AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

Minitab is a statistical package, a computer program that performs many statistical procedures. The versions of Minitab now available for use on personal computers are menu-driven and much easier to use than the main-frame version originally discussed in this text. Those sections are not included in this online edition of the text. At this time, the most recent version is Minitab 15, available at very reasonable prices for purchase or rental from:

www.e-academy.com/minitab

System Requirements

Processor:	PC with a 1 GHz 32- or 64-bit processor
Memory:	512 MB or more of available RAM
Disk Space:	125 MB free space available
Operating System:	Microsoft Windows 2000, XP, or Vista.
Display:	A display capable of 1024 X 768 or higher resolution
Software:	Adobe Acrobat Reader 5.0 or higher for Meet Minitab

CHAPTER 16

Inferences About Qualitative (or Categorical) Variables

IN THIS CHAPTER

- Chi-square goodness-of-fit test for a single qualitative variable
- Small-sample inference about a proportion
- Chi-square test of independence of two qualitative variables
- Chi-square test of homogeneity of a discrete distribution across populations
- Fisher's exact test

When does labor begin for pregnant women? Recording time of onset of labor for many women, investigators consider a day as 24 hour-long periods and ask: Is labor as likely to begin in any of these 24 time periods (Exercise 16-5)? Geneticists studying height (tall, dwarf) and leaf type (cut-leaf, potato-leaf) in a sample of tomato plants ask: Do the numbers of plants across the four combinations of observed height and leaf type suggest that the two characteristics are distributed independently to offspring (Exercise 16-3)? These two questions relate to inferences about a single qualitative or categorical variable, the subject of Sections 16-1 and 16-2. In Section 16-1, we discuss the chi-square goodness-of-fit test for comparing the observed distribution of a qualitative variable with some specified distribution.

> The **chi-square goodness-of-fit test** is a large-sample test for making inferences about a single qualitative variable. We compare the observed distribution of frequencies across categories with a specified distribution.

We consider small-sample inferences about a proportion (or, equivalently, testing for goodness of fit with small samples when a qualitative variable has exactly two categories) in Section 16-2.

Is the incidence of overwhelming infection among children with sickle cell disease different for those children treated with oral penicillin than for those treated with a placebo (Exercise 16-14)? Is there any association between hypertension (yes, no) and obesity (low, average, high) among adults in Western Australia (Exercise 16-9)? Is level of parental encouragement (high, low) related to a male high school senior's plans to attend college (yes, no)? Is the relationship between parental encouragement and college plans different for male and female high school seniors (Exercise 16-10)? These questions deal with inferences about the relationship between two qualitative variables. Sections 16-3 and 16-4 discuss large-sample tests of association between two qualitative variables, called chi-square tests of association.

> A **chi-square test of association** is a large-sample test for making inferences about the relationship between two qualitative variables.

Section 16-3 covers the chi-square test of independence of two qualitative variables.

> The **chi-square test of independence** is a large-sample test of the null hypothesis that two qualitative variables have independent distributions in a population.

We use the chi-square test of homogeneity, discussed in Section 16-4, to compare the distribution of a qualitative variable across populations. Comparing several proportions in a special case.

> The **chi-square test of homogeneity** is a large-sample test of the null hypothesis that a qualitative variable has the same distribution in each of several populations.

Fisher's exact test is a small-sample test of association in a 2×2 frequency table, when each variable has two categories. This test is discussed in Section 16-5.

We use **Fisher's exact test** to make inferences about the relationship between two qualitative variables, each having exactly two categories. Probabilities under the null hypothesis are conditional on the observed row and column totals in the 2 × 2 frequency table summarizing the observed results.

We begin in Section 16-1 with the chi-square goodness-of-fit test.

16-1

The Chi-Square Goodness-of-Fit Test

Suppose a qualitative variable Y has k possible values or categories, denoted by the numbers 1 through k. The probability distribution of Y is determined by the probabilities $P(Y = i)$, where i ranges from 1 to k. We want to test the null hypothesis that the probability distribution of Y is equal to some specified distribution. Equivalently, we want to test the *goodness of fit* of the specified distribution with what we observe in a sample. We call the test the *chi-square goodness-of-fit* test. Consider the following example.

EXAMPLE 16-1

Researchers wanted to study the relationship between tubal infertility and method of contraception (Cramer et al., 1987). (Tubal infertility results from damage to the Fallopian tubes, often caused by infection.) They interviewed 283 women with tubal infertility and no children, classifying the women according to method of contraception longest used. The results are shown below:

Method longest used	Observed frequency
None	36
Intrauterine device	39
Oral contraceptives	144
Barrier methods	64
Total	283

Barrier methods include diaphragms or condoms (or both).

The researchers also interviewed a large control group of women with children. These women were similar to the women with tubal infertility with respect to age, race, and income status. The researchers classified women in this control group according to method of contraception longest used:

Method longest used	Percent of control group
None	13.9
Intrauterine device	6.9
Oral contraceptives	52.7
Barrier methods	26.5

Why would medical workers be interested in looking for a possible relationship between tubal infertility and method of contraception? What would you want to know about the sample of women with tubal infertility and the control group to believe that comparisons are meaningful?

Considering method of contraception longest used, we ask: How does the sample distribution for women with tubal infertility compare with the control distribution? We will use the chi-square goodness-of-fit test to assess the null hypothesis that the distribution for women with tubal infertility is the same as for the control group. First we will outline the steps for carrying out the test.

The significance level approach to the chi-square goodness-of-fit test

1. The qualitative variable Y has k possible categories. The hypotheses are:

   Null hypothesis: $P(Y = i) = p_i$ for all i from 1 to k.
   Alternative hypothesis: $P(Y = i)$ does not equal p_i for all i from 1 to k.

   The specified probabilities p_1 through p_k sum to 1.
2. We have a sample of size n from the population. Let O_i denote the number of observations in category i. Define the expected frequency for category i by $E_i = n \times p_i$. E_i is the frequency we expect on average in category i under the null hypothesis. We want to measure how far the observed frequencies O_1 through O_k are from the null hypothesis expected frequencies E_1 through E_k. The test statistic is

$$\text{Test statistic} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

   If the observed and expected frequencies are close for each category, consistent with the null hypothesis, the test statistic will be small. If the observed and expected frequencies are very different for one or more categories, inconsistent with the null hypothesis, the test statistic will be large.
3. Assume that we have a random sample of size n from the population. Each observation is classified into exactly one of the k categories of variable Y. If n is large enough, the test statistic has approximately the chi-square distribution with $k - 1$ degrees of freedom under the null hypothesis.
4. Choose significance level α.
5. Let X denote a random variable having the chi-square distribution with $k - 1$ degrees of freedom. Find the number c from Table E such that $P(X \leq c) = 1 - \alpha$. The acceptance region is $[0, c)$; the rejection region is $[c, \infty)$.
6. The decision rule is:

   If test statistic $< c$, say the results are consistent with the null hypothesis that the distribution of Y in the population equals the specified distribution.

