN	5	8

Do these data show sufficient evidence to reject the claim that the choice of college major is independent of hand posture? Let $\alpha = 0.025$.

9.2-13. A study was conducted to determine the media credibility for reporting news. Those surveyed were asked to give their age, gender, education, and the most credible medium. The results of the survey are as follows:

	Most Credible Medium			
Age	Newspaper	Television	Radio	Totals
Under 35	30	68	10	108
35–54	61	79	20	160
Over 54	98	43	21	162
Totals	189	190	51	430

	Most Credible Medium			
Gender	Newspaper	Television	Radio	Totals
Male	92	108	19	219
Female	97	81	32	210
Totals	189	189	51	429

	Most Credible Medium			
Education	Newspaper	Television	Radio	Totals
Grade School	45	22	6	73
High School	94	115	30	239
College	49	52	13	114
Totals	188	189	49	426

(a) Test whether media credibility and age are independent.

(b) Test whether media credibility and gender are independent.

(c) Test whether media credibility and education are independent.

(d) Give the approximate p-value for each test.

9.3 ONE-FACTOR ANALYSIS OF VARIANCE

Frequently, experimenters want to compare more than two treatments: yields of several different corn hybrids; results due to three or more teaching techniques; or miles per gallon obtained from many different types of compact cars. Sometimes the different treatment distributions of the resulting observations are due to changing the level of a certain factor (e.g., different doses of a given drug). Thus, the consideration of the equality of the different means of the various distributions comes under the analysis of a **one-factor experiment**.

In Section 8.2, we discussed how to compare the means of two normal distributions. More generally, let us now consider m normal distributions with unknown means $\mu_1, \mu_2, \ldots, \mu_m$ and an unknown, but common, variance σ^2. One inference that we wish to consider is a test of the equality of the m means, namely, H_0: $\mu_1 = \mu_2 = \cdots = \mu_m = \mu$, with μ unspecified, against all possible alternative hypotheses H_1. In order to test this hypothesis, we shall take independent random samples from these distributions. Let $X_{i1}, X_{i2}, \ldots, X_{in_i}$ represent a random sample of size n_i from the normal distribution $N(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, m$. In Table 9.3-1, we have

Table 9.3-1 One-factor random samples

					Means
X_1:	X_{11}	X_{12}	$\cdots$	X_{1n_1}	$\overline{X}_{1\cdot}$
X_2:	X_{21}	X_{22}	$\cdots$	X_{2n_2}	$\overline{X}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_m:	X_{m1}	X_{m2}	$\cdots$	X_{mn_m}	$\overline{X}_{m\cdot}$
Grand Mean:					$\overline{X}_{\cdot\cdot}$

indicated these random samples along with the row means (sample means), where, with $n = n_1 + n_2 + \cdots + n_m$,

$$\overline{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij} \qquad \text{and} \qquad \overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \qquad i = 1, 2, \ldots, m.$$

The dot in the notation for the means, $\overline{X}_{\cdot\cdot}$ and $\overline{X}_{i\cdot}$, indicates the index over which the average is taken. Here $\overline{X}_{\cdot\cdot}$ is an average taken over both indices, while $\overline{X}_{i\cdot}$ is taken over just the index j.

To determine a critical region for a test of H_0, we shall first partition the sum of squares associated with the variance of the combined samples into two parts. This sum of squares is given by

$$\begin{aligned}
\text{SS(TO)} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot} + \overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 \\
&\quad + 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}).
\end{aligned}$$

The last term of the right-hand member of this identity may be written as

$$2\sum_{i=1}^{m}\left[(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})\right] = 2\sum_{i=1}^{m}(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})(n_i\overline{X}_{i\cdot} - n_i\overline{X}_{i\cdot}) = 0,$$

and the preceding term may be written as

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 = \sum_{i=1}^{m} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2.$$

Thus,

$$\text{SS(TO)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2 + \sum_{i=1}^{m} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2.$$

For notation, let

$\text{SS(TO)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{\cdot\cdot})^2$, the total sum of squares;

$\text{SS(E)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2$, the sum of squares within treatments, groups, or classes, often called the error sum of squares;

$\text{SS(T)} = \sum_{i=1}^{m} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$, the sum of squares among the different treatments, groups, or classes, often called the between-treatment sum of squares.

Hence,

$$\text{SS(TO)} = \text{SS(E)} + \text{SS(T)}.$$

When H_0 is true, we may regard X_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$, as a random sample of size $n = n_1 + n_2 + \cdots + n_m$ from the normal distribution $N(\mu, \sigma^2)$. Then $\text{SS(TO)}/(n-1)$ is an unbiased estimator of σ^2 because $\text{SS(TO)}/\sigma^2$ is $\chi^2(n-1)$, so that $E[\text{SS(TO)}/\sigma^2] = n - 1$ and $E[\text{SS(TO)}/(n-1)] = \sigma^2$. An unbiased estimator of σ^2 based only on the sample from the ith distribution is

$$W_i = \frac{\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2}{n_i - 1} \qquad \text{for } i = 1, 2, \ldots, m,$$

because $(n_i - 1)W_i/\sigma^2$ is $\chi^2(n_i - 1)$. Thus,

$$E\left[\frac{(n_i - 1)W_i}{\sigma^2}\right] = n_i - 1,$$

and so

$$E(W_i) = \sigma^2, \qquad i = 1, 2, \ldots, m.$$

It follows that the sum of m of these independent chi-square random variables, namely,

$$\sum_{i=1}^{m}\frac{(n_i - 1)W_i}{\sigma^2} = \frac{\text{SS(E)}}{\sigma^2},$$

is also chi-square with $(n_1 - 1) + (n_2 - 1) + \cdots + (n_m - 1) = n - m$ degrees of freedom. Hence, $\text{SS(E)}/(n - m)$ is an unbiased estimator of σ^2. We now have

$$\frac{\text{SS(TO)}}{\sigma^2} = \frac{\text{SS(E)}}{\sigma^2} + \frac{\text{SS(T)}}{\sigma^2},$$

where

$$\frac{\text{SS(TO)}}{\sigma^2} \text{ is } \chi^2(n-1) \quad \text{and} \quad \frac{\text{SS(E)}}{\sigma^2} \text{ is } \chi^2(n-m).$$

Because SS(T) ≥ 0, there is a theorem (see subsequent remark) which states that SS(E) and SS(T) are independent and the distribution of SS(T)/σ^2 is $\chi^2(m-1)$.

REMARK The sums of squares, SS(T), SS(E), and SS(TO), are examples of **quadratic forms** in the variables X_{ij}, $i = 1,2,\ldots,m$, $j = 1,2,\ldots,n_i$. That is, each term in these sums of squares is of second degree in X_{ij}. Furthermore, the coefficients of the variables are real numbers, so these sums of squares are called **real quadratic forms**. The next theorem, stated without proof, is used in this chapter. [For a proof, see Hogg, McKean, and Craig, *Introduction to Mathematical Statistics*, 7th ed. (Upper Saddle River: Prentice Hall, 2013).] ■

Theorem 9.3-1

Let $Q = Q_1 + Q_2 + \cdots + Q_k$, where $Q, Q_1, \ldots, Q_k$ are $k+1$ real quadratic forms in n mutually independent random variables normally distributed with the same variance σ^2. Let $Q/\sigma^2, Q_1/\sigma^2, \ldots, Q_{k-1}/\sigma^2$ have chi-square distributions with $r, r_1, \ldots, r_{k-1}$ degrees of freedom, respectively. If Q_k is nonnegative, then

(a) $Q_1, \ldots, Q_k$ are mutually independent, and hence,

(b) Q_k/σ^2 has a chi-square distribution with $r - (r_1 + \cdots + r_{k-1}) = r_k$ degrees of freedom.

Since, under H_0, SS(T)/σ^2 is $\chi^2(m-1)$, we have $E[\text{SS(T)}/\sigma^2] = m-1$ and it follows that $E[\text{SS(T)}/(m-1)] = \sigma^2$. Now, the estimator of σ^2, namely, SS(E)/$(n-m)$, which is based on SS(E), is always unbiased, whether H_0 is true or false. However, if the means $\mu_1, \mu_2, \ldots, \mu_m$ are not equal, the expected value of the estimator that is based on SS(T) will be greater than σ^2. To make this last statement clear, we have

$$\begin{aligned} E[\text{SS(T)}] &= E\left[\sum_{i=1}^{m} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2\right] = E\left[\sum_{i=1}^{m} n_i\overline{X}_{i\cdot}^2 - n\overline{X}_{\cdot\cdot}^2\right] \\ &= \sum_{i=1}^{m} n_i\{\text{Var}(\overline{X}_{i\cdot}) + [E(\overline{X}_{i\cdot})]^2\} - n\{\text{Var}(\overline{X}_{\cdot\cdot}) + [E(\overline{X}_{\cdot\cdot})]^2\} \\ &= \sum_{i=1}^{m} n_i\left\{\frac{\sigma^2}{n_i} + \mu_i^2\right\} - n\left\{\frac{\sigma^2}{n} + \overline{\mu}^2\right\} \\ &= (m-1)\sigma^2 + \sum_{i=1}^{m} n_i(\mu_i - \overline{\mu})^2, \end{aligned}$$

where $\overline{\mu} = (1/n)\sum_{i=1}^{m} n_i\mu_i$. If $\mu_1 = \mu_2 = \cdots = \mu_m = \mu$, then

$$E\left(\frac{\text{SS(T)}}{m-1}\right) = \sigma^2.$$

If the means are not all equal, then

$$E\left[\frac{\text{SS(T)}}{m-1}\right] = \sigma^2 + \sum_{i=1}^{m} n_i \frac{(\mu_i - \overline{\mu})^2}{m-1} > \sigma^2.$$

We can base our test of H_0 on the ratio of $\text{SS(T)}/(m-1)$ and $\text{SS(E)}/(n-m)$, both of which are unbiased estimators of σ^2, provided that H_0: $\mu_1 = \mu_2 = \cdots = \mu_m$ is true, so that, under H_0, the ratio would assume values near 1. However, in the case that the means $\mu_1, \mu_2, \ldots, \mu_m$ begin to differ, this ratio tends to become large, since $E[\text{SS(T)}/(m-1)]$ gets larger. Under H_0, the ratio

$$\frac{\text{SS(T)}/(m-1)}{\text{SS(E)}/(n-m)} = \frac{[\text{SS(T)}/\sigma^2]/(m-1)}{[\text{SS(E)}/\sigma^2]/(n-m)} = F$$

has an F distribution with $m-1$ and $n-m$ degrees of freedom because $\text{SS(T)}/\sigma^2$ and $\text{SS(E)}/\sigma^2$ are independent chi-square variables. We would reject H_0 if the observed value of F is too large because this would indicate that we have a relatively large SS(T), suggesting that the means are unequal. Thus, the critical region is of the form $F \geq F_\alpha(m-1, n-m)$.

The information used for tests of the equality of several means is often summarized in an **analysis-of-variance table**, or **ANOVA** table, like that given in Table 9.3-2, where the mean square (MS) is the sum of squares (SS) divided by its degrees of freedom.

Example 9.3-1

Let X_1, X_2, X_3, X_4 be independent random variables that have normal distributions $N(\mu_i, \sigma^2)$, $i = 1, 2, 3, 4$. We shall test

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

against all alternatives on the basis of a random sample of size $n_i = 3$ from each of the four distributions. A critical region of size $\alpha = 0.05$ is given by

$$F = \frac{\text{SS(T)}/(4-1)}{\text{SS(E)}/(12-4)} \geq 4.07 = F_{0.05}(3, 8).$$

The observed data are shown in Table 9.3-3. (Clearly, these data are not observations from normal distributions; they were selected to illustrate the calculations.)

Table 9.3-2 Analysis-of-variance table

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F Ratio
Treatment	SS(T)	$m-1$	$\text{MS(T)} = \dfrac{\text{SS(T)}}{m-1}$	$\dfrac{\text{MS(T)}}{\text{MS(E)}}$
Error	SS(E)	$n-m$	$\text{MS(E)} = \dfrac{\text{SS(E)}}{n-m}$	
Total	SS(TO)	$n-1$		

Table 9.3-3 Illustrative data

	Observations			$\overline{X}_{i\cdot}$
x_1:	13	8	9	10
x_2:	15	11	13	13
x_3:	8	12	7	9
x_4:	11	15	10	12
$\overline{x}_{\cdot\cdot}$				11

For the given data, the calculated SS(TO), SS(E), and SS(T) are

$$\text{SS(TO)} = (13-11)^2 + (8-11)^2 + \cdots + (15-11)^2 + (10-11)^2 = 80,$$

$$\text{SS(E)} = (13-10)^2 + (8-10)^2 + \cdots + (15-12)^2 + (10-12)^2 = 50,$$

$$\text{SS(T)} = 3[(10-11)^2 + (13-11)^2 + (9-11)^2 + (12-11)^2] = 30.$$

Note that since SS(TO) = SS(E) + SS(T), only two of the three values need to be calculated directly from the data. Here the computed value of F is

$$\frac{30/3}{50/8} = 1.6 < 4.07,$$

and H_0 is not rejected. The p-value is the probability, under H_0, of obtaining an F that is at least as large as this computed value of F. It is often given by computer programs.

The information for this example is summarized in Table 9.3-4. Again, we note that (here and elsewhere) the F statistic is the ratio of two appropriate mean squares. ■

Formulas that sometimes simplify the calculations of SS(TO), SS(T), and SS(E) (and also reduce roundoff errors created by subtracting the averages from the observations) are

Table 9.3-4 ANOVA table for illustrative data

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F Ratio	p-value
Treatment	30	3	30/3	1.6	0.264
Error	50	8	50/8		
Total	80	11			

$$\text{SS(TO)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{1}{n}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}\right]^2,$$

$$\text{SS(T)} = \sum_{i=1}^{m}\frac{1}{n_i}\left[\sum_{j=1}^{n_i} X_{ij}\right]^2 - \frac{1}{n}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}\right]^2,$$

and

$$\text{SS(E)} = \text{SS(TO)} - \text{SS(T)}.$$

It is interesting to note that in these formulas each square is divided by the number of observations in the sum being squared: X_{ij}^2 by 1, $(\sum_{j=1}^{n_i} X_{ij})^2$ by n_i, and $(\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij})^2$ by n. The preceding formulas are used in Example 9.3-2. Although they are useful, you are encouraged to use appropriate statistical packages on a computer to aid you with these calculations.

If the sample sizes are all at least equal to 7, insight can be gained by plotting box-and-whisker diagrams on the same figure, for each of the samples. This technique is also illustrated in Example 9.3-2.

Example 9.3-2

A window that is manufactured for an automobile has five studs for attaching it. A company that manufactures these windows performs "pullout tests" to determine the force needed to pull a stud out of the window. Let X_i, $i = 1, 2, 3, 4, 5$, equal the force required at position i, and assume that the distribution of X_i is $N(\mu_i, \sigma^2)$. We shall test the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, using seven independent observations at each position. At an $\alpha = 0.01$ significance level, H_0 is rejected if the computed

$$F = \frac{\text{SS(T)}/(5-1)}{\text{SS(E)}/(35-5)} \geq 4.02 = F_{0.01}(4, 30).$$

The observed data, along with certain sums, are given in Table 9.3-5. For these data,

Table 9.3-5 Pullout test data

	Observations							$\sum_{j=1}^{7} x_{ij}$	$\sum_{j=1}^{7} x_{ij}^2$
x_1:	92	90	87	105	86	83	102	645	59,847
x_2:	100	108	98	110	114	97	94	721	74,609
x_3:	143	149	138	136	139	120	145	970	134,936
x_4:	147	144	160	149	152	131	134	1017	148,367
x_5:	142	155	119	134	133	146	152	981	138,415
Totals								4334	556,174

$$\text{SS(TO)} = 556,174 - \frac{1}{35}(4334)^2 = 19{,}500.97,$$

$$\text{SS(T)} = \frac{1}{7}(645^2 + 721^2 + 970^2 + 1017^2 + 981^2) - \frac{1}{35}(4334)^2 = 16{,}672.11,$$

$$\text{SS(E)} = 19{,}500.97 - 16{,}672.11 = 2828.86.$$

Since the computed F is

$$F = \frac{16{,}672.11/4}{2828.86/30} = 44.20,$$

the null hypothesis is clearly rejected. This information obtained from the equations is summarized in Table 9.3-6.