   If test statistic $\geq c$, say the results are inconsistent with the null hypothesis, suggesting that the distribution of Y in the population does not equal the specified distribution.

7. Collect a random sample of observations on the variable Y. Calculate the test statistic in step 2. Use the decision rule in step 6 to decide whether the distribution of Y in the population appears to equal the specified distribution. Draw conclusions based on the experimental results.

We commonly say that the sample size is large enough to use the chi-square approximation if no expected frequency E_i is less than 1 and no more than 20% of the expected frequencies are less than 5. This is called *Cochran's rule.*

> **Cochran's rule** states that the chi-square approximation (for large-sample tests about one or more qualitative variables) is adequate if no expected frequency is less than 1 and no more than 20% of the expected frequencies are less than 5.

When the sample size is not large enough to justify the chi-square approximation, we may be able to formulate an exact test. For instance, in Section 16-2 we discuss small-sample inference about a qualitative variable that has two categories.

EXAMPLE 16-1 *(continued)*

In Example 16-1, the hypotheses are:

H_0: The distribution across categories for women with tubal infertility is the same as the control distribution.

H_a: The distribution for women with tubal infertility is not the same as the control distribution.

The population includes women with tubal infertility and no children, similar to women in the sample. However, the sample of 283 women is not a random sample from this population. Rather, they were women evaluated for infertility at seven infertility centers over a 30-month period. For our inferences to make sense, we must assume these women are representative of a larger population of interest. We must also assume that the observations for different women are independent. In addition, the control group women must make up a reasonable comparison group. If all these assumptions are met, then under the null hypothesis the test statistic has approximately the chi-square distribution with 3 degrees of freedom.

We will use significance level $\alpha = .005$. Looking in Table E, we find $c = 12.84$. The acceptance region is $[0, 12.84)$ and the rejection region is $[12.84, \infty)$. The decision rule is:

If test statistic < 12.84, say the results are consistent with the null hypothesis.
If test statistic ≥ 12.84, say the results are inconsistent with the null hypothesis.

The calculations for finding the test statistic are outlined in Table 16-1. The test statistic equals 21.5, in the rejection region. This suggests that the distribution for women with tubal infertility is not the same as the distribution for women in the control group.

Comparing the observed and expected frequencies in Table 16-1, we see that the largest differences are in categories 2 and 4. About 20 more women in

TABLE 16-1 Calculating the chi-square goodness-of-fit test statistic for Example 16-1

Category number	Category label	Observed frequency	Category probability under H_0	Expected frequency under H_0
1	None	36	.139	$283 \times .139 = 39.337$
2	Intrauterine device	39	.069	$283 \times .069 = 19.527$
3	Oral contraceptives	144	.527	$283 \times .527 = 149.141$
4	Barrier methods	64	.265	$283 \times .265 = 74.995$
	Totals	283	1.000	283

$$\text{Test statistic} = \frac{(36 - 39.337)^2}{39.337} + \frac{(39 - 19.527)^2}{19.527} + \frac{(144 - 149.141)^2}{149.141} + \frac{(64 - 74.995)^2}{74.995}$$
$$= .283 + 19.419 + .177 + 1.612 \doteq 21.5$$

Degrees of freedom = 4 − 1 = 3

the sample were in the intrauterine device category than expected from the control group distribution. About 11 fewer women in the sample were in the barrier methods category than expected from the control group distribution. Compared with the control group, a greater proportion of women with tubal infertility reported intrauterine device as the method of contraception longest used. A smaller proportion of women with tubal infertility reported barrier methods as longest used, compared with the control women.

What do these results suggest to you about a possible relationship between tubal infertility and method of contraception? How might you interpret these results? What suggestions would you have for further research?

When there are just two categories, we can think of the chi-square goodness-of-fit test as a large-sample test about a proportion. We compare the test statistic with the chi-square distribution with 1 degree of freedom. The resulting p-values are exactly the same as we get using the large-sample test for a proportion based on the standard Gaussian distribution (see Section 10-2).

Section 16-2 discusses inferences about a proportion when the sample size is small.

16-2

Small-Sample Inference About a Proportion Based on a Binomial Distribution

Suppose we have a random sample of observations from a population. Each observation has two possible outcomes; call them success and failure. We want to test hypotheses about the probability p of success, or the proportion p of successes in the population. When the sample size is small, we can base inferences on a binomial distribution. (We did this as a special case when we used the sign test in Section 10-5.) Consider the following example.

EXAMPLE 16-2

Researchers wanted to study the relationship between abnormal sex chromosome genotypes and criminal behavior among men in Denmark (Witkin et al., 1976). Since men with abnormal sex chromosome genotypes tend to be taller than average, the researchers included only men at least 184 centimeters in height. They identified 16 men with an extra X chromosome (the XXY genotype). Three of these 16 men had been convicted of at least one crime.

The researchers included more than 4,000 men who were at least 184 centimeters tall and had the normal XY genotype. Of these men, 9.3% had been convicted of one or more crimes.

Define the crime rate of a group to be the proportion in the group who have been convicted of at least one crime. Does this study suggest that the crime rate for men with the XXY genotype is different from the crime rate for men with the XY genotype? We will base our inferences on a binomial distribution. Let's outline the general approach to hypothesis testing and then apply it to this example.

The significance level approach to small-sample inference about a proportion p

1. The hypotheses are H_0: $p = p_0$ and H_a: $p \neq p_0$, where p_0 is a specified number between 0 and 1.
2. The test statistic is the observed number of successes in the sample.
3. We assume that we have a random sample of observations from a population. Each observation has two possible outcomes, success and failure. Then under the null hypothesis, the test statistic has the binomial(n, p_0) distribution.
4. Select significance level α.
5. Let X denote a random variable having the binomial(n, p_0) distribution. Find c_1 and c_2 so that $P(X \leq c_1) = \alpha/2$ and $P(X \geq c_2) = \alpha/2$. The acceptance region is (c_1, c_2). The rejection region includes $[0, c_1]$ and $[c_2, n]$.
6. The decision rule is:

   If $c_1 <$ test statistic $< c_2$, say the results are consistent with the null hypothesis that the proportion p of successes in the population equals p_0.