But why is H_0 rejected? The box-and-whisker diagrams shown in Figure 9.3-1 help to answer this question. It looks like the forces required to pull out studs in positions 1 and 2 are similar, and those in positions 3, 4, and 5 are quite similar, but different from, positions 1 and 2. (See Exercise 9.3-10.) An examination of the window would confirm that this is the case. ■

As with the two-sample t test, the F test works quite well even if the underlying distributions are nonnormal, unless they are highly skewed or the variances are quite different. In these latter cases, we might need to transform the observations

Table 9.3-6 ANOVA table for pullout tests

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F
Treatment	16,672.11	4	4,168.03	44.20
Error	2,828.86	30	94.30	
Total	19,500.97	34		

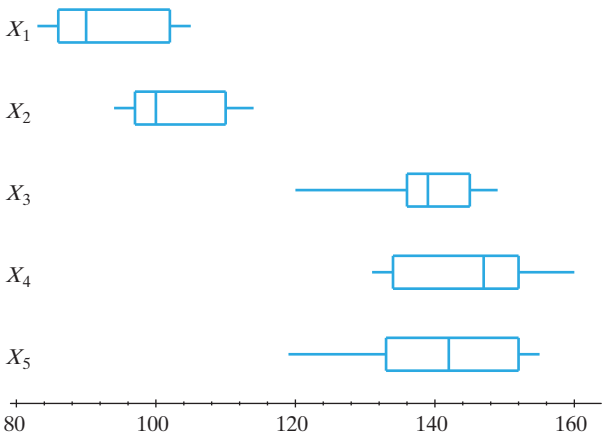

Figure 9.3-1 Box plots for pullout tests

to make the data more symmetric with about the same variances or to use certain nonparametric methods that are beyond the scope of this text.

Exercises

(In some of the exercises that follow, we must make assumptions, such as normal distributions with equal variances.)

9.3-1. Let μ_1, μ_2, μ_3 be, respectively, the means of three normal distributions with a common, but unknown, variance σ^2. In order to test, at the $\alpha = 0.05$ significance level, the hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$ against all possible alternative hypotheses, we take a random sample of size 4 from each of these distributions. Determine whether we accept or reject H_0 if the observed values from the three distributions are, respectively, as follows:

x_1:	5	9	6	8
x_2:	11	13	10	12
x_3:	10	6	9	9

9.3-2. Let μ_i be the average yield in bushels per acre of variety i of corn, $i = 1, 2, 3, 4$. In order to test the hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ at the 5% significance level, four test plots for each of the four varieties of corn are planted. Determine whether we accept or reject H_0 if the yield in bushels per acre of the four varieties of corn are, respectively, as follows:

x_1:	158.82	166.99	164.30	168.73
x_2:	176.84	165.69	167.87	166.18
x_3:	180.16	168.84	170.65	173.58
x_4:	151.58	163.51	164.57	160.75

9.3-3. Four groups of three pigs each were fed individually four different feeds for a specified length of time to test the hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, where μ_i, $i = 1, 2, 3, 4$, is the mean weight gain for each of the feeds. Determine whether the null hypothesis is accepted or rejected at a 5% significance level if the observed weight gains in pounds are, respectively, as follows:

x_1:	194.11	182.80	187.43
x_2:	216.06	203.50	216.88
x_3:	178.10	189.20	181.33
x_4:	197.11	202.68	209.18

9.3-4. Ledolter and Hogg (see References) report that a civil engineer wishes to compare the strengths of three different types of beams, one (A) made of steel and two (B and C) made of different and more expensive alloys. A certain deflection (in units of 0.001 inch) was measured for each beam when submitted to a given force; thus, a small deflection would indicate a beam of great strength. The order statistics for the three samples, of respective sizes $n_1 = 8$, $n_2 = 6$, and $n_3 = 6$, are as follows:

A:	79	82	83	84	85	86	86	87
B:	74	75	76	77	78	82		
C:	77	78	79	79	79	82		

(a) Use these data, $\alpha = 0.05$, and the F test to test the equality of the three means.

(b) For each set of data, construct box-and-whisker diagrams on the same figure and give an interpretation of your diagrams.

9.3-5. The female cuckoo lays her eggs in other birds' nests. The "foster parents" are usually deceived, probably because of the similarity in sizes of their own eggs and cuckoo eggs. Latter (see References) investigated this possible explanation and measured the lengths of cuckoo eggs (in mm) that were found in the nests of three species. Following are his results:

Hedge sparrow:	22.0	23.9	20.9	23.8	25.0
	24.0	21.7	23.8	22.8	23.1
	23.1	23.5	23.0	23.0	
Robin:	21.8	23.0	23.3	22.4	23.0
	23.0	23.0	22.4	23.9	22.3
	22.0	22.6	22.0	22.1	21.1
	23.0				
Wren:	19.8	22.1	21.5	20.9	22.0
	21.0	22.3	21.0	20.3	20.9
	22.0	20.0	20.8	21.2	21.0

(a) Construct an ANOVA table to test the equality of the three means.

(b) For each set of data, construct box-and-whisker diagrams on the same figure.

(c) Interpret your results.

9.3-6. Let X_1, X_2, X_3, X_4 equal the cholesterol level of a woman under the age of 50, a man under 50, a woman 50

or older, and a man 50 or older, respectively. Assume that the distribution of X_i is $N(\mu_i, \sigma^2)$, $i = 1, 2, 3, 4$. We shall test the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, using seven observations of each X_i.

(a) Give a critical region for an $\alpha = 0.05$ significance level.

(b) Construct an ANOVA table and state your conclusion, using the following data:

x_1:	221	213	202	183	185	197	162
x_2:	271	192	189	209	227	236	142
x_3:	262	193	224	201	161	178	265
x_4:	192	253	248	278	232	267	289

(c) Give bounds on the p-value for this test.

(d) For each set of data, construct box-and-whisker diagrams on the same figure and give an interpretation of your diagram.

9.3-7. Montgomery (see References) examines the strengths of a synthetic fiber that may be affected by the percentage of cotton in the fiber. Five levels of this percentage are considered, with five observations taken at each level.

Percentage of Cotton	Tensile Strength in lb/in^2				
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

Use the F test, with $\alpha = 0.05$, to see if there are differences in the breaking strengths due to the percentages of cotton used.

9.3-8. Different sizes of nails are packaged in "1-pound" boxes. Let X_i equal the weight of a box with nail size $(4i)C$, $i = 1, 2, 3, 4, 5$, where $4C, 8C, 12C, 16C$, and $20C$ are the sizes of the sinkers from smallest to largest. Assume that the distribution of X_i is $N(\mu_i, \sigma^2)$. To test the null hypothesis that the mean weights of "1-pound" boxes are all equal for different sizes of nails, we shall use random samples of size 7, weighing the nails to the nearest hundredth of a pound.

(a) Give a critical region for an $\alpha = 0.05$ significance level.

(b) Construct an ANOVA table and state your conclusion, using the following data:

x_1:	1.03	1.04	1.07	1.03	1.08	1.06	1.07
x_2:	1.03	1.10	1.08	1.05	1.06	1.06	1.05
x_3:	1.03	1.08	1.06	1.02	1.04	1.04	1.07
x_4:	1.10	1.10	1.09	1.09	1.06	1.05	1.08
x_5:	1.04	1.06	1.07	1.06	1.05	1.07	1.05

(c) For each set of data, construct box-and-whisker diagrams on the same figure and give an interpretation of your diagrams.

9.3-9. Let X_i, $i = 1, 2, 3, 4$, equal the distance (in yards) that a golf ball travels when hit from a tee, where i denotes the index of the ith manufacturer. Assume that the distribution of X_i is $N(\mu_i, \sigma^2)$, $i = 1, 2, 3, 4$, when a ball is hit by a certain golfer. We shall test the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, using three observations of each random variable.

(a) Give a critical region for an $\alpha = 0.05$ significance level.

(b) Construct an ANOVA table and state your conclusion, using the following data:

x_1:	240	221	265
x_2:	286	256	272
x_3:	259	245	232
x_4:	239	215	223

(c) What would your conclusion be if $\alpha = 0.025$?

(d) What is the approximate p-value of this test?

9.3-10. From the box-and-whisker diagrams in Figure 9.3-1, it looks like the means of X_1 and X_2 could be equal and also that the means of X_3, X_4, and X_5 could be equal but different from the first two.

(a) Using the data in Example 9.3-2, as well as a t test and an F test, test H_0: $\mu_1 = \mu_2$ against a two-sided alternative hypothesis. Let $\alpha = 0.05$. Do the F and t tests give the same result?

(b) Using the data in Example 9.3-2, test H_0: $\mu_3 = \mu_4 = \mu_5$. Let $\alpha = 0.05$.

9.3-11. The driver of a diesel-powered automobile decided to test the quality of three types of diesel fuel sold in the area. The test is to be based on miles per gallon (mpg). Make the usual assumptions, take $\alpha = 0.05$, and use the following data to test the null hypothesis that the three means are equal:

Brand A:	38.7	39.2	40.1	38.9	
Brand B:	41.9	42.3	41.3		
Brand C:	40.8	41.2	39.5	38.9	40.3

9.3-12. A particular process puts a coating on a piece of glass so that it is sensitive to touch. Randomly throughout the day, pieces of glass are selected from the production line and the resistance is measured at 12 different locations on the glass. On each of three different days, December 6, December 7, and December 22, the following data give the means of the 12 measurements on each of 11 pieces of glass:

December 6:	175.05	177.44	181.94	176.51	182.12	164.34
	163.20	168.12	171.26	171.92	167.87	
December 7:	175.93	176.62	171.39	173.90	178.34	172.90
	174.67	174.27	177.16	184.13	167.21	
December 22:	167.27	161.48	161.86	173.83	170.75	172.90
	173.27	170.82	170.93	173.89	177.68	

(a) Use these data to test whether the means on all three days are equal.

(b) Use box-and-whisker diagrams to confirm your answer.

9.3-13. For an aerosol product, there are three weights: the tare weight (container weight), the concentrate weight, and the propellant weight. Let X_1, X_2, X_3 denote the propellant weights on three different days. Assume that each of these independent random variables has a normal distribution with common variance and respective means μ_1, μ_2, and μ_3. We shall test the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$, using nine observations of each of the random variables.

(a) Give a critical region for an $\alpha = 0.01$ significance level.

(b) Construct an ANOVA table and state your conclusion, using the following data:

x_1:	43.06	43.32	42.63	42.86	43.05
	42.87	42.94	42.80	42.36	
x_2:	42.33	42.81	42.13	42.41	42.39
	42.10	42.42	41.42	42.52	
x_3:	42.83	42.57	42.96	43.16	42.25
	42.24	42.20	41.97	42.61	

(c) For each set of data, construct box-and-whisker diagrams on the same figure and give an interpretation of your diagrams.

9.3-14. Ledolter and Hogg (see References) report the comparison of three workers with different experience who manufacture brake wheels for a magnetic brake. Worker A has four years of experience, worker B has seven years, and worker C has one year. The company is concerned about the product's quality, which is measured by the difference between the specified diameter and the actual diameter of the brake wheel. On a given day, the supervisor selects nine brake wheels at random from the output of each worker. The following data give the differences between the specified and actual diameters in hundredths of an inch:

Worker A:	2.0	3.0	2.3	3.5	3.0	2.0	4.0	4.5	3.0
Worker B:	1.5	3.0	4.5	3.0	3.0	2.0	2.5	1.0	2.0
Worker C:	2.5	3.0	2.0	2.5	1.5	2.5	2.5	3.0	3.5

(a) Test whether there are statistically significant differences in the quality among the three different workers.

(b) Do box plots of the data confirm your answer in part (a)?

9.3-15. Ledolter and Hogg (see References) report that an operator of a feedlot wants to compare the effectiveness of three different cattle feed supplements. He selects a random sample of 15 one-year-old heifers from his lot of over 1000 and divides them into three groups at random. Each group gets a different feed supplement. Upon noting that one heifer in group A was lost due to an accident, the operator records the gains in weight (in pounds) over a six-month period as follows:

Group A:	500	650	530	680	
Group B:	700	620	780	830	860
Group C:	500	520	400	580	410

(a) Test whether there are differences in the mean weight gains due to the three different feed supplements.

(b) Do box plots of the data confirm your answer in part (a)?

9.4 TWO-WAY ANALYSIS OF VARIANCE

The test of the equality of several means, considered in Section 9.3, is an example of a statistical inference method called the analysis of variance (ANOVA). This method derives its name from the fact that the quadratic form SS(TO) $= (n-1)S^2$—the total

sum of squares about the combined sample mean—is decomposed into its components and analyzed. In this section, other problems in the analysis of variance will be investigated; here we restrict our considerations to the two-factor case, but the reader can see how it can be extended to three-factor and other cases.

Consider a situation in which it is desirable to investigate the effects of two factors that influence an outcome of an experiment. For example, a teaching method (lecture, discussion, computer assisted, television, etc.) and the size of a class might influence a student's score on a standard test; or the type of car and the grade of gasoline used might change the number of miles per gallon. In this latter example, if the number of miles per gallon is not affected by the grade of gasoline, we would no doubt use the least expensive grade.

The first analysis-of-variance model that we discuss is referred to as a **two-way classification with one observation per cell**. Assume that there are two factors (attributes), one of which has a levels and the other b levels. There are thus $n = ab$ possible combinations, each of which determines a cell. Let us think of these cells as being arranged in a rows and b columns. Here we take one observation per cell, and we denote the observation in the ith row and jth column by X_{ij}. Assume further that X_{ij} is $N(\mu_{ij}, \sigma^2)$, $i = 1, 2, \ldots, a$, and $j = 1, 2, \ldots, b$; and the $n = ab$ random variables are independent. [The assumptions of normality and homogeneous (same) variances can be somewhat relaxed in applications, with little change in the significance levels of the resulting tests.] We shall assume that the means μ_{ij} are composed of a row effect, a column effect, and an overall effect in some additive way, namely, $\mu_{ij} = \mu + \alpha_i + \beta_j$, where $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$. The parameter α_i represents the ith row effect, and the parameter β_j represents the jth column effect.