   If test statistic $\leq c_1$ or test statistic $\geq c_2$, say the results are inconsistent with the null hypothesis, suggesting that the proportion p of successes in the population does not equal p_0.
7. Collect a sample that satisfies the assumptions in step 3. Find the test statistic in step 2. Use the decision rule in step 6 to decide whether the proportion of successes in the population equals p_0. Draw conclusions based on the experimental results.

EXAMPLE 16-2 *(continued)*

In Example 16-2, let p denote the proportion of Danish men with the XXY genotype who have been convicted of at least one crime. We want to test the hypotheses H_0: $p = .093$ and H_a: $p \neq .093$.

The 16 men in the sample were not selected at random from all Danish men with the XXY genotype. Instead, they were identified from a group of over 4,000 Danish men at least 184 centimeters in height. For our purposes, we will assume that the 16 observations are independent and come from the same

TABLE 16-2 Binomial probability distribution for sample size 16 and probability of success .093

Possible value, k	P(test statistic = k when H_0 is true) $= \binom{16}{k}(.093)^k(.907)^{16-k}$
0	.2098
1	.3441
2	.2646
3	.1266
4	.0422
5	.0104
6	.0020
7	.0003
8	.0000
9	.0000
10	.0000
11	.0000
12	.0000
13	.0000
14	.0000
15	.0000
16	.0000

probability distribution. The test statistic equals the number in the sample who have been convicted of at least one crime. With our assumptions, the test statistic has the binomial(16, .093) distribution under the null hypothesis. The probabilities for the binomial(16, .093) distribution are shown in Table 16-2.

We will use significance level $\alpha = .05$. From Table 16-2, we see that the probability of 0 successes under the null hypothesis is .2098, much larger than $\alpha/2 = .025$. Therefore, only large values of the test statistic are inconsistent with the null hypothesis. We will let $c = 4$. Under the null hypothesis, the probability of seeing a test statistic greater than or equal to 4 is .0549. The acceptance region is [0, 3], the rejection region is [4, 16], and the decision rule is:

If test statistic ≤ 3, say the results are consistent with the null hypothesis.
If test statistic ≥ 4, say the results are inconsistent with the null hypothesis.

Since three of the 16 men with the XXY genotype had been convicted of at least one crime, the test statistic equals 3, in the acceptance region. The p-value = .1815, the probability of a test statistic at least as extreme as 3 under the null hypothesis. (To get the p-value for a two-sided test about a proportion based on a binomial distribution, we add all the probabilities in the null hypothesis distribution that are less than or equal to the probability for the observed test statistic.) In this study of Danish men at least 184 centimeters tall, the difference in crime rate for men with the XXY genotype and for men with the normal XY genotype is not statistically significant. The sample results are consistent with the null hypothesis that the crime rate for men with the XXY genotype is the same as for men with the normal XY genotype.

In the remainder of this chapter, we consider inferences about the relationship between two qualitative variables.

16-3

The Chi-Square Test of Independence of Two Qualitative Variables

Suppose we have a random sample of observations selected from a population. An observation is classified according to each of two qualitative variables. We want to test the null hypothesis that the two variables are distributed independently in the population. If the variables are independent, we say there is no association between them. We test our hypotheses using a chi-square test of association called the *chi-square test of independence.* Consider the following example.

EXAMPLE 16-3

Investigators wanted to evaluate a no-smoking policy in a health maintenance organization with over 6,000 employees (Rosenstock, Stergachis, and Heaney, 1986). Four months after establishment of a smoking ban, the investigators selected 687 employees at random for an opinion survey. They mailed each of these employees a questionnaire to be answered and returned anonymously. Sixty-three percent, or 434, of these 687 employees returned the questionnaire. Distributions of age, sex, and length of employment for the respondents were similar to those for the entire employee population. Respondents provided information on smoking status and approval of the no-smoking policy, with results shown below.

	Approval of the smoking ban			
Smoking status	Approve	Do not approve	Not sure	Total
Never smoked	237	3	10	250
Ex-smoker	106	4	7	117
Current smoker	24	32	11	67
Total	367	39	28	434

Was there an association between smoking status and approval of the smoking ban in this employee population? Equivalently, was approval of the smoking ban independent of smoking status? We will use the chi-square test of independence to evaluate these questions.

Before outlining the test procedure, we need some notation. Suppose variable I has r categories and variable II has c categories. Let O_{ij} denote the number of observations classified into category i of variable I and category j

of variable II. We might arrange our observations into a two-way frequency table having r rows and c columns (called an $r \times c$ frequency table), with the r categories of variable I forming the rows and the c categories of variable II forming the columns. Then O_{ij} appears in row i and column j of the frequency table:

	Variable II						
Variable I	Category 1	Category 2	$\cdots$	Category j	$\cdots$	Category c	Total
Category 1	O_{11}	O_{12}	$\cdots$	O_{1j}	$\cdots$	O_{1c}	R_1
Category 2	O_{21}	O_{22}	$\cdots$	O_{2j}	$\cdots$	O_{2c}	R_2
.							
.							
.							
Category i	O_{i1}	O_{i2}	$\cdots$	O_{ij}	$\cdots$	O_{ic}	R_i
.							
.							
.							
Category r	O_{r1}	O_{r2}	$\cdots$	O_{rj}	$\cdots$	O_{rc}	R_r
Total	C_1	C_2	$\cdots$	C_j	$\cdots$	C_c	n

Let R_i denote the total number of observations in category i of variable I, C_j the total number of observations in category j of Variable II, and n the total sample size. With this notation, we can now outline the steps for the chi-square test of independence.

The significance level approach to the chi-square test of independence of two qualitative variables

1. The null hypothesis states that the two variables are independent. The alternative hypothesis states that the two variables are not independent.
2. The expected frequency in category i of variable I and category j of variable II under the null hypothesis is

$$E_{ij} = \frac{R_i \times C_j}{n}$$

The test statistic is a measure of how far the observed frequencies differ from what we would expect under the null hypothesis:

$$\text{Test statistic} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Small values of the test statistic occur when observed and expected frequencies are close to one another, consistent with the null hypothesis. Large values of the test statistic occur when some observed and expected frequencies are not close, inconsistent with the null hypothesis.

3. Assume that we have a random sample of size n from the population. Each observation can be classified into exactly one of the r categories of variable

I and exactly one of the c categories of variable II. Under the null hypothesis that variables I and II are independent, the test statistic has approximately the chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom.

4. Select significance level α.
5. Let X denote a random variable having the chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom. Find c from Table E such that $P(X \leq c) = 1 - \alpha$. The acceptance region is $[0, c)$; the rejection region is $[c, \infty)$.
6. The decision rule is:

   If test statistic $< c$, say the results are consistent with the null hypothesis that variables I and II are independently distributed in the population.