REMARK There is no loss in generality in assuming that

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0.$$

To see this, let $\mu_{ij} = \mu' + \alpha_i' + \beta_j'$. Write

$$\overline{\alpha}' = \left(\frac{1}{a}\right)\sum_{i=1}^{a} \alpha_i' \quad \text{and} \quad \overline{\beta}' = \left(\frac{1}{b}\right)\sum_{j=1}^{b} \beta_j'.$$

We have

$$\mu_{ij} = (\mu' + \overline{\alpha}' + \overline{\beta}') + (\alpha_i' - \overline{\alpha}') + (\beta_j' - \overline{\beta}') = \mu + \alpha_i + \beta_j,$$

where $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$. The reader is asked to find μ, α_i, and β_j for one display of μ_{ij} in Exercise 9.4-2. ■

To test the hypothesis that there is no row effect, we would test H_A: $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$, since $\sum_{i=1}^{a} \alpha_i = 0$. Similarly, to test that there is no column effect, we would test H_B: $\beta_1 = \beta_2 = \cdots = \beta_b = 0$, since $\sum_{j=1}^{b} \beta_j = 0$. To test these hypotheses, we shall again partition the total sum of squares into several components. Letting

$$\overline{X}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} X_{ij}, \quad \overline{X}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} X_{ij}, \quad \overline{X}_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij},$$

we have

$$\begin{aligned}
\text{SS(TO)} &= \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{..})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}[(\overline{X}_{i.}-\overline{X}_{..})+(\overline{X}_{.j}-\overline{X}_{..})+(X_{ij}-\overline{X}_{i.}-\overline{X}_{.j}+\overline{X}_{..})]^2 \\
&= b\sum_{i=1}^{a}(\overline{X}_{i.}-\overline{X}_{..})^2 + a\sum_{j=1}^{b}(\overline{X}_{.j}-\overline{X}_{..})^2 \\
&\quad + \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{i.}-\overline{X}_{.j}+\overline{X}_{..})^2 \\
&= \text{SS(A)} + \text{SS(B)} + \text{SS(E)},
\end{aligned}$$

where SS(A) is the sum of squares among levels of factor A, or among rows; SS(B) is the sum of squares among levels of factor B, or among columns; and SS(E) is the error or residual sum of squares. In Exercise 9.4-4, the reader is asked to show that the three cross-product terms in the square of the trinomial sum to zero. The distribution of the error sum of squares does not depend on the mean μ_{ij}, provided that the additive model is correct. Hence, its distribution is the same whether H_A or H_B is true or not, and thus SS(E) acts as a "measuring stick," as did SS(E) in Section 9.3. This can be seen more clearly by writing

$$\begin{aligned}
\text{SS(E)} &= \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{i.}-\overline{X}_{.j}+\overline{X}_{..})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}[X_{ij}-(\overline{X}_{i.}-\overline{X}_{..})-(\overline{X}_{.j}-\overline{X}_{..})-\overline{X}_{..}]^2
\end{aligned}$$

and noting the similarity of the summand in the right-hand member to

$$X_{ij}-\mu_{ij} = X_{ij}-\alpha_i-\beta_j-\mu.$$

We now show that $\text{SS(A)}/\sigma^2$, $\text{SS(B)}/\sigma^2$, and $\text{SS(E)}/\sigma^2$ are independent chi-square variables, provided that both H_A and H_B are true—that is, when all the means μ_{ij} have a common value μ. To do this, we first note that $\text{SS(TO)}/\sigma^2$ is $\chi^2(ab-1)$. In addition, from Section 9.3, we see that expressions such as $\text{SS(A)}/\sigma^2$ and $\text{SS(B)}/\sigma^2$ are chi-square variables, namely, $\chi^2(a-1)$ and $\chi^2(b-1)$, by replacing the n_i of Section 9.3 by a and b, respectively. Obviously, $\text{SS(E)} \geq 0$, and hence by Theorem 9.3-1, $\text{SS(A)}/\sigma^2$, $\text{SS(B)}/\sigma^2$, and $\text{SS(E)}/\sigma^2$ are independent chi-square variables with $a-1$, $b-1$, and $ab-1-(a-1)-(b-1) = (a-1)(b-1)$ degrees of freedom, respectively.

To test the hypothesis H_A: $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$, we shall use the row sum of squares SS(A) and the residual sum of squares SS(E). When H_A is true, $\text{SS(A)}/\sigma^2$ and $\text{SS(E)}/\sigma^2$ are independent chi-square variables with $a-1$ and $(a-1)(b-1)$ degrees of freedom, respectively. Thus, $\text{SS(A)}/(a-1)$ and $\text{SS(E)}/[(a-1)(b-1)]$ are both unbiased estimators of σ^2 when H_A is true. However, $E[\text{SS(A)}/(a-1)] > \sigma^2$ when H_A is not true, and hence we would reject H_A when

$$F_A = \frac{\text{SS(A)}/[\sigma^2(a-1)]}{\text{SS(E)}/[\sigma^2(a-1)(b-1)]} = \frac{\text{SS(A)}/(a-1)}{\text{SS(E)}/[(a-1)(b-1)]}$$

is "too large." Since F_A has an F distribution with $a-1$ and $(a-1)(b-1)$ degrees of freedom when H_A is true, H_A is rejected if the observed value of F_A equals or exceeds $F_\alpha[a-1,(a-1)(b-1)]$.

Similarly, the test of the hypothesis H_B: $\beta_1 = \beta_2 = \cdots = \beta_b = 0$ against all alternatives can be based on

$$F_B = \frac{\text{SS(B)}/[\sigma^2(b-1)]}{\text{SS(E)}/[\sigma^2(a-1)(b-1)]} = \frac{\text{SS(B)}/(b-1)}{\text{SS(E)}/[(a-1)(b-1)]},$$

which has an F distribution with $b-1$ and $(a-1)(b-1)$ degrees of freedom, provided that H_B is true.

Table 9.4-1 is the ANOVA table that summarizes the information needed for these tests of hypotheses. The formulas for F_A and F_B show that each of them is a ratio of two mean squares.

Example 9.4-1

Each of three cars is driven with each of four different brands of gasoline. The number of miles per gallon driven for each of the $ab = (3)(4) = 12$ different combinations is recorded in Table 9.4-2.

We would like to test whether we can expect the same mileage for each of these four brands of gasoline. In our notation, we test the hypothesis

$$H_B: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

Table 9.4-1 Two-way ANOVA table, one observation per cell

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F
Factor A (row)	SS(A)	$a-1$	$\text{MS(A)} = \frac{\text{SS(A)}}{a-1}$	$\frac{\text{MS(A)}}{\text{MS(E)}}$
Factor B (column)	SS(B)	$b-1$	$\text{MS(B)} = \frac{\text{SS(B)}}{b-1}$	$\frac{\text{MS(B)}}{\text{MS(E)}}$
Error	SS(E)	$(a-1)(b-1)$	$\text{MS(E)} = \frac{\text{SS(E)}}{(a-1)(b-1)}$	
Total	SS(TO)	$ab-1$		

Table 9.4-2 Gas mileage data

	Gasoline				
Car	1	2	3	4	$\overline{X}_{i\cdot}$
1	26	28	31	31	29
2	24	25	28	27	26
3	25	25	28	26	26
$\overline{X}_{\cdot j}$	25	26	29	28	27

against all alternatives. At a 1% significance level, we shall reject H_B if the computed F, namely,

$$\frac{\text{SS(B)}/(4-1)}{\text{SS(E)}/[(3-1)(4-1)]} \geq 9.78 = F_{0.01}(3,6).$$

We have

$$\text{SS(B)} = 3[(25-27)^2 + (26-27)^2 + (29-27)^2 + (28-27)^2] = 30;$$

$$\text{SS(E)} = (26-29-25+27)^2 + (24-26-25+27)^2 + \cdots$$

$$+ (26-26-28+27)^2 = 4.$$

Hence, the computed F is

$$\frac{30/3}{4/6} = 15 > 9.78,$$

and the hypothesis H_B is rejected. That is, the gasolines seem to give different performances (at least with these three cars).

The information for this example is summarized in Table 9.4-3. ■

In a two-way classification problem, particular combinations of the two factors might interact differently from what is expected from the additive model. For instance, in Example 9.4-1, gasoline 3 seemed to be the best gasoline and car 1 the best car; however, it sometimes happens that the two best do not "mix" well and the joint performance is poor. That is, there might be a strange interaction between this combination of car and gasoline, and accordingly, the joint performance is not as good as expected. Sometimes it happens that we get good results from a combination of some of the poorer levels of each factor. This phenomenon is called interaction, and it frequently occurs in practice (e.g., in chemistry). In order to test for possible interaction, we shall consider a two-way classification problem in which $c > 1$ independent observations per cell are taken.

Assume that X_{ijk}, $i = 1,2,\ldots,a$; $j = 1,2,\ldots,b$; and $k = 1,2,\ldots,c$, are $n = abc$ random variables that are mutually independent and have normal distributions with a common, but unknown, variance σ^2. The mean of each X_{ijk}, $k = 1,2,\ldots,c$, is $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, where $\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$, $\sum_{i=1}^{a} \gamma_{ij} = 0$, and $\sum_{j=1}^{b} \gamma_{ij} = 0$. The parameter γ_{ij} is called the **interaction** associated with cell (i,j). That is, the interaction between the ith level of one classification and the jth level of

Table 9.4-3 ANOVA table for gas mileage data

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F	p-value
Row (A)	24	2	12	18	0.003
Column (B)	30	3	10	15	0.003
Error	4	6	2/3		
Total	58	11			

the other classification is γ_{ij}. In Exercise 9.4-6, the reader is asked to determine μ, α_i, β_j, and γ_{ij} for some given μ_{ij}.

To test the hypotheses that (a) the row effects are equal to zero, (b) the column effects are equal to zero, and (c) there is no interaction, we shall again partition the total sum of squares into several components. Letting

$$\overline{X}_{ij\cdot} = \frac{1}{c}\sum_{k=1}^{c} X_{ijk},$$

$$\overline{X}_{i\cdot\cdot} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk},$$

$$\overline{X}_{\cdot j\cdot} = \frac{1}{ac}\sum_{i=1}^{a}\sum_{k=1}^{c} X_{ijk},$$

$$\overline{X}_{\cdots} = \frac{1}{abc}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk},$$

we have

$$\begin{aligned}
\text{SS(TO)} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk}-\overline{X}_{\cdots})^2 \\
&= bc\sum_{i=1}^{a}(\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdots})^2 + ac\sum_{j=1}^{b}(\overline{X}_{\cdot j\cdot}-\overline{X}_{\cdots})^2 \\
&\quad + c\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{ij\cdot}-\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdot j\cdot}+\overline{X}_{\cdots})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk}-\overline{X}_{ij\cdot})^2 \\
&= \text{SS(A)} + \text{SS(B)} + \text{SS(AB)} + \text{SS(E)},
\end{aligned}$$

where SS(A) is the row sum of squares, or the sum of squares among levels of factor A; SS(B) is the column sum of squares, or the sum of squares among levels of factor B; SS(AB) is the interaction sum of squares; and SS(E) is the error sum of squares. Again, we can show that the cross-product terms sum to zero.

To consider the joint distribution of SS(A), SS(B), SS(AB), and SS(E), let us assume that all the means equal the same value μ. Of course, we know that $\text{SS(TO)}/\sigma^2$ is $\chi^2(abc-1)$. Also, by letting the n_i of Section 9.3 equal bc and ac, respectively, we know that $\text{SS(A)}/\sigma^2$ and $\text{SS(B)}/\sigma^2$ are $\chi^2(a-1)$ and $\chi^2(b-1)$. Moreover,

$$\frac{\sum_{k=1}^{c}(X_{ijk}-\overline{X}_{ij\cdot})^2}{\sigma^2}$$

is $\chi^2(c-1)$; hence, $\text{SS(E)}/\sigma^2$ is the sum of ab independent chi-square variables such as this and thus is $\chi^2[ab(c-1)]$. Of course $\text{SS(AB)} \geq 0$; so, according to Theorem 9.3-1, $\text{SS(A)}/\sigma^2$, $\text{SS(B)}/\sigma^2$, $\text{SS(AB)}/\sigma^2$, and $\text{SS(E)}/\sigma^2$ are mutually independent chi-square variables with $a-1$, $b-1$, $(a-1)(b-1)$, and $ab(c-1)$ degrees of freedom, respectively.

To test the hypotheses concerning row, column, and interaction effects, we form F statistics in which the numerators are affected by deviations from the respective hypotheses, whereas the denominator is a function of SS(E), whose distribution

depends only on the value of σ^2 and not on the values of the cell means. Hence, SS(E) acts as our measuring stick here.

The statistic for testing the hypothesis

$$H_{AB}: \gamma_{ij} = 0,\ i = 1,2,\ldots,a;\ j = 1,2,\ldots,b,$$

against all alternatives is

$$F_{AB} = \frac{c\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X}_{\cdots})^2/[\sigma^2(a-1)(b-1)]}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \overline{X}_{ij\cdot})^2/[\sigma^2 ab(c-1)]}$$

$$= \frac{\text{SS(AB)}/[(a-1)(b-1)]}{\text{SS(E)}/[ab(c-1)]},$$

which has an F distribution with $(a-1)(b-1)$ and $ab(c-1)$ degrees of freedom when H_{AB} is true. If the computed $F_{AB} \geq F_{\alpha}[(a–1)(b–1), ab(c–1)]$, we reject H_{AB} and say that there is a difference among the means, since there seems to be interaction. Most statisticians do *not* proceed to test row and column effects if H_{AB} is rejected.

The statistic for testing the hypothesis

$$H_A: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

against all alternatives is

$$F_A = \frac{bc\sum_{i=1}^{a}(\overline{X}_{i\cdot\cdot} - \overline{X}_{\cdots})^2/[\sigma^2(a-1)]}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \overline{X}_{ij\cdot})^2/[\sigma^2 ab(c-1)]} = \frac{\text{SS(A)}/(a-1)}{\text{SS(E)}/[ab(c-1)]},$$

which has an F distribution with $a-1$ and $ab(c-1)$ degrees of freedom when H_A is true. The statistic for testing the hypothesis

$$H_B: \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

against all alternatives is

$$F_B = \frac{ac\sum_{j=1}^{b}(\overline{X}_{\cdot j\cdot} - \overline{X}_{\cdots})^2/[\sigma^2(b-1)]}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \overline{X}_{ij\cdot})^2/[\sigma^2 ab(c-1)]} = \frac{\text{SS(B)}/(b-1)}{\text{SS(E)}/[ab(c-1)]},$$

which has an F distribution with $b-1$ and $ab(c-1)$ degrees of freedom when H_B is true. Each of these hypotheses is rejected if the observed value of F is greater than a given constant that is selected to yield the desired significance level.

Table 9.4-4 is the ANOVA table that summarizes the information needed for these tests of hypotheses.

Example 9.4-2

Consider the following experiment: One hundred eight people were randomly divided into 6 groups with 18 people in each group. Each person was given sets of three numbers to add. The three numbers were either in a "down array" or an "across array," representing the two levels of factor A. The levels of factor B are determined by the number of digits in the numbers to be added: one-digit, two-digit, or three-digit numbers. Table 9.4-5 illustrates this experiment with a sample problem for each cell; note, however, that an individual person works problems only of one of these types. Each person was placed in one of the six groups and was told to work as many problems as possible in 90 seconds. The measurement that was recorded was the average number of problems worked correctly in two trials.

Table 9.4-4 Two-way ANOVA table, c observations per cell

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F
Factor A (row)	SS(A)	$a-1$	$\text{MS(A)} = \dfrac{\text{SS(A)}}{a-1}$	$\dfrac{\text{MS(A)}}{\text{MS(E)}}$
Factor B (column)	SS(B)	$b-1$	$\text{MS(B)} = \dfrac{\text{SS(B)}}{b-1}$	$\dfrac{\text{MS(B)}}{\text{MS(E)}}$
Factor AB (interaction)	SS(AB)	$(a-1)(b-1)$	$\text{MS(AB)} = \dfrac{\text{SS(AB)}}{(a-1)(b-1)}$	$\dfrac{\text{MS(AB)}}{\text{MS(E)}}$
Error	SS(E)	$ab(c-1)$	$\text{MS(E)} = \dfrac{\text{SS(E)}}{ab(c-1)}$	
Total	SS(TO)	$abc-1$		

Table 9.4-5 Illustration of arrays for numbers of digits

	Number of Digits		
Type of Array	1	2	3
Down	5	25	259
	3	69	567
	8	37	130
Across	5 + 3 + 8 =	25 + 69 + 37 =	259 + 567 + 130 =

Whenever this many subjects are used, a computer becomes an invaluable tool. A computer program provided the summary shown in Table 9.4-6 of the sample means of the rows, the columns, and the six cells. Each cell mean is the average for 18 people.