   If test statistic $\geq c$, say the results are inconsistent with the null hypothesis, suggesting that variables I and II are not independently distributed in the population.
7. Collect a sample that satisfies the assumptions in step 3. Calculate the test statistic in step 2. Use the decision rule in step 6 to decide whether variables I and II appear to be independent in the population. Draw conclusions based on the experimental results.

Because the test statistic is based on frequencies, its exact probability distribution is discrete. We say the sample size is large enough to use the chi-square approximation if no expected frequency is less than 1 and no more than 20% of the expected frequencies are less than 5 (Cochran's rule).

EXAMPLE 16-3 *(continued)*

In Example 16-3, we want to test the hypotheses:

H_0: Smoking status and approval of the smoking ban are independent in the employee population.

H_a: Smoking status and approval of the smoking ban are not independent in the employee population.

The investigators selected 687 employees at random to receive questionnaires. However, the 434 actual respondents compose a self-selected subgroup of this random sample. In all survey situations, we must worry about how nonrespondents may differ from respondents. Perhaps in this example, people approving of the smoking ban were more likely to respond to the survey. Or perhaps they were less likely to respond. In any case, we must be careful in interpreting results of the survey.

We will assume that the respondents are representative of a larger group of employees who would have responded to the questionnaire if they had received it. Our inferences apply only to this subgroup of the entire employee population. We also assume that the respondents answered independently. If these assumptions hold, then under the null hypothesis the test statistic has approximately the chi-square distribution with $(3 - 1)(3 - 1) = 4$ degrees of freedom.

We will use significance level $\alpha = .005$. Referring to Table E, we find $c = 14.86$. The acceptance region is $[0, 14.86)$, the rejection region is $[14.86, \infty)$, and the decision rule is:

If test statistic < 14.86, say the results are consistent with the null hypothesis that smoking status and approval of the smoking ban are independent.
If test statistic ≥ 14.86, say the results are inconsistent with the null hypothesis, suggesting that smoking status and approval of the smoking ban are not independent.

If smoking status and approval of the smoking ban were independent in this employee population, we would expect the same distribution for smoking status within each approval category. Equivalently, we would expect the same distribution for approval of the ban within each smoking status category. The expected frequencies under the null hypothesis are shown below:

	Approval of the smoking ban		
Smoking status	Approve	Do not approve	Not sure
Never smoked	$\frac{250 \times 367}{434} = 211.4$	$\frac{250 \times 39}{434} = 22.5$	$\frac{250 \times 28}{434} = 16.1$
Ex-smoker	$\frac{117 \times 367}{434} = 98.9$	$\frac{117 \times 39}{434} = 10.5$	$\frac{117 \times 28}{434} = 7.5$
Current smoker	$\frac{67 \times 367}{434} = 56.7$	$\frac{67 \times 39}{434} = 6.0$	$\frac{67 \times 28}{434} = 4.3$

From these observed and expected frequencies, we calculate the test statistic:

$$\begin{aligned}\text{Test statistic} &= \frac{(237 - 211.4)^2}{211.4} + \frac{(3 - 22.5)^2}{22.5} + \frac{(10 - 16.1)^2}{16.1} \\ &\quad + \frac{(106 - 98.9)^2}{98.9} + \frac{(4 - 10.5)^2}{10.5} + \frac{(7 - 7.5)^2}{7.5} \\ &\quad + \frac{(24 - 56.7)^2}{56.7} + \frac{(32 - 6.0)^2}{6.0} + \frac{(11 - 4.3)^2}{4.3} \\ &= 3.1 + 16.9 + 2.3 + .5 + 4.0 + .0 + 18.9 + 112.7 + 10.4 \\ &= 168.8\end{aligned}$$

This large test statistic is in the rejection region. The survey results strongly suggest that smoking status and approval of the smoking ban are not independent in the population of employees who would have responded to the questionnaire.

Another way to look at the responses is with the two-way frequency table in Table 16-3. Looking at row percentages, we see that 95% of the respondents who had never smoked and 91% of the former smokers approved the smoking ban. Only 36% of the current smokers approved.

The column percentages in Table 16-3 show that among respondents who approved the smoking ban, 65% had never smoked, 29% were former smokers, and 7% were current smokers. In contrast, 82% of respondents who

TABLE 16-3 Frequency table for Example 16-3. Row percentages are shown in parentheses to the right of the observed frequencies. Column percentages are shown in parentheses below the observed frequencies.

	Approval of the smoking ban			
Smoking status	Approve	Do not approve	Not sure	Total
Never smoked	237 (95) (65)	3 (1) (8)	10 (4) (36)	250 (58)
Ex-smoker	106 (91) (29)	4 (3) (10)	7 (6) (25)	117 (27)
Current smoker	24 (36) (7)	32 (48) (82)	11 (16) (39)	67 (15)
Total	367 (85)	39 (9)	28 (6)	434

did not approve the smoking ban were current smokers. This is strong evidence that smoking status and approval of the smoking ban were not independent in the employee population of potential respondents.

In Section 16-4, we discuss the chi-square test of homogeneity for comparing the distribution of a qualitative variable across several populations.

16-4

Comparing the Distribution of a Qualitative Variable Across Populations

Suppose we have independent random samples, one from each of c populations. We can classify an observation into exactly one of r categories of a qualitative variable we will call variable I. We want to test the null hypothesis that the probability distribution of variable I is the same in each population. The alternative states that the probability distribution of variable I is not the same in all c populations. If the null hypothesis is true, we say the populations are *homogeneous* (alike or the same) with respect to the probability distribution of variable I. We test our hypotheses using the *chi-square test of homogeneity*.

We can think of population as a second qualitative variable in this situation, say variable II. The test statistic and the steps for carrying out the chi-square test of homogeneity are exactly the same as we outlined for the chi-square test of independence, in Section 16-3.

Let's illustrate the chi-square test of homogeneity with an example.

EXAMPLE 16-4 We will compare the distributions of body weight classification for 20–24-year-old women in Britain, Canada, and the United States. We base our inferences on results of surveys carried out during 1976–1981 (Millar and Stephens,

1987). Investigators obtained the samples from the noninstitutionalized populations in the three countries. They used similar techniques in all three surveys.

The investigators defined body weight classification in terms of the Quetelet index. The Quetelet index is weight in kilograms divided by the square of height in meters (so it has units kilograms/meter^2).

A woman is classified as underweight if her Quetelet index is less than or equal to 20 kg/m^2. She has a normal body weight classification if her Quetelet index is greater than 20 kg/m^2 and less than or equal to 25 kg/m^2. She is classified as overweight if her Quetelet index is greater than 25 kg/m^2 and less than or equal to 30 kg/m^2. She has the obese body weight classification if her Quetelet index is greater than 30 kg/m^2.

The researchers studied 547 20–24-year-old women from Britain, 873 from Canada, and 624 from the United States. The women in the United States sample were all white. (Why?) When each woman is classified by country and body weight, we have the following results:

Body weight classification	Britain	Canada	United States	Total
Underweight	126	297	156	579
Normal	306	498	349	1,153
Overweight	88	61	75	224
Obese	27	17	44	88
Total	547	873	624	2,044

We want to test the hypotheses:

H_0: The distribution of 20–24-year-old women across body weight classifications is the same for Britain, Canada, and the United States.

H_a: The distribution of 20–24-year-old women across body weight classifications is not the same for Britain, Canada, and the United States.

We must assume that the sample from each country is a random sample from the population of noninstitutionalized 20–24-year-old (white) women in that country. We also assume that the measurement techniques for determining the Quetelet index were uniform and independent for different women. Then under the null hypothesis, the test statistic has approximately the chi-square distribution with $(4 - 1)(3 - 1) = 6$ degrees of freedom.

We will use significance level $\alpha = .005$. From Table E, we find $c = 18.55$. The acceptance region is $[0, 18.55)$, the rejection region is $[18.55, \infty)$, and the decision rule is:

If test statistic < 18.55, say the results are consistent with the null hypothesis.
If test statistic ≥ 18.55, say the results are inconsistent with the null hypothesis.

The expected frequencies under the null hypothesis are:

Body weight	Britain	Canada	United States
Underweight	$\frac{579 \times 547}{2{,}044} = 154.9$	$\frac{579 \times 873}{2{,}044} = 247.3$	$\frac{579 \times 624}{2{,}044} = 176.8$
Normal	$\frac{1{,}153 \times 547}{2{,}044} = 308.6$	$\frac{1{,}153 \times 873}{2{,}044} = 492.5$	$\frac{1{,}153 \times 624}{2{,}044} = 352.0$
Overweight	$\frac{224 \times 547}{2{,}044} = 59.9$	$\frac{224 \times 873}{2{,}044} = 95.7$	$\frac{224 \times 624}{2{,}044} = 68.4$
Obese	$\frac{88 \times 547}{2{,}044} = 23.5$	$\frac{88 \times 873}{2{,}044} = 37.6$	$\frac{88 \times 624}{2{,}044} = 26.9$

From these observed and expected frequencies, we calculate the test statistic:

$$\begin{aligned}\text{Test statistic} = {} & \frac{(126 - 154.9)^2}{154.9} + \frac{(297 - 247.3)^2}{247.3} + \frac{(156 - 176.8)^2}{176.8} \\ & + \frac{(306 - 308.6)^2}{308.6} + \frac{(498 - 492.5)^2}{492.5} + \frac{(349 - 352.0)^2}{352.0} \\ & + \frac{(88 - 59.9)^2}{59.9} + \frac{(61 - 95.7)^2}{95.7} + \frac{(75 - 68.4)^2}{68.4} \\ & + \frac{(27 - 23.5)^2}{23.5} + \frac{(17 - 37.6)^2}{37.6} + \frac{(44 - 26.9)^2}{26.9} \\ = {} & 5.39 + 9.99 + 2.45 + .02 + .06 + .03 \\ & + 13.18 + 12.58 + .64 + .52 + 11.29 + 10.87 = 67.0\end{aligned}$$

The test statistic is in the rejection region. This comparison of survey results strongly suggests that the distribution of body weight classification for 20–24-year-old women is not the same for Britain, Canada, and the United States.

Let's compare the observed and expected frequencies. In the samples from Britain and the United States, there were fewer underweight women and more overweight women than expected under the null hypothesis. In the sample from Canada, there were more underweight women and fewer overweight women than expected under the null hypothesis. We could also look at a frequency table showing row and column percentages, as we did in Table 16-3. What conclusions would you draw based on an analysis of these surveys?

Suppose we have independent random samples from each of several populations. An observation can be classified into exactly one of two categories of a qualitative variable. Call these categories success and failure. We want to compare the proportion in the success category across the populations. Comparing proportions across several populations is a special case of the situation we have discussed in this section, so we can apply the chi-square test of homogeneity.

Now suppose we want to compare two proportions. That is, we want to compare the proportion in the success category for two populations. Then we

use the chi-square test of homogeneity, with 1 degree of freedom. Resulting *p*-values are the same as *p*-values we get using the large-sample test for two proportions based on the standard Gaussian distribution (Section 11-2).

In Section 16-5, we discuss an exact test for association in a 2 × 2 frequency table, based on a hypergeometric distribution.

16-5

Testing for Association in a 2 × 2 Frequency Table, Using a Hypergeometric Distribution

We will consider two sampling situations. In each situation, we want to test for association in a 2 × 2 frequency table.

In the first situation, we have a random sample from a population. An observation is classified according to each of two qualitative variables—variables I and II. Each variable has two categories. We want to test the hypotheses:

Null hypothesis: The two variables are independent in the population.
Alternative hypothesis: The two variables are not independent in the population.

In the second situation, we have two independent random samples, one from each of two populations. An observation is classified into one of two categories of a qualitative variable—call it variable I. We can think of population as variable II. We want to test the hypotheses:

Null hypothesis: The distribution of variable I is the same in the two populations.
Alternative hypothesis: The distribution of variable I is not the same in the two populations.

For both sampling situations, we can use a test of association called *Fisher's exact test.* The test is based on a hypergeometric distribution, as outlined below.

The p-value approach to Fisher's exact test

1. We state one of the two sets of hypotheses just given, depending on the sampling situation.
2. Suppose we write the 2 × 2 frequency table for our sample as follows:

	Variable II		
Variable I	Category 1	Category 2	Total
Category 1	O_{11}	O_{12}	R_1
Category 2	O_{21}	O_{22}	R_2
Total	C_1	C_2	n

The test statistic equals the smallest of the observed frequencies O_{11}, O_{12}, O_{21}, and O_{22}.

3. Make the assumptions for one of the two sampling situations. Then under the null hypothesis, the conditional probability of the observed results, given the row totals R_1 and R_2 and the column totals C_1 and C_2, is a hypergeometric probability that we can write as

$$P(\text{observed results under } H_0 \text{ given observed row and column totals}) = \frac{R_1!\, R_2!\, C_1!\, C_2!}{O_{11}!\, O_{12}!\, O_{21}!\, O_{22}!\, n!}$$

4. Collect observations that satisfy the appropriate set of assumptions. Let the test statistic equal the smallest of the observed frequencies, O_{11}, O_{12}, O_{21}, and O_{22}.
5. Let m be the minimum of R_1, R_2, C_1, and C_2. Then m is the smallest frequency observed in a category for either variable I or variable II. There are $m + 1$ possible 2×2 frequency tables having row totals R_1 and R_2 and column totals C_1 and C_2. Given these fixed row and column totals, find the probability under the null hypothesis of each of these 2×2 frequency tables, using the formula given in step 3.