Simply considering these means, we can see clearly that there is a column effect: It is not surprising that it is easier to add one-digit than three-digit numbers.

The most interesting feature of these results is that they show the possibility of interaction. The largest cell mean occurs for those adding one-digit numbers in an across array. Note, however, that for two- and three-digit numbers, the down arrays have larger means than the across arrays.

The computer provided the ANOVA table given in Table 9.4-7. The number of degrees of freedom for SS(E) is not in our F table in Appendix B. However, the rightmost column, obtained from the computer printout, provides the p-value of each test, namely, the probability of obtaining an F as large as or larger than the calculated F ratio. Note, for example, that, to test for interaction, $F = 5.51$ and the p-value is 0.0053. Thus, the hypothesis of no interaction would be rejected at the $\alpha = 0.05$ or $\alpha = 0.01$ significance level, but it would not be rejected with $\alpha = 0.001$.

Table 9.4-6 Cell, row, and column means for adding numbers

	Number of Digits			
Type of Array	1	2	3	Row Means
Down	23.806	10.694	6.278	13.593
Across	26.056	6.750	3.944	12.250
Column means	24.931	8.722	5.111	

Table 9.4-7 ANOVA table for adding numbers

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Factor A (array)	48.678	1	48.669	2.885	0.0925
Factor B (number of digits)	8022.73	2	4011.363	237.778	<0.0001
Interaction	185.92	2	92.961	5.510	0.0053
Error	1720.76	102	16.870		
Total	9978.08	107			

Exercises

(In some of the exercises that follow, we must make assumptions, such as normal distributions with equal variances.)

9.4-1. For the data given in Example 9.4-1, test the hypothesis H_A: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ against all alternatives at the 5% significance level.

9.4-2. With $a = 3$ and $b = 4$, find μ, α_i, and β_j if μ_{ij}, $i = 1, 2, 3$ and $j = 1, 2, 3, 4$, are given by

6	3	7	8
10	7	11	12
8	5	9	10

Note that in an "additive" model such as this one, one row (column) can be determined by adding a constant value to each of the elements of another row (column).

9.4-3. We wish to compare compressive strengths of concrete corresponding to $a = 3$ different drying methods (treatments). Concrete is mixed in batches that are just large enough to produce three cylinders. Although care is taken to achieve uniformity, we expect some variability among the $b = 5$ batches used to obtain the following compressive strengths (there is little reason to suspect interaction; hence, only one observation is taken in each cell):

	Batch				
Treatment	B_1	B_2	B_3	B_4	B_5
A_1	52	47	44	51	42
A_2	60	55	49	52	43
A_3	56	48	45	44	38

(a) Use the 5% significance level and test H_A: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ against all alternatives.

(b) Use the 5% significance level and test H_B: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ against all alternatives. (See Ledolter and Hogg in References.)

9.4-4. Show that the cross-product terms formed from $(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})$, $(\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})$, and $(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot})$ sum to zero, $i = 1, 2, \ldots a$ and $j = 1, 2, \ldots, b$. HINT: For example, write

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot})$$

$$= \sum_{j=1}^{b}(\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})\sum_{i=1}^{a}[(X_{ij} - \overline{X}_{\cdot j}) - (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})]$$

and sum each term in the inner summation, as grouped here, to get zero.

9.4-5. A psychology student was interested in testing how food consumption by rats would be affected by a particular drug. She used two levels of one attribute, namely, drug and placebo, and four levels of a second attribute, namely, male (M), castrated (C), female (F), and ovariectomized (O). For each cell, she observed five rats. The amount of food consumed in grams per 24 hours is listed in the following table:

	M	C	F	O
Drug	22.56	16.54	18.58	18.20
	25.02	24.64	15.44	14.56
	23.66	24.62	16.12	15.54
	17.22	19.06	16.88	16.82
	22.58	20.12	17.58	14.56
Placebo	25.64	22.50	17.82	19.74
	28.84	24.48	15.76	17.48
	26.00	25.52	12.96	16.46
	26.02	24.76	15.00	16.44
	23.24	20.62	19.54	15.70

(a) Use the 5% significance level and test H_{AB}: $\gamma_{ij} = 0$, $i = 1, 2, j = 1, 2, 3, 4$.

(b) Use the 5% significance level and test H_A: $\alpha_1 = \alpha_2 = 0$.

(c) Use the 5% significance level and test H_B: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

(d) How could you modify this model so that there are three attributes of classification, each with two levels?

9.4-6. With $a = 3$ and $b = 4$, find μ, α_i, β_j, and γ_{ij} if μ_{ij}, $i = 1, 2, 3$ and $j = 1, 2, 3, 4$, are given by

6	7	7	12
10	3	11	8
8	5	9	10

Note the difference between the layout here and that in Exercise 9.4-2. Does the interaction help explain the difference?

9.4-7. In order to test whether four brands of gasoline give equal performance in terms of mileage, each of three cars was driven with each of the four brands of gasoline. Then each of the (3)(4) = 12 possible combinations was repeated four times. The number of miles per gallon for each of the four repetitions in each cell is recorded in the following table:

	Brand of Gasoline							
Car	1		2		3		4	
1	31.0	24.9	26.3	30.0	25.8	29.4	27.8	27.3
	26.2	28.8	25.2	31.6	24.5	24.8	28.2	30.4
2	30.6	29.5	25.5	26.8	26.6	23.7	28.1	27.1
	30.8	28.9	27.4	29.4	28.2	26.1	31.5	29.1
3	24.2	23.1	27.4	28.1	25.2	26.7	26.3	26.4
	26.8	27.4	26.4	26.9	27.7	28.1	27.9	28.8

Test the hypotheses H_{AB}: no interaction, H_A: no row effect, and H_B: no column effect, each at the 5% significance level.

9.4-8. There is another way of looking at Exercise 9.3-6, namely, as a two-factor analysis-of-variance problem with the levels of gender being female and male, the levels of age being less than 50 and at least 50, and the measurement for each subject being their cholesterol level. The data would then be set up as follows:

	Age	
Gender	<50	≥50
	221	262
	213	193
	202	224
Female	183	201
	185	161
	197	178
	162	265
	271	192
	192	253
	189	248
Male	209	278
	227	232
	236	267
	142	289

(a) Test H_{AB}: $\gamma_{ij} = 0, i = 1,2; j = 1,2$ (no interaction).

(b) Test H_A: $\alpha_1 = \alpha_2 = 0$ (no row effect).

(c) Test H_B: $\beta_1 = \beta_2 = 0$ (no column effect).

Use a 5% significance level for each test.

9.4-9. Ledolter and Hogg (see References) report that volunteers who had a smoking history classified as heavy, moderate, and nonsmoker were accepted until nine men were in each category. Three men in each category were randomly assigned to each of the following three stress tests: bicycle ergometer, treadmill, and step tests. The time until maximum oxygen uptake was recorded in minutes as follows:

	Test		
Smoking History	Bicycle	Treadmill	Step Test
Nonsmoker	12.8, 13.5, 11.2	16.2, 18.1, 17.8	22.6, 19.3, 18.9
Moderate	10.9, 11.1, 9.8	15.5, 13.8, 16.2	20.1, 21.0, 15.9
Heavy	8.7, 9.2, 7.5	14.7, 13.2, 8.1	16.2, 16.1, 17.8

(a) Analyze the results of this experiment. Obtain the ANOVA table and test for main effects and interactions.

(b) Use box plots to compare the data graphically.

9.5* GENERAL FACTORIAL AND 2^k FACTORIAL DESIGNS

In Section 9.4, we studied two-factor experiments in which the A factor is performed at a levels and the B factor has b levels. Without replications, we need ab-level combinations, and with c replications with each of these combinations, we need a total of abc experiments.

Let us now consider a situation with three factors—say, A, B, and C, with a, b, and c levels, respectively. Here there are a total of abc-level combinations, and if, at each of these combinations, we have d replications, there is a need for $abcd$ experiments. Once these experiments are run, in some random order, and the data collected, there are computer programs available to calculate the entries in the ANOVA table, as in Table 9.5-1.

The main effects (A, B, and C) and the two-factor interactions (AB, AC, and BC) have the same interpretations as in the two-factor ANOVA. The three-factor interaction represents that part of the model for the means μ_{ijh}, $i = 1,2,\ldots,a$; $j = 1,2,\ldots,b$; $h = 1,2,\ldots,c$, that cannot be explained by a model including only the main effects and two-factor interactions. In particular, if, for each fixed h, the "plane" created by μ_{ijh} is "parallel" to the "plane" created by every other fixed h, then the three-factor interaction is equal to zero. Usually, higher-order interactions tend to be small.

Table 9.5-1 ANOVA table

Source	SS	d.f.	MS	F
A	SS(A)	$a-1$	MS(A)	MS(A)/MS(E)
B	SS(B)	$b-1$	MS(B)	MS(B)/MS(E)
C	SS(C)	$c-1$	MS(C)	MS(C)/MS(E)
AB	SS(AB)	$(a-1)(b-1)$	MS(AB)	MS(AB)/MS(E)
AC	SS(AC)	$(a-1)(c-1)$	MS(AC)	MS(AC)/MS(E)
BC	SS(BC)	$(b-1)(c-1)$	MS(BC)	MS(BC)/MS(E)
ABC	SS(ABC)	$(a-1)(b-1)(c-1)$	MS(ABC)	MS(ABC)/MS(E)
Error	SS(E)	$abc(d-1)$	MS(E)	
Total	SS(TO)	$abcd-1$		

In the testing sequence, we test the three-factor interaction first by checking to see whether or not

$$\text{MS(ABC)/MS(E)} \geq F_{\alpha}[(a-1)(b-1)(c-1), abc(d-1)].$$

If this inequality holds, the ABC interaction is significant at the α level. We would then not continue testing the two-factor interactions and the main effects with those F values, but analyze the data otherwise. For example, for each fixed h, we could look at a two-factor ANOVA for factors A and B. Of course, if the inequality does not hold, we next check the two-factor interactions with the appropriate F values. If these are not significant, we check the main effects, A, B, and C.

Factorial analyses with three or more factors require many experiments, particularly if each factor has several levels. Often, in the health, social, and physical sciences, experimenters want to consider several factors (maybe as many as 10, 20, or even hundreds), and they cannot afford to run that many experiments. This is particularly true with preliminary or screening investigations, in which they want to detect the factors that seem most important. In these cases, they often consider factorial experiments such that each of k factors is run at just two levels, frequently without replication. We consider only this situation, although the reader should recognize that it has many variations. In particular, there are methods for investigating only *fractions of these* 2^k *designs*. The reader interested in more information should refer to a good book on the design of experiments, such as that by Box, Hunter, and Hunter (see References). Many statisticians in industry believe that these statistical methods are the most useful in improving product and process designs. Hence, this is clearly an extremely important topic, as many industries are greatly concerned about the quality of their products.

In factorial experiments in which each of the k factors is considered at only two levels, those levels are selected at some reasonable low and high values. That is, with the help of someone in the field, the typical range of each factor is considered. For instance, if we are considering baking temperatures in the range from 300° to 375°, a representative low is selected—say, 320°—and a representative high is selected—say 355°. There is no formula for these selections, and someone familiar with the

experiment would help make them. Often, it happens that only two different types of a material (e.g., fabric) are considered and one is called low and the other high.

Thus, we select a low and high for each factor and code them as -1 and $+1$ or, more simply, $-$ and $+$, respectively. We give three 2^k designs, for $k = 2, 3$, and 4, in standard order in Tables 9.5-2, 9.5-3, and 9.5-4, respectively. From these three tables, we can easily note what is meant by standard order. The A column starts with a minus sign and then the sign alternates. The B column begins with two minus signs and then the signs alternate in blocks of two. The C column has 4 minus signs and then 4 plus signs, and so on. The D column starts with 8 minus signs and then 8 plus signs. It is easy to extend this idea to 2^k designs, where $k \geq 5$. To illustrate, under the E column in a 2^5 design, we have 16 minus signs followed by 16 plus signs, which together account for the 32 experiments.

To be absolutely certain what these runs mean, consider run number 12 in Table 9.5-4: A is set at its high level, B at its high, C at its low, and D at its high level. The value X_{12} is the random observation resulting from this one combination of these four settings. It must be emphasized that the runs are not necessarily performed in the order $1, 2, 3, \ldots, 2^k$; in fact, they should be performed in a random

Table 9.5-2 2^2 Design

	2^2 Design		
Run	A	B	Observation
1	–	–	X_1
2	+	–	X_2
3	–	+	X_3
4	+	+	X_4

Table 9.5-3 2^3 Design

	2^3 Design			
Run	A	B	C	Observation
1	–	–	–	X_1
2	+	–	–	X_2
3	–	+	–	X_3
4	+	+	–	X_4
5	–	–	+	X_5
6	+	–	+	X_6
7	–	+	+	X_7
8	+	+	+	X_8

Table 9.5-4 2^4 Design

	2^4 Design				
Run	A	B	C	D	Observation
1	−	−	−	−	X_1
2	+	−	−	−	X_2
3	−	+	−	−	X_3
4	+	+	−	−	X_4
5	−	−	+	−	X_5
6	+	−	+	−	X_6
7	−	+	+	−	X_7
8	+	+	+	−	X_8
9	−	−	−	+	X_9
10	+	−	−	+	X_{10}
11	−	+	−	+	X_{11}
12	+	+	−	+	X_{12}
13	−	−	+	+	X_{13}
14	+	−	+	+	X_{14}
15	−	+	+	+	X_{15}
16	+	+	+	+	X_{16}

order if at all possible. That is, in a 2^3 design, we might perform the experiment in the order 3, 2, 8, 6, 5, 1, 4, 7 if this, in fact, was a random selection of a permutation of the first eight positive integers.

Once all 2^k experiments have been run, it is possible to consider the total sum of squares

$$\sum_{i=1}^{2^k} (X_i - \overline{X})^2$$

and decompose it very easily into $2^k - 1$ parts, which represent the respective measurements (estimators) of the k main effects, $\binom{k}{2}$ two-factor interactions, $\binom{k}{3}$ three-factor interactions, and so on, until we have the one k-factor interaction. We illustrate this decomposition with the 2^3 design in Table 9.5-5. Note that column AB is found by formally multiplying the elements of column A by the corresponding ones in B. Likewise, AC is found by multiplying the elements of column A by the corresponding ones in column C, and so on, until column ABC is the product of the corresponding elements of columns A, B, and C. Next, we construct

Table 9.5-5 2^3 Design decomposition

	2^3 Design							
Run	A	B	C	AB	AC	BC	ABC	Observation
1	−	−	−	+	+	+	−	X_1
2	+	−	−	−	−	+	+	X_2
3	−	+	−	−	+	−	+	X_3
4	+	+	−	+	−	−	−	X_4
5	−	−	+	+	−	−	+	X_5
6	+	−	+	−	+	−	−	X_6
7	−	+	+	−	−	+	−	X_7
8	+	+	+	+	+	+	+	X_8

seven linear forms, using these seven columns of signs with the corresponding observations. The resulting measures (estimates) of the main effects (A, B, C), the two-factor interactions (AB, AC, BC), and the three-factor interaction (ABC) are then found by dividing the linear forms by $2^k = 2^3 = 8$. (Some statisticians divide by $2^{k-1} = 2^{3-1} = 4$.) These are denoted by

$$
\begin{aligned}
[\mathrm{A}] &= (-X_1 + X_2 - X_3 + X_4 - X_5 + X_6 - X_7 + X_8)/8,\\
[\mathrm{B}] &= (-X_1 - X_2 + X_3 + X_4 - X_5 - X_6 + X_7 + X_8)/8,\\
[\mathrm{C}] &= (-X_1 - X_2 - X_3 - X_4 + X_5 + X_6 + X_7 + X_8)/8,\\
[\mathrm{AB}] &= (+X_1 - X_2 - X_3 + X_4 + X_5 - X_6 - X_7 + X_8)/8,\\
[\mathrm{AC}] &= (+X_1 - X_2 + X_3 - X_4 - X_5 + X_6 - X_7 + X_8)/8,\\
[\mathrm{BC}] &= (+X_1 + X_2 - X_3 - X_4 - X_5 - X_6 + X_7 + X_8)/8,\\
[\mathrm{ABC}] &= (-X_1 + X_2 + X_3 - X_4 + X_5 - X_6 - X_7 + X_8)/8.
\end{aligned}
$$

With assumptions of normality, mutual independence, and common variance σ^2, under the overall null hypothesis of the equality of all the means, each of these measures has a normal distribution with mean zero and variance $\sigma^2/8$ (in general, $\sigma^2/2^k$). This implies that the square of each measure divided by $\sigma^2/8$ is $\chi^2(1)$. Moreover, it can be shown (see Exercise 9.5-2) that

$$
\sum_{i=1}^{8} \left(X_i - \overline{X}\right)^2 = 8\left([\mathrm{A}]^2 + [\mathrm{B}]^2 + [\mathrm{C}]^2 + [\mathrm{AB}]^2 + [\mathrm{AC}]^2 + [\mathrm{BC}]^2 + [\mathrm{ABC}]^2\right).
$$

So, by Theorem 9.3-1, the terms on the right-hand side, divided by σ^2, are mutually independent random variables, each being $\chi^2(1)$. While it requires a little more theory, it follows that the linear forms [A], [B], [C], [AB], [AC], [BC], and [ABC] are mutually independent $N(0, \sigma^2/8)$ random variables.

Since we have assumed that we have not run any replications, how can we obtain an estimate of σ^2 to see if any of the main effects or interactions are significant? To help us, we fall back on the use of a q–q plot because, under the overall null hypothesis, those seven measures are mutually independently, normally distributed variables with the same mean and variance. Thus, a q–q plot of the normal percentiles against the corresponding ordered values of the measures should be about on a straight line if, in fact, the null hypothesis is true. If one of these points is "out of line," we might believe that the overall null hypothesis is not true and that the effect associated with the factor represented with that point is significant. It is possible that two or three points might be out of line; then all corresponding effects (main or interaction) should be investigated. Clearly, this is not a formal test, but it has been extremely successful in practice.

As an illustration, we use the data from an experiment designed to evaluate the effects of laundering on a certain fire-retardant treatment for fabrics. These data, somewhat modified, were taken from *Experimental Statistics, National Bureau of Standards Handbook 91*, by Mary G. Natrella (Washington, DC: U.S. Government Printing Office, 1963). Factor A is the type of fabric (sateen or monk's cloth), factor B corresponds to two different fire-retardant treatments, and factor C describes the laundering conditions (no laundering, after one laundering). The observations are

Table 9.5-6 Seven measures ordered

Identity of Effect	Ordered Effect	Percentile	Percentile from $N(0,1)$
[A]	−8.06	12.5	−1.15
[AB]	−2.19	25.0	−0.67
[AC]	−0.31	37.5	−0.32
[ABC]	0.31	50.0	0.00
[C]	0.56	62.5	0.32
[BC]	0.81	75.0	0.67
[B]	1.56	87.5	1.15

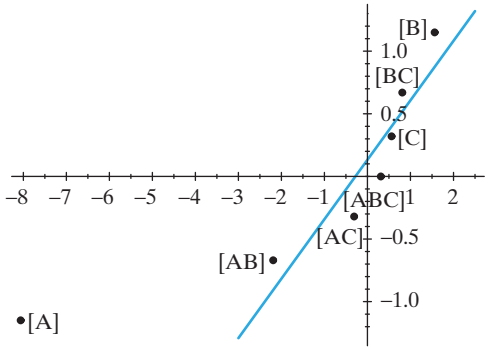

Figure 9.5-1 A q–q plot of normal percentiles versus estimated effects

inches burned, measured on a standard-size fabric after a flame test. They are as follows, in standard order:

$$x_1 = 41.0, \quad x_2 = 30.5, \quad x_3 = 47.5, \quad x_4 = 27.0,$$
$$x_5 = 39.5, \quad x_6 = 26.5, \quad x_7 = 48.0, \quad x_8 = 27.5.$$

Thus, the measures of the effects are

$$[\text{A}] = (-41.0 + 30.5 - 47.5 + 27.0 - 39.5 + 26.5 - 48.0 + 27.5)/8 = -8.06,$$
$$[\text{B}] = (-41.0 - 30.5 + 47.5 + 27.0 - 39.5 - 26.5 + 48.0 + 27.5)/8 = 1.56,$$
$$[\text{C}] = (-41.0 - 30.5 - 47.5 - 27.0 + 39.5 + 26.5 + 48.0 + 27.5)/8 = 0.56,$$
$$[\text{AB}] = (+41.0 - 30.5 - 47.5 + 27.0 + 39.5 - 26.5 - 48.0 + 27.5)/8 = -2.19,$$
$$[\text{AC}] = (+41.0 - 30.5 + 47.5 - 27.0 - 39.5 + 26.5 - 48.0 + 27.5)/8 = -0.31,$$
$$[\text{BC}] = (+41.0 + 30.5 - 47.5 - 27.0 - 39.5 - 26.5 + 48.0 + 27.5)/8 = 0.81,$$
$$[\text{ABC}] = (-41.0 + 30.5 + 47.5 - 27.0 + 39.5 - 26.5 - 48.0 + 27.5)/8 = 0.31.$$

In Table 9.5-6, we order these seven measures, determine their percentiles, and find the corresponding percentiles of the standard normal distribution.

The q–q plot is given in Figure 9.5-1. Each point has been identified with its effect. A straight line fits six of those points reasonably well, but the point associated with $[\text{A}] = -8.06$ is far from this straight line. Hence, the main effect of factor A (the type of fabric) seems to be significant. It is interesting to note that the laundering factor, C, does not seem to be a significant factor.

Exercises

9.5-1. Write out a 2^2 design, displaying the A, B, and AB columns for the four runs.

(a) If X_1, X_2, X_3, and X_4 are the four observations for the respective runs in standard order, write out the three linear forms, [A], [B], and [AB], that measure the two main effects and the interaction. These linear forms should include the divisor $2^2 = 4$.

(b) Show that $\sum_{i=1}^{4} (X_i - \overline{X})^2 = 4([\text{A}]^2 + [\text{B}]^2 + [\text{AB}]^2)$.

(c) Under the null hypothesis that all the means are equal and with the usual assumptions (normality, mutual independence, and common variance), what can you say about the distributions of the expressions in (b) after each is divided by σ^2?

9.5-2. Show that, in a 2^3 design,

$$\sum_{i=1}^{8} (X_i - \overline{X})^2 = 8\left([\text{A}]^2 + [\text{B}]^2 + [\text{C}]^2 + [\text{AB}]^2 + [\text{AC}]^2 + [\text{BC}]^2 + [\text{ABC}]^2\right).$$

HINT: Since both the right and the left members of this equation are symmetric in the variables $X_1, X_2, \ldots, X_8$, it is necessary to show only that the corresponding coefficients of $X_1 X_i$, $i = 1, 2, \ldots, 8$, are the same in each member of the equation. Of course, recall that $\overline{X} = (X_1 + X_2 + \cdots + X_8)/8$.

9.5-3. Show that the unbiased estimator of the variance σ^2 from a sample of size $n = 2$ is one half of the square of the difference of the two observations. Thus, show that, if a 2^k design is replicated, say, with X_{i1} and X_{i2}, $i = 1, 2, \ldots, 2^k$, then the estimate of the common σ^2 is

$$\frac{1}{2^{k+1}} \sum_{i=1}^{2^k} (X_{i1} - X_{i2})^2 = \text{MS(E)}.$$

Under the usual assumptions, this equation implies that each of $2^k[\text{A}]^2/\text{MS(E)}$, $2^k[\text{B}]^2/\text{MS(E)}$, $2^k[\text{AB}]^2/\text{MS(E)}$, and so on has an $F(1, 2^k)$ distribution under the null

hypothesis. This approach, of course, would provide tests for the significance of the various effects, including interactions.

9.5-4. Ledolter and Hogg (see References) note that percent yields from a certain chemical reaction for changing temperature (factor A), reaction time (factor B), and concentration (factor C) are $x_1 = 79.7$, $x_2 = 74.3$, $x_3 = 76.7$, $x_4 = 70.0$, $x_5 = 84.0$, $x_6 = 81.3$, $x_7 = 87.3$, and $x_8 = 73.7$, in standard order with a 2^3 design.

(a) Estimate the main effects, the three two-factor interactions, and the three-factor interaction.

(b) Construct an appropriate q–q plot to see if any of these effects seem to be significantly larger than the others.

9.5-5. Box, Hunter, and Hunter (see References) studied the effects of catalyst charge (10 pounds = −1, 20 pounds = +1), temperature (220 °C = − 1, 240 °C = +1), pressure (50 psi = −1, 80 psi = +1), and concentration (10% = −1, 12% = +1) on percent conversion (X) of a certain chemical. The results of a 2^4 design, in standard order, are

$$x_1 = 71,\ x_2 = 61,\ x_3 = 90,\ x_4 = 82,\ x_5 = 68,\ x_6 = 61,$$
$$x_7 = 87,\ x_8 = 80,\ x_9 = 61,\ x_{10} = 50,\ x_{11} = 89,$$
$$x_{12} = 83,\ x_{13} = 59,\ x_{14} = 51,\ x_{15} = 85,\ x_{16} = 78.$$

(a) Estimate the main effects and the two-, three-, and four-factor interactions.

(b) Construct an appropriate q–q plot and assess the significance of the various effects.

9.6* TESTS CONCERNING REGRESSION AND CORRELATION

In Section 6.5, we considered the estimation of the parameters of a very simple regression curve, namely, a straight line. We can use confidence intervals for the parameters to test hypotheses about them. For example, with the same model as that in Section 6.5, we could test the hypothesis H_0: $\beta = \beta_0$ by using a t random variable that was used for a confidence interval with β replaced by β_0, namely,

$$T_1 = \frac{\widehat{\beta} - \beta_0}{\sqrt{\dfrac{n\widehat{\sigma^2}}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}}.$$

The null hypothesis, along with three possible alternative hypotheses, is given in Table 9.6-1; these tests are equivalent to stating that we reject H_0 if β_0 is not in certain confidence intervals. For example, the first test is equivalent to rejecting H_0 if β_0 is not in the one-sided confidence interval with lower bound

$$\widehat{\beta} - t_\alpha(n-2)\sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Often we let $\beta_0 = 0$ and test the hypothesis H_0: $\beta = 0$. That is, we test the null hypothesis that the slope is equal to zero.

Table 9.6-1 Tests about the slope of the regression line

H_0	H_1	Critical Region
$\beta = \beta_0$	$\beta > \beta_0$	$t_1 \geq t_\alpha(n-2)$
$\beta = \beta_0$	$\beta < \beta_0$	$t_1 \leq -t_\alpha(n-2)$
$\beta = \beta_0$	$\beta \neq \beta_0$	$\lvert t_1 \rvert \geq t_{\alpha/2}(n-2)$

Example 9.6-1 Let x equal a student's preliminary test score in a psychology course and y equal the same student's score on the final examination. With $n = 10$ students, we shall test H_0: $\beta = 0$ against H_1: $\beta \neq 0$. At the 0.01 significance level, the critical region is $|t_1| \geq t_{0.005}(8) = 3.355$. Using the data in Example 6.5-1, we find that the observed value of T_1 is

$$t_1 = \frac{0.742 - 0}{\sqrt{10(21.7709)/8(756.1)}} = \frac{0.742}{0.1897} = 3.911.$$

Thus, we reject H_0 and conclude that a student's score on the final examination is related to his or her preliminary test score. ■

We consider tests about the correlation coefficient ρ of a bivariate normal distribution. Let X and Y have a bivariate normal distribution. We know that if the correlation coefficient ρ is zero, then X and Y are independent random variables. Furthermore, the value of ρ gives a measure of the linear relationship between X and Y. We now give methods for using the sample correlation coefficient to test the hypothesis H_0: $\rho = 0$ and also to form a confidence interval for ρ.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote a random sample from a bivariate normal distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ. That is, the n pairs of (X, Y) are independent, and each pair has the same bivariate normal distribution. The **sample correlation coefficient** is

$$R = \frac{[1/(n-1)]\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{[1/(n-1)]\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{[1/(n-1)]\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{S_{XY}}{S_X S_Y}.$$

We note that

$$R\frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2} = \frac{[1/(n-1)]\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{[1/(n-1)]\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

is exactly the solution that we obtained for $\widehat{\beta}$ in Section 6.5 when the X-values were fixed at $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Let us consider these values fixed temporarily so that we are considering conditional distributions, given $X_1 = x_1, \ldots, X_n = x_n$. Moreover, if H_0: $\rho = 0$ is true, then $Y_1, Y_2, \ldots, Y_n$ are independent of $X_1, X_2, \ldots, X_n$ and $\beta = \rho\sigma_Y/\sigma_X = 0$. Under these conditions, the conditional distribution of

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

given that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, is $N[0, \sigma_Y^2/(n-1)s_x^2]$ when $s_x^2 > 0$. Moreover, recall from Section 6.5 that the conditional distribution of

$$\frac{\sum_{i=1}^{n}[Y_i - \overline{Y} - (S_{XY}/S_X^2)(X_i - \overline{X})]^2}{\sigma_Y^2} = \frac{(n-1)S_Y^2(1-R^2)}{\sigma_Y^2},$$

given that $X_1 = x_1, \ldots, X_n = x_n$, is $\chi^2(n-2)$ and is independent of $\widehat{\beta}$. (See Exercise 9.6-6.) Thus, when $\rho = 0$, the conditional distribution of

$$T = \frac{(RS_Y/S_X)/(\sigma_Y/\sqrt{n-1}\,S_X)}{\sqrt{[(n-1)\,S_Y^2\,(1-R^2)/\sigma_Y^2][1/(n-2)]}} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

is t with $n-2$ degrees of freedom. However, since the conditional distribution of T, given that $X_1 = x_1, \ldots, X_n = x_n$, does not depend on $x_1, x_2, \ldots, x_n$, the unconditional

distribution of T must be t with $n-2$ degrees of freedom, and T and $(X_1, X_2, \ldots, X_n)$ are independent when $\rho = 0$.

REMARK It is interesting to note that in the discussion about the distribution of T, the assumption that (X, Y) has a bivariate normal distribution can be relaxed. Specifically, if X and Y are independent and Y has a normal distribution, then T has a t distribution regardless of the distribution of X. Obviously, the roles of X and Y can be reversed in all of this development. In particular, if X and Y are independent, then T and $Y_1, Y_2, \ldots, Y_n$ are also independent. ■

Now T can be used to test H_0: $\rho = 0$. If the alternative hypothesis is H_1: $\rho > 0$, we would use the critical region defined by the observed $T \geq t_\alpha(n-2)$, since large T implies large R. Obvious modifications would be made for the alternative hypotheses H_1: $\rho < 0$ and H_1: $\rho \neq 0$, the latter leading to a two-sided test.