   Possible results as extreme as or more extreme (in the direction of the alternative) than those observed are those corresponding to frequency tables with probabilities less than or equal to the probability of the observed table. The *p-value* is the sum of all the probabilities that are less than or equal to the probability for the observed table. If we have a one-sided alternative, the p-value is the sum of only the probabilities in the direction indicated by the alternative.
6. If the p-value is large, say the results are consistent with the null hypothesis. If the p-value is small, say the results are inconsistent with the null hypothesis. Draw conclusions based on the experimental results.

Let's look at an example of the first sampling situation. We have a random sample from a population. Observations are classified according to each of two qualitative variables (each with two categories). We want to test the null hypothesis that the two variables are independent in the population.

EXAMPLE 16-5

Investigators studied the relationship between abnormal sex chromosome genotypes and criminal behavior among men in Denmark (Witkin et al., 1976). As indicated in Example 16-2, they included only men at least 184 centimeters tall. The investigators identified 16 men with an extra X chromosome (genotype XXY) and 12 men with an extra Y chromosome (genotype XYY).

Three of the 16 men with the XXY genotype and 5 of the 12 men with the XYY genotype had been convicted of at least one crime. Does this suggest any difference between men with the two genotypes with respect to conviction of a crime? We will use Fisher's exact test to address this question.

Define a variable called convictions, with two categories: none and at

least one. The variable genotype also has two categories: XXY and XYY. We want to test the hypotheses:

H_0: The variables genotype and convictions are independent in the population.
H_a: The variables genotype and convictions are not independent in the population.

As we know from Example 16-2, the investigators did not select these 28 men as a random sample from a larger population of men with XXY or XYY genotypes. Instead, they were identified from among over 4,000 Danish men at least 184 centimeters tall. For our purposes, we will assume that the relationship between genotype and convictions in the sample of 28 men is representative of the relationship between these two variables in the larger population of Danish men at least 184 centimeters tall, with either the XXY or XYY genotype. We also assume that the 28 observations are independent of one another. Then we can calculate conditional probabilities based on a hypergeometric distribution, under the null hypothesis.

We summarize the observations in a 2×2 frequency table:

	Genotype		
Convictions	XXY	XYY	Total
None	13	7	20
At least one	3	5	8
Total	16	12	28

The test statistic equals 3, the smallest frequency in the table. The conditional probability for the observed frequency table under the null hypothesis is

$$P(\text{observed results under } H_0, \text{ given } R_1 = 20, R_2 = 8, C_1 = 16, C_2 = 12) = \frac{\binom{8}{3}\binom{20}{13}}{\binom{28}{16}} = \frac{20!\,8!\,16!\,12!}{13!\,7!\,3!\,5!\,28!} = .1427$$

This hypergeometric probability equals the number of ways we could select 3 of the 8 men with at least one conviction and 13 of the 20 men with no convictions to have the XXY genotype, divided by the number of ways we could select 16 of the 28 men to have the XXY genotype. We could calculate this same probability as

$$P(\text{observed results under } H_0, \text{ given } R_1 = 20, R_2 = 8, C_1 = 16, C_2 = 12) = \frac{\binom{16}{3}\binom{12}{5}}{\binom{28}{8}} = \frac{20!\,8!\,16!\,12!}{13!\,7!\,3!\,5!\,28!} = .1427$$

This hypergeometric probability is the number of ways we could select 3 of the 16 XXY men and 5 of the 12 XYY men to have at least one conviction, divided by the number of ways we could select 8 of the 28 men to have at least one conviction.

The smallest row or column total is 8. There are $8 + 1 = 9$ possible 2×2 tables having the same row and column totals as the observed table. These nine possible frequency tables are listed in Table 16-4, along with the associated conditional probability under the null hypothesis.

TABLE 16-4 For Example 16-5, the conditional probability of each possible frequency table under the null hypothesis, given 16 men with the XXY genotype and 20 men with no convictions

Possible frequency table		Conditional probability under the null hypothesis
16	4	$\frac{20!\,8!\,16!\,12!}{16!\,4!\,0!\,8!\,28!} = .0002$
0	8	
15	5	$\frac{20!\,8!\,16!\,12!}{15!\,5!\,1!\,7!\,28!} = .0041$
1	7	
14	6	$\frac{20!\,8!\,16!\,12!}{14!\,6!\,2!\,6!\,28!} = .0357$
2	6	
13	7	$\frac{20!\,8!\,16!\,12!}{13!\,7!\,3!\,5!\,28!} = .1427$
3	5	
12	8	$\frac{20!\,8!\,16!\,12!}{12!\,8!\,4!\,4!\,28!} = .2899$
4	4	
11	9	$\frac{20!\,8!\,16!\,12!}{11!\,9!\,5!\,3!\,28!} = .3092$
5	3	
10	10	$\frac{20!\,8!\,16!\,12!}{10!\,10!\,6!\,2!\,28!} = .1700$
6	2	
9	11	$\frac{20!\,8!\,16!\,12!}{9!\,11!\,7!\,1!\,28!} = .0442$
7	1	
8	12	$\frac{20!\,8!\,16!\,12!}{8!\,12!\,8!\,0!\,28!} = .0041$
8	0	

Possible frequency tables at least as extreme (in the direction of the alternative) as observed are those with probabilities in Table 16-4 less than or equal to .1427. The p-value is the sum of the corresponding probabilities, so p-value = .231. This p-value is fairly large, so our results are consistent with the null hypothesis. From this sample, we cannot see an association between convictions (none or at least one) and genotype (XXY or XYY).

Now let's look at an example of the second sampling situation. We have two independent random samples, one from each of two populations. An observation can be classified into exactly one of two categories of a qualitative variable. We want to test the null hypothesis that the variable has the same distribution in both populations. Equivalently, we want to test whether the proportion in the "success" category is the same in the two populations.

EXAMPLE 16-6 Researchers wanted to study modes of transmission of common cold viruses (Dick et al., 1986). Twenty-four uninfected men and 16 men infected with the common cold participated. Volunteers played poker at tables of five, with two infected and three uninfected men at each table, for 12 hours. Twelve uninfected volunteers were restrained with large collars or arm braces to prevent infection by contact; thus, they could be infected only by airborne viruses. The other 12 uninfected volunteers were not restrained; they could be infected through either airborne viruses or hand contamination. The researchers monitored all 24 uninfected volunteers after the experiment, for subsequent infection with a common cold. The results of the experiment follow.