Using the pdf $h(t)$ of T, we can find the distribution function and pdf of R when $-1 < r < 1$, provided that $\rho = 0$:

$$\begin{aligned} G(r) = P(R \leq r) &= P\left(T \leq \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right) \\ &= \int_{-\infty}^{r\sqrt{n-2}/\sqrt{1-r^2}} h(t)\,dt \\ &= \int_{-\infty}^{r\sqrt{n-2}/\sqrt{1-r^2}} \frac{\Gamma[(n-1)/2]}{\Gamma(1/2)\,\Gamma[(n-2)/2]} \frac{1}{\sqrt{n-2}} \left(1 + \frac{t^2}{n-2}\right)^{-(n-1)/2} dt. \end{aligned}$$

The derivative of $G(r)$, with respect to r, is (see Appendix D.4)

$$g(r) = h\left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right) \frac{d(r\sqrt{n-2}/\sqrt{1-r^2})}{dr},$$

which equals

$$g(r) = \frac{\Gamma[(n-1)/2]}{\Gamma(1/2)\,\Gamma[(n-2)/2]} (1-r^2)^{(n-4)/2}, \qquad -1 < r < 1.$$

Thus, to test the hypothesis H_0: $\rho = 0$ against the alternative hypothesis H_1: $\rho \neq 0$ at a significance level α, select either a constant $r_{\alpha/2}(n-2)$ or a constant $t_{\alpha/2}(n-2)$ so that

$$\alpha = P(|R| \geq r_{\alpha/2}(n-2); H_0) = P(|T| \geq t_{\alpha/2}(n-2); H_0),$$

depending on the availability of R or T tables.

It is interesting to graph the pdf of R. Note in particular that if $n = 4$, $g(r) = 1/2$, $-1 < r < 1$, and if $n = 6$, $g(r) = (3/4)(1-r^2)$, $-1 < r < 1$. The graphs of the pdf of R when $n = 8$ and when $n = 14$ are given in Figure 9.6-1. Recall that this is the pdf of R when $\rho = 0$. As n increases, R is more likely to equal values close to 0.

Table IX in Appendix B lists selected values of the distribution function of R when $\rho = 0$. For example, if $n = 8$, then the number of degrees of freedom is 6 and $P(R \leq 0.7887) = 0.99$. Also, if $\alpha = 0.10$, then $r_{\alpha/2}(6) = r_{0.05}(6) = 0.6215$. [See Figure 9.6-1(a).]

It is also possible to obtain an approximate test of size α by using the fact that

$$W = \frac{1}{2} \ln \frac{1+R}{1-R}$$

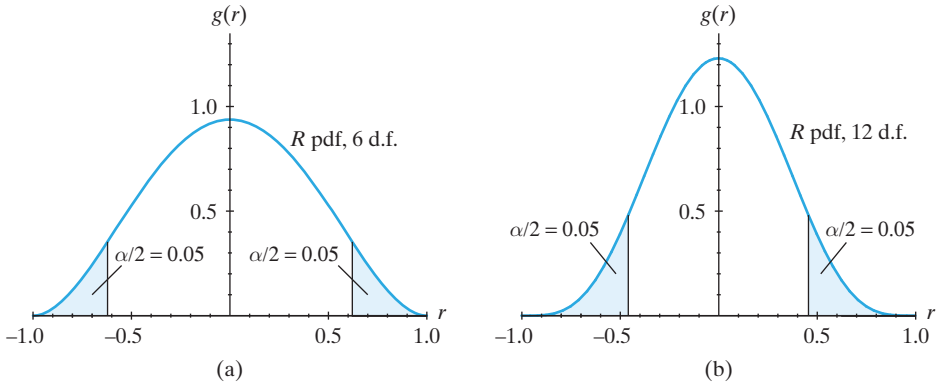

Figure 9.6-1 R pdfs when $n = 8$ and $n = 14$

has an approximate normal distribution with mean $(1/2)\ln[(1+\rho)/(1-\rho)]$ and variance $1/(n-3)$. We accept this statement without proof. (See Exercise 9.6-8.) Thus, a test of H_0: $\rho = \rho_0$ can be based on the statistic

$$Z = \frac{\frac{1}{2}\ln\frac{1+R}{1-R} - \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}}{\sqrt{\frac{1}{n-3}}},$$

which has a distribution that is approximately $N(0,1)$ under H_0. Notice that this approximate size-α test can be used to test a null hypothesis specifying a nonzero population correlation coefficient, whereas the exact size-α test may be used only in conjunction with the null hypothesis H_0: $\rho = 0$. Also, notice that the sample size must be at least $n = 4$ for the approximate test, but $n = 3$ is sufficient for the exact test.

Example 9.6-2

We would like to test the hypothesis H_0: $\rho = 0$ against H_1: $\rho \neq 0$ at an $\alpha = 0.05$ significance level. A random sample of size 18 from a bivariate normal distribution yielded a sample correlation coefficient of $r = 0.35$. From Table XI in Appendix B, since $0.35 < 0.4683$, H_0 is accepted (not rejected) at an $\alpha = 0.05$ significance level. Using the t distribution, we would reject H_0 if $|t| \geq 2.120 = t_{0.025}(16)$. Since

$$t = \frac{0.35\sqrt{16}}{\sqrt{1-(0.35)^2}} = 1.495,$$

H_0 is not rejected. If we had used the normal approximation for Z, H_0 would be rejected if $|z| \geq 1.96$. Because

$$z = \frac{(1/2)\ln[(1+0.35)/(1-0.35)] - 0}{\sqrt{1/(18-3)}} = 1.415,$$

H_0 is not rejected. ■

To develop an approximate $100(1-\alpha)\%$ confidence interval for ρ, we use the normal approximation for the distribution of Z. Thus, we select a constant $c = z_{\alpha/2}$ from Table V in Appendix B so that

$$P\left(-c \le \frac{(1/2)\ln[(1+R)/(1-R)] - (1/2)\ln[(1+\rho)/(1-\rho)]}{\sqrt{1/(n-3)}} \le c\right) \approx 1-\alpha.$$

After several algebraic manipulations, this formula becomes

$$P\left(\frac{1+R-(1-R)\exp(2c/\sqrt{n-3})}{1+R+(1-R)\exp(2c/\sqrt{n-3})} \le \rho \le \frac{1+R-(1-R)\exp(-2c/\sqrt{n-3})}{1+R+(1-R)\exp(-2c/\sqrt{n-3})}\right) \approx 1-\alpha.$$

Example 9.6-3 Suppose that a random sample of size 12 from a bivariate normal distribution yielded a correlation coefficient of $r = 0.6$. An approximate 95% confidence interval for ρ would be

$$\left[\frac{1+0.6-(1-0.6)\exp\left(\frac{2(1.96)}{3}\right)}{1+0.6+(1-0.6)\exp\left(\frac{2(1.96)}{3}\right)}, \frac{1+0.6-(1-0.6)\exp\left(\frac{-2(1.96)}{3}\right)}{1+0.6+(1-0.6)\exp\left(\frac{-2(1.96)}{3}\right)}\right]$$

$$= [0.040, 0.873].$$

If the sample size had been $n = 39$ and $r = 0.6$, the approximate 95% confidence interval would have been $[0.351, 0.770]$. ■

Exercises

(In some of the exercises that follow, we must make assumptions of normal distributions with the usual notation.)

9.6-1. For the data given in Exercise 6.5-3, use a t test to test H_0: $\beta = 0$ against H_1: $\beta > 0$ at the $\alpha = 0.025$ significance level.

9.6-2. For the data given in Exercise 6.5-4, use a t test to test H_0: $\beta = 0$ against H_1: $\beta > 0$ at the $\alpha = 0.025$ significance level.

9.6-3. A random sample of size $n = 27$ from a bivariate normal distribution yielded a sample correlation coefficient of $r = -0.45$. Would the hypothesis H_0: $\rho = 0$ be rejected in favor of H_1: $\rho \neq 0$ at an $\alpha = 0.05$ significance level?

9.6-4. In bowling, it is often possible to score well in the first game and then bowl poorly in the second game, or vice versa. The following six pairs of numbers give the scores of the first and second games bowled by the same person on six consecutive Tuesday evenings:

Game 1:	170	190	200	183	187	178
Game 2:	197	178	150	176	205	153

Assume a bivariate normal distribution, and use these scores to test the hypothesis H_0: $\rho = 0$ against H_1: $\rho \neq 0$ at $\alpha = 0.10$.

9.6-5. A random sample of size 28 from a bivariate normal distribution yielded a sample correlation coefficient of $r = 0.65$. Find an approximate 90% confidence interval for ρ.

9.6-6. By squaring the binomial expression $[(Y_i - \overline{Y}) - (S_{xY}/s_x^2)(x_i - \overline{x})]$, show that

$$\sum_{i=1}^{n}[(Y_i-\overline{Y}) - (S_{xY}/s_x^2)(x_i-\overline{x})]^2 = \sum_{i=1}^{n}(Y_i-\overline{Y})^2 - 2\left(\frac{S_{xY}}{s_x^2}\right)\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y}) + \frac{S_{xY}^2}{s_x^4}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

equals $(n-1)S_Y^2(1-R^2)$, where $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. HINT: Replace $S_{xY} = \sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})/(n-1)$ by Rs_xS_Y.

9.6-7. To help determine whether gallinules selected their mate on the basis of weight, 14 pairs of gallinules were captured and weighed. Test the null hypothesis H_0: $\rho = 0$ against a two-sided alternative at an $\alpha = 0.01$ significance level. Given that the male and female weights for the $n = 14$ pairs of birds yielded a sample correlation coefficient of $r = -0.252$, would H_0 be rejected?

9.6-8. In sampling from a bivariate normal distribution, it is true that the sample correlation coefficient R has an approximate normal distribution $N[\rho, (1-\rho^2)^2/n]$ if the sample size n is large. Since, for large n, R is close to ρ, use two terms of the Taylor's expansion of $u(R)$ about ρ and determine that function $u(R)$ such that it has a variance which is (essentially) free of p. (The solution of this exercise explains why the transformation $(1/2)\ln[(1+R)/(1-R)]$ was suggested.)

9.6-9. Show that when $\rho = 0$,

(a) The points of inflection for the graph of the pdf of R are at $r = \pm 1/\sqrt{n-5}$ for $n \geq 7$.

(b) $E(R) = 0$.

(c) $\text{Var}(R) = 1/(n-1), n \geq 3$. HINT: Note that $E(R^2) = E[1-(1-R^2)]$.

9.6-10. In a college health fitness program, let X equal the weight in kilograms of a female freshman at the beginning of the program and let Y equal her change in weight during the semester. We shall use the following data for $n = 16$ observations of (x, y) to test the null hypothesis H_0: $\rho = 0$ against a two-sided alternative hypothesis:

(61.4, −3.2)	(62.9, 1.4)	(58.7, 1.3)	(49.3, 0.6)
(71.3, 0.2)	(81.5,−2.2)	(60.8, 0.9)	(50.2, 0.2)
(60.3, 2.0)	(54.6, 0.3)	(51.1, 3.7)	(53.3, 0.2)
(81.0, −0.5)	(67.6, −0.8)	(71.4, −0.1)	(72.1, −0.1)

(a) What is the conclusion if $\alpha = 0.10$?

(b) What is the conclusion if $\alpha = 0.05$?

9.6-11. Let X and Y have a bivariate normal distribution with correlation coefficient ρ. To test H_0: $\rho = 0$ against H_1: $\rho \neq 0$, a random sample of n pairs of observations is selected. Suppose that the sample correlation coefficient is $r = 0.68$. Using a significance level of $\alpha = 0.05$, find the smallest value of the sample size n so that H_0 is rejected.

9.6-12. In Exercise 6.5-5, data are given for horsepower, the time it takes a car to go from 0 to 60, and the weight in pounds of a car, for 14 cars. Those data are repeated here:

Horsepower	0–60	Weight	Horsepower	0–60	Weight
230	8.1	3516	282	6.2	3627
225	7.8	3690	300	6.4	3892
375	4.7	2976	220	7.7	3377
322	6.6	4215	250	7.0	3625
190	8.4	3761	315	5.3	3230
150	8.4	2940	200	6.2	2657
178	7.2	2818	300	5.5	3518

(a) Let ρ be the correlation coefficient of horsepower and weight. Test H_0: $\rho = 0$ against H_1: $\rho \neq 0$.

(b) Let ρ be the correlation coefficient of horsepower and "0–60." Test H_0: $\rho = 0$ against H_1: $\rho < 0$.

(c) Let ρ be the correlation coefficient of weight and "0–60." Test H_0: $\rho = 0$ against H_1: $\rho \neq 0$.

9.7* STATISTICAL QUALITY CONTROL

Statistical methods can be used in many scientific fields, such as medical research, engineering, chemistry, and psychology. Often, it is necessary to compare two ways of doing something—say, the old way and a possible new way. We collect data on each way, quite possibly in a laboratory situation, and try to decide whether the new way is actually better than the old. Needless to say, it would be terrible to change to the new way at great expense, only to find out that it is really not any better than the old. That is, suppose the lab results indicate, by some statistical method, that the new is seemingly better than the old. Can we actually extrapolate those outcomes in the lab to the situations in the real world? Clearly, statisticians cannot make these decisions, but they should be made by some professional who knows both statistics and the specialty in question very well. The statistical analysis might provide helpful guidelines, but we still need the expert to make the final decision.

However, even before investigating possible changes in any process, it is extremely important to determine exactly what the process in question is doing at the present time. Often, people in charge of an organization do not understand the capabilities of many of its processes. Simply measuring what is going on frequently leads to improvements. In many cases, measurement is easy, such as determining the diameter of a bolt, but sometimes it is extremely difficult, as in evaluating good teaching or many other service activities. But if at all possible, we encourage those involved to begin to "listen" to their processes; that is, they should measure what is going on in their organization. These measurements alone often are the beginning of desirable improvements. While most of our remarks in this chapter concern measurements made in manufacturing, service industries frequently find them just as useful.

At one time, some manufacturing plants would make parts to be used in the construction of some piece of equipment. Say a particular line in the plant, making a certain part, might produce several hundreds of them each day. These items would then be sent on to an inspection cage, where they would be checked for goodness, often several days or even weeks later. Occasionally, the inspectors would discover many defectives among the items made, say, two weeks ago. There was little that could be done at that point except scrap or rework the defective parts, both expensive outcomes.

In the 1920s, W. A. Shewhart, who was working for AT&T Bell Labs, recognized that this was an undesirable situation and suggested that, with some suitable frequency, a sample of these parts should be taken as they were being made. If the sample indicated that the items were satisfactory, the manufacturing process would continue. But if the sampled parts were not satisfactory, corrections should be made then so that things became satisfactory. This idea led to what are commonly called *Shewhart control charts*—the basis of what was called *statistical quality control* in those early days; today it is often referred to as *statistical process control*.

Shewhart control charts consist of calculated values of a statistic, say, $\bar{x}$, plotted in sequence. That is, in making products, every so often (each hour, each day, or each week, depending upon how many items are being produced) a sample of size n of them is taken, and they are measured, resulting in the observations $x_1, x_2, \ldots, x_n$. The average $\bar{x}$ and the standard deviation s are computed. This is done k times, and the k values of $\bar{x}$ and s are averaged, resulting in $\bar{\bar{x}}$ and $\bar{s}$, respectively; usually, k is equal to some number between 10 and 30.

The central limit theorem states that if the true mean μ and standard deviation σ of the process were known, then almost all of the $\bar{x}$-values would plot between $\mu - 3\sigma/\sqrt{n}$ and $\mu + 3\sigma/\sqrt{n}$, unless the system has actually changed. However, suppose we know neither μ nor σ, and thus μ is estimated by $\bar{\bar{x}}$ and $3\sigma/\sqrt{n}$ by $A_3\bar{s}$, where $\bar{\bar{x}}$ and $\bar{s}$ are the respective means of the k observations of $\bar{x}$ and s, and where A_3 is a factor depending upon n that can be found in books on statistical quality control. A few values of A_3 (and some other constants that will be used later) are given in Table 9.7-1 for typical values of n.

The estimates of $\mu \pm 3\sigma/\sqrt{n}$ are called the *upper control limit* (UCL), $\bar{\bar{x}} + A_3\bar{s}$, and the *lower control limit* (LCL), $\bar{\bar{x}} - A_3\bar{s}$, and $\bar{\bar{x}}$ provides the estimate of the centerline. A typical plot is given in Figure 9.7-1. Here, in the 13th sampling period, $\bar{x}$ is outside the control limits, indicating that the process has changed and that some investigation and action are needed to correct this change, which seems like an upward shift in the process.

Note that there is a control chart for the s values, too. From sampling distribution theory, values of B_3 and B_4 have been determined and are given in Table 9.7-1, so we know that almost all the s-values should be between $B_3\bar{s}$ and $B_4\bar{s}$ if there is no

Table 9.7-1 Some constants used with control charts

n	A_3	B_3	B_4	A_2	D_3	D_4
4	1.63	0	2.27	0.73	0	2.28
5	1.43	0	2.09	0.58	0	2.11
6	1.29	0.03	1.97	0.48	0	2.00
8	1.10	0.185	1.815	0.37	0.14	1.86
10	0.98	0.28	1.72	0.31	0.22	1.78
20	0.68	0.51	1.49	0.18	0.41	1.59

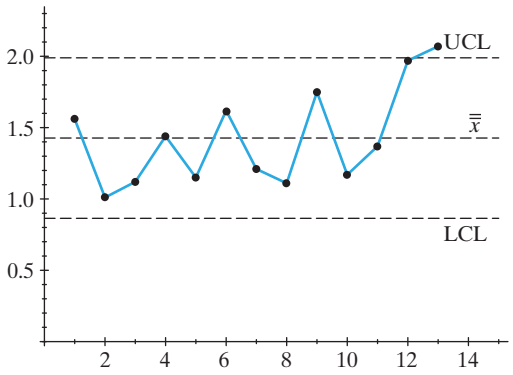

Figure 9.7-1 Typical control chart

change in the underlying distribution. So again, if an individual s-value is outside these control limits, some action should be taken, as it seems as if there has been a change in the variation of the underlying distribution.

Often, when these charts are first constructed after $k = 10$ to 30 sampling periods, many points fall outside the control limits. A team consisting of workers, the manager of the process, the supervisor, an engineer, and even a statistician should try to find the reasons that this has occurred, and the situation should be corrected. After this is done and the points plot within the control limits, the process is "in statistical control." However, being in statistical control is not a guarantee of satisfaction with the products. Since $A_3\bar{s}$ is an estimate of $3\sigma/\sqrt{n}$, it follows that $\sqrt{n}A_3\bar{s}$ is an estimate of 3σ, and with an underlying distribution close to a normal one, almost all items would be between $\bar{\bar{x}} \pm \sqrt{n}A_3\bar{s}$. If these limits are too wide, then corrections must be made again.

If the variation is under control (i.e., if $\bar{x}$ and s are within their control limits), we say that the variations seen in $\bar{x}$ and s are due to common causes. If products made under such a system with these existing common causes are satisfactory, then production continues. If either $\bar{x}$ or s, however, is outside the control limits, that is an indication that some special causes are at work, and they must be corrected. That is, a team should investigate the problem and some action should be taken.

Table 9.7-2 Console opening times

Group	x_1	x_2	x_3	x_4	x_5	$\bar{x}$	s	R
1	1.2	1.8	1.7	1.3	1.4	1.480	0.259	0.60
2	1.5	1.2	1.0	1.0	1.8	1.300	0.346	0.80
3	0.9	1.6	1.0	1.0	1.0	1.100	0.283	0.70
4	1.3	0.9	0.9	1.2	1.0	1.060	0.182	0.40
5	0.7	0.8	0.9	0.6	0.8	0.760	0.114	0.30
6	1.2	0.9	1.1	1.0	1.0	1.040	0.104	0.30
7	1.1	0.9	1.1	1.0	1.4	1.100	0.187	0.50
8	1.4	0.9	0.9	1.1	1.0	1.060	0.207	0.50
9	1.3	1.4	1.1	1.5	1.6	1.380	0.192	0.50
10	1.6	1.5	1.4	1.3	1.5	1.460	0.114	0.30
						$\bar{\bar{x}} = 1.174$	$\bar{s} = 0.200$	$\bar{R} = 0.49$

Example 9.7-1

A company produces a storage console. Twice a day, nine critical characteristics are tested on five consoles that are selected randomly from the production line. One of these characteristics is the time it takes the lower storage component door to open completely. Table 9.7-2 lists the opening times in seconds for the consoles that were tested during one week. Also included in the table are the sample means, the sample standard deviations, and the ranges.

The upper control limit (UCL) and the lower control limit (LCL) for $\bar{x}$ are found using A_3 in Table 9.7-1 with $n = 5$ as follows:

$$\text{UCL} = \bar{\bar{x}} + A_3\bar{s} = 1.174 + 1.43(0.20) = 1.460$$

and

$$\text{LCL} = \bar{\bar{x}} - A_3\bar{s} = 1.174 - 1.43(0.20) = 0.888.$$

These control limits and the sample means are plotted on the $\bar{x}$ chart in Figure 9.7-2. There should be some concern about the fifth sampling period; thus, there should be an investigation to determine why that particular $\bar{x}$ is below the LCL.

The UCL and LCL for s are found using B_3 and B_4 in Table 9.7-1 with $n = 5$ as follows:

$$\text{UCL} = B_4\bar{s} = 2.09(0.200) = 0.418$$

and

$$\text{LCL} = B_3\bar{s} = 0(0.200) = 0.$$

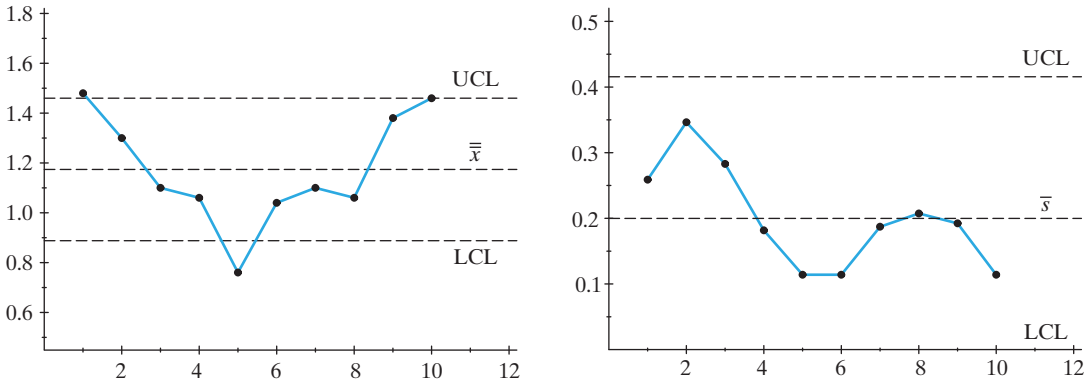

Figure 9.7-2 The $\bar{x}$ chart and s chart

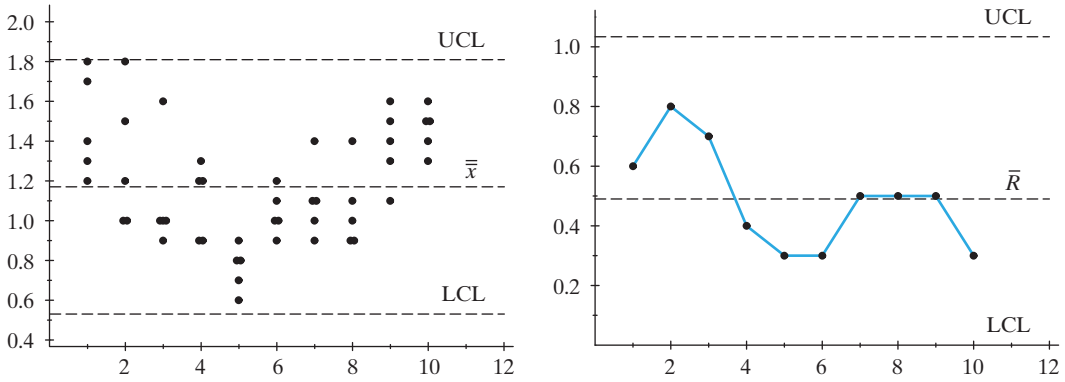

Figure 9.7-3 Plot of 50 console opening times and R chart

These control limits and the sample standard deviations are plotted on the s chart in Figure 9.7-2.

Almost all of the observations should lie between $\bar{\bar{x}} \pm \sqrt{n}\, A_3 \bar{s}$; namely,

$$1.174 + \sqrt{5}\,(1.43)(0.20) = 1.814$$

and

$$1.174 - \sqrt{5}\,(1.43)(0.20) = 0.534.$$

This situation is illustrated in Figure 9.7-3, in which all 50 observations do fall within these control limits. ■

In most books on statistical quality control, there is an alternative way of constructing the limits on an $\bar{x}$ chart. For each sample, we compute the range, R, which is the absolute value of the difference of the extremes of the sample. This computation is much easier than that for calculating s. After k samples are taken, we compute the average of these R-values, obtaining $\overline{R}$ as well as $\bar{\bar{x}}$. The statistic $A_2\,\overline{R}$ serves as an estimate of $3\sigma/\sqrt{n}$, where A_2 is found in Table 9.7-1. Thus, the estimates of $\mu \pm 3\sigma/\sqrt{n}$, namely, $\bar{\bar{x}} \pm A_2\,\overline{R}$, can be used as the UCL and LCL of an $\bar{x}$ chart.

In addition, $\sqrt{n}A_2\,\overline{R}$ is an estimate of 3σ; so, with an underlying distribution that is close to a normal one, we find that almost all observations are within the limits $\overline{\overline{x}} \pm \sqrt{n}A_2\,\overline{R}$.

Moreover, an R chart can be constructed with centerline $\overline{R}$ and control limits equal to $D_3\overline{R}$ and $D_4\overline{R}$, where D_3 and D_4 are given in Table 9.7-1 and were determined so that almost all R-values should be between the control limits if there is no change in the underlying distribution. Thus, a value of R falling outside those limits would indicate a change in the spread of the underlying distribution, and some corrective action should be considered.

The use of R, rather than s, is illustrated in the next example.

Example 9.7-2

Using the data in Example 9.7-1, we compute UCL and LCL for an $\overline{x}$ chart. We use $\overline{\overline{x}} \pm A_2\,\overline{R}$ as follows:

$$\text{UCL} = \overline{\overline{x}} + A_2\,\overline{R} = 1.174 + 0.58(0.49) = 1.458$$

and

$$\text{LCL} = \overline{\overline{x}} - A_2\,\overline{R} = 1.174 - 0.58(0.49) = 0.890.$$

Note that these values are very close to the limits that we found for the $\overline{x}$ chart in Figure 9.7-2 using $\overline{\overline{x}} \pm A_3\overline{s}$. In addition, almost all of the observations should lie within the limits $\overline{x} + \sqrt{n}A_2\,\overline{R}$, which are

$$\text{UCL} = 1.174 + \sqrt{5}\,(0.58)(0.49) = 1.809$$

and

$$\text{LCL} = 1.174 - \sqrt{5}\,(0.58)(0.49) = 0.539.$$

Note that these are almost the same as the limits found in Example 9.7-1 and plotted in Figure 9.7-3.

An R chart can be constructed with centerline $\overline{R} = 0.49$ and control limits given by

$$\text{UCL} = D_4\overline{R} = 2.11(0.49) = 1.034$$

and

$$\text{LCL} = D_3\overline{R} = 0(0.49) = 0.$$

Figure 9.7-3 illustrates this control chart for the range, and we see that its pattern is similar to that of the s chart in Figure 9.7-2. ■

There are two other Shewhart control charts: the p and c charts. The central limit theorem, which provided a justification for the three-sigma limits in the $\overline{x}$ chart, also justifies the control limits in the p chart. Suppose the number of defectives among n items that are selected randomly—say, D—has a binomial distribution $b(n,p)$. Then the limits $p \pm 3\sqrt{p(1-p)/n}$ should include almost all of the D/n-values.

However, p must be approximated by observing k values of D—say, $D_1, D_2, \ldots, D_k$—and computing what is called $\overline{p}$ in the statistical quality control literature, namely,

$$\overline{p} = \frac{D_1 + D_2 + \cdots + D_k}{kn}.$$

Thus, the LCL and UCL for the fraction defective, D/n, are respectively given by

$$\text{LCL} = \overline{p} - 3\sqrt{\overline{p}(1-\overline{p})/n}$$

and

$$\text{UCL} = \overline{p} + 3\sqrt{\overline{p}(1-\overline{p})/n}.$$

If the process is in control, almost all D/n-values are between the LCL and UCL. Still, this may not be satisfactory and improvements might be needed to decrease $\overline{p}$. If it is satisfactory, however, let the process continue under these common causes of variation until a point, D/n, outside the control limits would indicate that some special cause has changed the variation. [Incidentally, if D/n is below the LCL, this might very well indicate that some type of change for the better has been made, and we want to find out why. In general, outlying statistics can often suggest that good (as well as bad) breakthroughs have been made.]

The next example gives the results of a simple experiment that you can easily duplicate.

Example 9.7-3

Let D_i equal the number of yellow candies in a 1.69-ounce bag. Because the number of pieces of candy varies slightly from bag to bag, we shall use an average value for n when we construct the control limits. Table 9.7-3 lists, for 20 packages, the number of pieces of candy in the package, the number of yellow ones, and the proportion of yellow ones.