Developed a common cold	Experimental restraint		Total
	Restrained	Unrestrained	
Yes	6	11	17
No	6	1	7
Total	12	12	24

Does the restraint affect the likelihood of developing a cold? We will test the hypotheses:

H_0: The proportion developing a common cold is the same for restrained and unrestrained volunteers.

H_a: The proportion developing a common cold is smaller for restrained than for unrestrained volunteers.

The two populations sampled are hypothetical. We assume that the volunteers are representative of a larger group of interest. The first experimental population consists of the (hypothetical) responses of this larger group if they were all exposed under the restrained condition. The second experimental

population consists of their responses if they were all exposed under the unrestrained condition.

We must assume that the observations are independent. That is, one volunteer developing a cold (or not) does not in any way affect the likelihood of another volunteer developing a cold. We also assume that the volunteers in the restrained and unrestrained experimental groups have similar likelihoods of developing colds if exposed to similar conditions. The best way to ensure this is to select volunteers as similar as possible and then randomly divide them into two experimental groups. We have no way of checking any of these assumptions without more information on how the experiment was conducted.

The test statistic equals 1, the smallest frequency in the table. The conditional probability under the null hypothesis of the observed frequency table is

$$P(\text{observed results under } H_0, \text{ given } R_1 = 17, R_2 = 7, C_1 = 12, C_2 = 12) = \frac{\binom{12}{1}\binom{12}{6}}{\binom{24}{7}} = \frac{12!\,12!\,7!\,17!}{1!\,11!\,6!\,6!\,24!} = .032$$

This hypergeometric probability equals the number of ways we could select 1 of the 12 unrestrained volunteers and 6 of the 12 restrained volunteers to remain cold-free, divided by the number of ways we could select 7 of the 24 volunteers to remain cold-free.

What possible results are more extreme (in the direction of the alternative) than those observed? The alternative hypothesis states that restrained volunteers are less likely to develop a cold than unrestrained volunteers. Therefore, seeing no unrestrained volunteers remain free of a cold (test statistic = 0) would be more extreme in the direction of the alternative. The associated frequency table is shown in Table 16-5.

TABLE 16-5 This table displays results that are more extreme in the direction of the alternative hypothesis (that restrained volunteers are less likely to develop colds than unrestrained volunteers) than the observed results in Example 16-6

Developed a common cold	Experimental restraint		Total
	Restrained	Unrestrained	
Yes	5	12	17
No	7	0	7
Total	12	12	24

Under the null hypothesis, the conditional probability for the frequency table in Table 16-5 is

$$\frac{\binom{12}{0}\binom{12}{7}}{\binom{24}{7}} = \frac{12!\ 12!\ 7!\ 17!}{0!\ 12!\ 7!\ 5!\ 24!} = .002$$

This hypergeometric probability is the number of ways we could select none of the 12 unrestrained volunteers and 7 of the 12 restrained volunteers to remain cold-free, divided by the number of ways we could select 7 of the 24 volunteers to remain cold-free.

The p-value equals .032 + .002 = .034. A p-value of .034 is fairly small, so we say the results are inconsistent with the null hypothesis. For the conditions and type of volunteer involved, restrained men seem to be less likely to develop colds than unrestrained men. This experiment suggests that people subject to both hand contamination and airborne transmission are more likely to develop colds than are people exposed only to airborne transmission of viruses.

In most hypothesis testing situations, we greatly simplify the goals of the experiment. The investigators in Example 16-6 were interested in the comparison we made. However, they were also interested in seeing whether the restrained men developed colds at all. (Some previous studies had suggested that airborne viruses were not likely to cause cold infection.) The finding that 6 of 12 restrained volunteers did subsequently develop colds suggested that airborne viruses might be a significant mode of cold transmission.

Summary of Chapter 16

The chi-square goodness-of-fit test is used to make large-sample inferences about the distribution of a qualitative variable. We compare the observed distribution of frequencies across categories with a specified distribution. When there are two categories, we can use the chi-square goodness-of-fit test for large-sample inferences about a proportion. This test is equivalent to the large-sample test based on the standard Gaussian distribution. When the sample size is small, we can base inferences about a proportion on a binomial distribution.

We consider inferences about two qualitative variables for two sampling situations. In the first sampling situation, we have a random sample from a population. An observation is classified according to each of two qualitative variables. We want to test the null hypothesis that the two variables are independent in the population. If the sample size is large enough, we can use the chi-square test of independence.

In the second sampling situation, we have independent random samples, one from each of several populations. Each observation can be classified according to a qualitative variable. (We can think of population as a second qualitative variable.) We want to test the null hypothesis that the variable has the same distribution in each of the populations. With large enough sample sizes, we can use the chi-square test of homogeneity. We can use this test to compare two proportions when sample sizes are large; this test is equivalent to the large-sample test based on the standard Gaussian distribution.

The same large-sample test statistic is used for both sampling situations. If the sample size is large enough, we can compare the test statistic with the chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom. Here, r is the number of rows (categories of the first variable) and c the number of columns (categories of the second variable) in a two-way frequency table summarizing the observed results.

To decide whether the sample size is large enough to use a large-sample test about qualitative variables, we can apply Cochran's rule: No expected frequency should be less than 1 and no more than 20% of the expected frequencies less than 5.

If each of two qualitative variables has two categories, we can test for association using Fisher's exact test. Probabilities are based on a hypergeometric distribution. These null hypothesis probabilities are conditional on the observed row and column totals in the frequency table.

```
Expected counts are printed below observed counts

             C1        C2        C3    Total
   1        237         3        10      250
         211.41     22.47     16.13

   2        106         4         7      117
          98.94     10.51      7.55

   3         24        32        11       67
          56.66      6.02      4.32

Total       367        39        28      434

ChiSq =  3.099 + 16.866 +  2.329 +
         0.504 +  4.036 +  0.040 +
        18.823 +112.100 + 10.315 = 168.111
df = 4
1 cells with expected counts less than 5.0
```

FIGURE M16-2 CHISQUARE output for Example 16-3

Carrying Out a Chi-Square Test of Association

Suppose we want to use the chi-square test of association in a two-way frequency table, as in Sections 16-3 and 16-4. We put the table into Minitab and use the CHISQUARE command to calculate the test statistic. Consider the data in Example 16-3. We enter the data onto our worksheet and then use the CHISQUARE command:

```
MTB>  read c1-c3
DATA> 237 3 10
DATA> 106 4 7
DATA> 24 32 11
DATA> END
MTB>  chisquare c1-c3
```

We get the results shown in Figure M16-2.