For these data,

$$\sum_{i=1}^{20} n_i = 1124 \qquad \text{and} \qquad \sum_{i=1}^{20} D_i = 219.$$

It follows that

$$\overline{p} = \frac{219}{1124} = 0.195 \qquad \text{and} \qquad \overline{n} = \frac{1124}{20} \approx 56.$$

Thus, the LCL and UCL are respectively given by

$$\text{LCL} = \overline{p} - 3\sqrt{\overline{p}(1-\overline{p})/56} = 0.195 - 3\sqrt{0.195(0.805)/56} = 0.036$$

and

$$\text{UCL} = \overline{p} + 3\sqrt{\overline{p}(1-\overline{p})/56} = 0.195 + 3\sqrt{0.195(0.805)/56} = 0.354.$$

The control chart for p is depicted in Figure 9.7-4. (For your information the "true" value for p is 0.20.) ■

Consider the following explanation of the c chart: Suppose the number of flaws, say, C, on some product has a Poisson distribution with parameter λ. If λ is

Table 9.7-3 Data on yellow candies

Package	n_i	D_i	D_i/n_i	Package	n_i	D_i	D_i/n_i
1	56	8	0.14	11	57	10	0.18
2	55	13	0.24	12	59	8	0.14
3	58	12	0.21	13	54	10	0.19
4	56	13	0.23	14	55	11	0.20
5	57	14	0.25	15	56	12	0.21
6	54	5	0.09	16	57	11	0.19
7	56	14	0.25	17	54	6	0.11
8	57	15	0.26	18	58	7	0.12
9	54	11	0.20	19	58	12	0.21
10	55	13	0.24	20	58	14	0.24

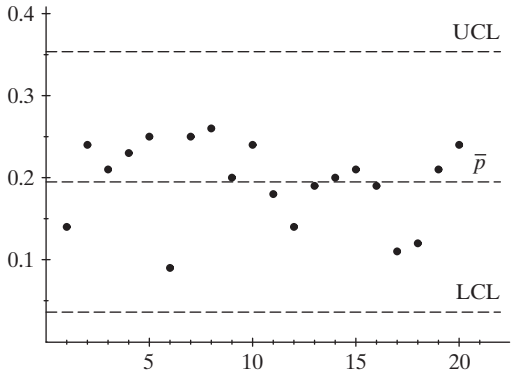

Figure 9.7-4 The p chart

sufficiently large, as in Example 5.7-5, we consider approximating the discrete Poisson distribution with the continuous $N(\lambda, \lambda)$ distribution. Thus, the interval from $\lambda - 3\sqrt{\lambda}$ to $\lambda + 3\sqrt{\lambda}$ contains virtually all of the C-values. Since λ is unknown, however, it must be approximated by $\bar{c}$, the average of the k values, $c_1, c_2, \ldots, c_k$. Hence the two control limits for C are computed as

$$\text{LCL} = \bar{c} - 3\sqrt{\bar{c}} \qquad \text{and} \qquad \text{UCL} = \bar{c} + 3\sqrt{\bar{c}}.$$

The remarks made about the $\bar{x}$ and $\bar{p}$ charts apply to the c chart as well, but we must remember that each c-value is the number of flaws on one manufactured item, not an average $\bar{x}$ or a fraction defective D/n.

Exercises

9.7-1. It is important to control the viscosity of liquid dishwasher soap so that it flows out of the container but does not run out too rapidly. Thus, samples are taken randomly throughout the day and the viscosity is measured. Use the following 20 sets of 5 observations for this exercise:

Observations					$\bar{x}$	s	R
158	147	158	159	169	158.20	7.79	22
151	166	151	143	169	156.00	11.05	26
153	174	151	164	185	165.40	14.33	34
168	140	180	176	154	163.60	16.52	40
160	187	145	164	158	162.80	15.29	42
169	153	149	144	157	154.40	9.48	25
156	183	157	140	162	159.60	15.47	43
158	160	180	154	160	162.40	10.14	26
164	168	154	158	164	161.60	5.55	14
159	153	170	158	170	162.00	7.65	17
150	161	169	166	154	160.00	7.97	19
157	138	155	134	165	149.80	13.22	31
161	172	156	145	153	157.40	10.01	27
143	152	152	156	163	153.20	7.26	20
179	157	135	172	143	157.20	18.63	44
154	165	145	152	145	152.20	8.23	20
171	189	144	154	147	161.00	18.83	45
187	147	159	167	151	162.20	15.85	40
153	168	148	188	152	161.80	16.50	40
165	155	140	157	176	158.60	13.28	36

(a) Calculate the values of $\bar{\bar{x}}$, $\bar{s}$, and $\bar{R}$.

(b) Use the values of A_3 and $\bar{s}$ to construct an $\bar{x}$ chart.

(c) Construct an s chart.

(d) Use the values of A_2 and $\bar{R}$ to construct an $\bar{x}$ chart.

(e) Construct an R chart.

(f) Do the charts indicate that viscosity is in statistical control?

9.7-2. It is necessary to control the percentage of solids in a product, so samples are taken randomly throughout the day and the percentage of solids is measured. Use the following 20 sets of 5 observations for this exercise:

Observations					$\bar{x}$	s	R
69.8	71.3	65.6	66.3	70.1	68.62	2.51	5.7
71.9	69.6	71.9	71.1	71.7	71.24	0.97	2.3
71.9	69.8	66.8	68.3	64.4	68.24	2.86	7.5
64.2	65.1	63.7	66.2	61.9	64.22	1.61	4.3
66.1	62.9	66.9	67.3	63.3	65.30	2.06	4.4
63.4	67.2	67.4	65.5	66.2	65.94	1.61	4.0
67.5	67.3	66.9	66.5	65.5	66.74	0.79	2.0
63.9	64.6	62.3	66.2	67.2	64.84	1.92	4.9
66.0	69.8	69.7	71.0	69.8	69.26	1.90	5.0
66.0	70.3	65.5	67.0	66.8	67.12	1.88	4.8
67.6	68.6	66.5	66.2	70.4	67.86	1.71	4.2
68.1	64.3	65.2	68.0	65.1	66.14	1.78	3.8
64.5	66.6	65.2	69.3	62.0	65.52	2.69	7.3
67.1	68.3	64.0	64.9	68.2	66.50	1.96	4.3
67.1	63.8	71.4	67.5	63.7	66.70	3.17	7.7
60.7	63.5	62.9	67.0	69.6	64.74	3.53	8.9
71.0	68.6	68.1	67.4	71.7	69.36	1.88	4.3
69.5	61.5	63.7	66.3	68.6	65.92	3.34	8.0
66.7	75.2	79.0	75.3	79.2	75.08	5.07	12.5
77.3	67.2	69.3	67.9	65.6	69.46	4.58	11.7

(a) Calculate the values of $\bar{\bar{x}}$, $\bar{s}$, and $\bar{R}$.

(b) Use the values of A_3 and $\bar{s}$ to construct an $\bar{x}$ chart.

(c) Construct an s chart.

(d) Use the values of A_2 and $\bar{R}$ to construct an $\bar{x}$ chart.

(e) Construct an R chart.

(f) Do the charts indicate that the percentage of solids in this product is in statistical control?

9.7-3. It is important to control the net weight of a packaged item; thus, items are selected randomly throughout

the day from the production line and their weights are recorded. Use the following 20 sets of 5 weights (in grams) for this exercise (note that a weight recorded here is the actual weight minus 330):

Observations					$\bar{x}$	s	R
7.97	8.10	7.73	8.26	7.30	7.872	0.3740	0.96
8.11	7.26	7.99	7.88	8.88	8.024	0.5800	1.62
7.60	8.23	8.07	8.51	8.05	8.092	0.3309	0.91
8.44	4.35	4.33	4.48	3.89	5.098	1.8815	4.55
5.11	4.05	5.62	4.13	5.01	4.784	0.6750	1.57
4.79	5.25	5.19	5.23	3.97	4.886	0.5458	1.28
4.47	4.58	5.35	5.86	5.61	5.174	0.6205	1.39
5.82	4.51	5.38	5.01	5.54	5.252	0.5077	1.31
5.06	4.98	4.13	4.58	4.35	4.620	0.3993	0.93
4.74	3.77	5.05	4.03	4.29	4.376	0.5199	1.28
4.05	3.71	4.73	3.51	4.76	4.152	0.5748	1.25
3.94	5.72	5.07	5.09	4.61	4.886	0.6599	1.78
4.63	3.79	4.69	5.13	4.66	4.580	0.4867	1.34
4.30	4.07	4.39	4.63	4.47	4.372	0.2079	0.56
4.05	4.14	4.01	3.95	4.05	4.040	0.0693	0.19
4.20	4.50	5.32	4.42	5.24	4.736	0.5094	1.12
4.54	5.23	4.32	4.66	3.86	4.522	0.4999	1.37
5.02	4.10	5.08	4.94	5.18	4.864	0.4360	1.08
4.80	4.73	4.82	4.69	4.27	4.662	0.2253	0.55
4.55	4.76	4.45	4.85	4.02	4.526	0.3249	0.83

(a) Calculate the values of $\bar{\bar{x}}$, $\bar{s}$, and $\bar{R}$.

(b) Use the values of A_3 and $\bar{s}$ to construct an $\bar{x}$ chart.

(c) Construct an s chart.

(d) Use the values of A_2 and $\bar{R}$ to construct an $\bar{x}$ chart.

(e) Construct an R chart.

(f) Do the charts indicate that these fill weights are in statistical control?

9.7-4. A company has been producing bolts that are about $\bar{p} = 0.02$ defective, and this is satisfactory. To monitor the quality of the process, 100 bolts are selected at random each hour and the number of defective bolts counted. With $\bar{p} = 0.02$, compute the UCL and LCL of the $\bar{p}$ chart. Then suppose that, over the next 24 hours, the following numbers of defective bolts are observed:

4 1 1 0 5 2 1 3 4 3 1 0 0 4 1 1 6 2 0 0 2 8 7 5

Would any action have been required during this time?

9.7-5. To give some indication of how the values in Table 9.7-1 are calculated, values of A_3 are found in this exercise. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the normal distribution $N(\mu, \sigma^2)$. Let S^2 equal the sample variance of this random sample.

(a) Use the fact that $Y = (n-1)S^2/\sigma^2$ has a distribution that is $\chi^2(n-1)$ to show that $E[S^2] = \sigma^2$.

(b) Using the $\chi^2(n-1)$ pdf, find the value of $E(\sqrt{Y})$.

(c) Show that

$$E\left[\frac{\sqrt{n-1}\,\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)} S\right] = \sigma.$$

(d) Verify that

$$\frac{3}{\sqrt{n}}\left[\frac{\sqrt{n-1}\,\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}\right] = A_3,$$

found in Table 9.7-1 for $n = 5$ and $n = 6$. Thus, $A_3\,\bar{s}$ approximates $3\sigma/\sqrt{n}$.

9.7-6. In a woolen mill, 100-yard pieces are inspected. In the last 20 observations, the following numbers of flaws were found:

2 4 0 1 0 3 4 1 1 2 4 0 0 1 0 3 2 3 5 0

(a) Compute the control limits of the c chart and draw this control chart.

(b) Is the process in statistical control?

9.7-7. In the past, $n = 50$ fuses are tested each hour and $\bar{p} = 0.03$ have been found defective. Calculate the UCL and LCL. After a production error, say the true p shifts to $p = 0.05$.

(a) What is the probability that the next observation exceeds the UCL?

(b) What is the probability that at least one of the next five observations exceeds the UCL? HINT: Assume independence and compute the probability that none of the next five observations exceeds the UCL.

9.7-8. Snee (see References) has measured the thickness of the "ears" of paint cans. (The "ear" of a paint can is the tab that secures the lid of the the can.) At periodic intervals, samples of five paint cans are taken from a hopper that collects the production from two machines, and the thickness of each ear is measured. The results (in inches $\times$ 1000) of 30 such samples are as follows:

Observations					$\bar{x}$	s	R
29	36	39	34	34	34.4	3.64692	10
29	29	28	32	31	29.8	1.64317	4
34	34	39	38	37	36.4	2.30217	5
35	37	33	38	41	36.8	3.03315	8
30	29	31	38	29	31.4	3.78153	9
34	31	37	39	36	35.4	3.04959	8
30	35	33	40	36	34.8	3.70135	10
28	28	31	34	30	30.2	2.48998	6
32	36	38	38	35	35.8	2.48998	6
35	30	37	35	31	33.6	2.96648	7
35	30	35	38	35	34.6	2.88097	8
38	34	35	35	31	34.6	2.50998	7
34	35	33	30	34	33.2	1.92354	5
40	35	34	33	35	35.4	2.70185	7
34	35	38	35	30	34.4	2.88097	8
35	30	35	29	37	33.2	3.49285	8
40	31	38	35	31	35.0	4.06202	9
35	36	30	33	32	33.2	2.38747	6
35	34	35	30	36	34.0	2.34521	6
35	35	31	38	36	35.0	2.54951	7
32	36	36	32	36	34.4	2.19089	4
36	37	32	34	34	34.6	1.94936	5
29	34	33	37	35	33.6	2.96648	8
36	36	35	37	37	36.2	0.83666	2
36	30	35	33	31	33.0	2.54951	6
35	30	29	38	35	33.4	3.78153	9
35	36	30	34	36	34.2	2.48998	6
35	30	36	29	35	33.0	3.24037	7
38	36	35	31	31	34.2	3.11448	7
30	34	40	28	30	32.4	4.77493	12

(a) Calculate the values of $\bar{\bar{x}}$, $\bar{s}$, and $\bar{R}$.

(b) Use the values of A_3 and $\bar{s}$ to construct an $\bar{x}$ chart.

(c) Construct an s chart.

(d) Use the values of A_2 and $\bar{R}$ to construct an $\bar{x}$ chart.

(e) Construct an R chart.

(f) Do the charts indicate that these fill weights are in statistical control?

9.7-9. Ledolter and Hogg (see References) report that, in the production of stainless steel pipes, the number of defects per 100 feet should be controlled. From 15 randomly selected pipes of length 100 feet, the following data on the number of defects were observed:

6 10 8 1 7 9 7 4 5 10 3 4 9 8 5

(a) Compute the control limits of the c chart and draw this control chart.

(b) Is the process in statistical control?

9.7-10. Suppose we find that the number of blemishes in 50-foot tin strips averages about $\bar{c} = 1.4$. Calculate the control limits. Say the process has gone out of control and this average has increased to 3.

(a) What is the probability that the next observation will exceed the UCL?

(b) What is the probability that at least 1 of the next 10 observations will exceed the UCL?

HISTORICAL COMMENTS Chi-square tests were the invention of Karl Pearson, except that he had it wrong in the case in which parameters are estimated. When R. A. Fisher was a brash young man, he told his senior, Pearson, that he should reduce the number of degrees of freedom of the chi-square distribution by 1 for every parameter that was estimated. Pearson never believed this (of course, Fisher was correct), and, as editor of the very prestigious journal *Biometrika*, Pearson blocked Fisher in his later professional life from publishing in that journal. Fisher was disappointed, and the two men battled during their lifetimes; however, later Fisher saw this conflict to be to his advantage, as it made him consider applied journals in which to publish, and thus he became a better, more well-rounded scientist.

Another important item in this chapter is the analysis of variance (ANOVA). This is just the beginning of what is called the design of experiments, developed by R. A. Fisher. In our simple cases in this section, he shows how to test for the best levels of factors in the one-factor and two-factor cases. We study a few important generalizations in Section 9.5. The analysis of designed experiments was a huge contribution by Fisher.

Quality improvement made a substantial change in manufacturing beginning in the 1920s, with Walter A. Shewhart's control charts. In fairness, it should be noted that the British started a similar program about the same time. Statistical quality control, as described in Section 9.7, really had a huge influence during World War II, with many universities giving short courses in the subject. These courses continued after the war, but the development of the importance of total quality improvement lagged behind. W. Edwards Deming complained that the Japanese used his quality ideas beginning in the 1950s, but the Americans did not adopt them until 1980. That year NBC televised a program entitled *If Japan Can, Why Can't We?*, and Deming was the "star" of that broadcast. He related that the next day his phone "started ringing off the hook." Various companies requested that he spend one day with them to get them started on the right path. According to Deming, they all wanted "instant pudding," and he noted that he had asked the Japanese to give him five years to make the improvements he pioneered. Actually, using his philosophy, many of these companies did achieve substantial results in quality sooner than that. However, it was after the NBC program that Deming started his famous four-day courses, and he taught his last one in December of 1993, about 10 days before his death at the age of 93.

Many of these quality efforts in the 1970s and 1980s used the name "Total Quality Management" or, later, "Continuous Process Improvements." However, it was Motorola's Six Sigma program, which started in the late 1980s and has continued for over 20 years since then, that has had the biggest impact. In addition to Motorola, GE, Allied, and a large number of companies have used this system. In our opinion, Six Sigma is the leading development in the quality improvement effort.