We find the *p*-value with the CDF command:

```
MTB>  cdf 168.111 k1;
SUBC> chisquare 4.
MTB>  let k2=1-k1
MTB>  print k1 k2
K1      1.00000
K2      0
```

Minitab does not have a command for Fisher's exact test (Section 16-5).

Exercises for Chapter 16

For each exercise, describe the population(s) sampled, whether real or hypothetical. For each procedure, state the assumptions that make it valid. Do these

assumptions seem reasonable? What additional information would you like to have about the experiment? Describe the results of your analysis.

EXERCISE 16-1 In a genetic study of the relationship between tobacco mosaic virus and the 30-kD protein gene, scientists introduced a gene encoding the 30-kD protein into tobacco plants (Deom, Oliver, and Beachy, 1987). Of 40 seedlings from one such plant, 29 expressed the 30-kD protein and 11 did not. Of 100 seedlings from another plant, 93 expressed the 30-kD protein and 7 did not.

Let p denote the probability that a plant expresses the 30-kD protein. If the 30-kD protein gene is expressed from a single genetic locus, scientists expect a 3:1 ratio (with the 30-kD protein:without the 30-kD protein). That is, they expect $p = \frac{3}{4}$. (See Exercise 6-30.)

If the 30-kD protein gene is expressed at two genetic loci, scientists expect a 15:1 ratio (with the 30-kD protein:without the 30-kD protein). That is, they expect $p = \frac{15}{16}$.

State and test appropriate hypotheses, separately for the two plants. Discuss your findings.

EXERCISE 16-2 In a study of T-DNA insertion mutagenesis in *Arabidopsis thaliana* plants, scientists studied two traits: resistance to kanamycin (resistant or susceptible) and height (dwarf or tall). Scientists recorded the phenotypes (or observed characteristics) of offspring in five experiments with parent plants of the same genotype (or genetic make-up for these two traits). The numbers of offspring with each phenotype are shown below (Feldmann et al., 1989).

Experiment	Category 1 resistant/dwarf	Category 2 resistant/tall	Category 3 susceptible/tall
1	41	81	34
2	54	104	52
3	62	142	54
4	50	125	50
5	46	131	52

In each experiment, the scientists observed no offspring with the susceptible/dwarf phenotype. The researchers wanted to compare the observed results with the 1:2:1 ratio expected under Mendelian genetics. (That is, Mendelian genetics predicts $\frac{1}{4}$ of offspring in category 1, $\frac{1}{2}$ in category 2, and $\frac{1}{4}$ in category 3.)

State and test appropriate hypotheses for each of the five experiments. Discuss your findings.

EXERCISE 16-3 Researchers studied two characteristics of tomatoes: height (tall or dwarf) and leaf (cut-leaf or potato-leaf). The dominant characteristics are tall height and cut-leaf. In a dihybrid cross, both parents have one dominant and one recessive

gene for both characteristics. In one such dihybrid cross, the researchers observed the following results (Devore, 1982, page 525; from "Linkage Studies of the Tomato," *Trans. Royal Canadian Institute,* 1931, pages 1–19):

Category:	Tall, cut-leaf	Tall, potato-leaf	Dwarf, cut-leaf	Dwarf, potato-leaf
Observed frequency:	926	288	293	104

Mendelian genetics tells us that if the height and leaf characteristics are distributed independently in such a dihybrid cross, we expect a 9:3:3:1 ratio of the four phenotypes (see Exercise 6-30). That is, we expect $\frac{9}{16}$ of the offspring in the tall, cut-leaf category, $\frac{3}{16}$ of the offspring in the tall, potato-leaf category, $\frac{3}{16}$ in the dwarf, cut-leaf category, and $\frac{1}{16}$ in the dwarf, potato-leaf category.

State and test appropriate hypotheses. Discuss your findings.

EXERCISE 16-4

In a study of the cereal crop sorghum, scientists self-crossed red-seeded plants to produce offspring with red, yellow, and white seeds (Devore, 1982, page 529; from "A Genetic and Biochemical Study on Pericarp Pigments in a Cross Between Two Cultivars of Grain Sorghum, Sorghum Bicolor," *Heredity,* 1976, pages 413–416). Genetic theory predicts red, yellow, and white seeds in a ratio of 9:3:4. That is, the theory predicts that $\frac{9}{16}$ of the seeds will be red, $\frac{3}{16}$ will be yellow, and $\frac{4}{16}$ will be white. The results are shown below.

Category:	Red seeds	Yellow seeds	White seeds
Observed frequency:	195	73	100

State and test appropriate hypotheses. Discuss your findings.

EXERCISE 16-5

Is time of onset of labor in pregnant women uniform across the 24 hours? To address this question, researchers recorded the time of onset of labor for 1,186 pregnant women. They considered 24 1-hour time categories, beginning at midnight. Their observations are shown below (Devore, 1982, pages 527–528; from "The Hour of Birth," *British J. Preventive and Social Medicine,* 1953, pages 43–59).

Hour	Frequency	Hour	Frequency	Hour	Frequency	Hour	Frequency
1	52	7	58	13	21	19	47
2	73	8	47	14	31	20	34
3	89	9	48	15	40	21	36
4	88	10	53	16	24	22	44
5	68	11	47	17	37	23	78
6	47	12	34	18	31	24	59

a. Plot the observations.

b. Test the null hypothesis that labor is just as likely to begin in any of the 24 1-hour periods.

c. Discuss your findings.

EXERCISE 16-6

In Exercise 11-2, we considered a study of two drugs that suppress abnormal heart rhythms, encainide and flecainide (*Science News,* April 29, 1989, volume 135, page 260). In this study, 730 patients received one of these two experimental drugs for cardiac arrhythmia and 730 patients received placebo. In an early review of results, a safety monitoring board found 33 of the 730 patients in the experimental drug group had experienced either sudden cardiac death or a nonfatal heart attack, compared with 9 of the 730 patients in the placebo group.

a. Use a large-sample test based on a chi-square distribution to test the null hypothesis that the probability of suffering heart attack or cardiac death is the same for the two treatments.

b. Compare these results with the results based on the standard Gaussian distribution (Exercise 11-2).

EXERCISE 16-7

Researchers wanted to investigate the response of acute myelogenous leukemia patients to an immunotherapy (Granatek et al., 1981). The researchers measured antibody response on 13 patients before and after immunotherapy treatment. At the end of the study period, they classified patients by the percent change in antibody response following treatment (greater than 40%, less than or equal to 40%) and by survival time (at least 160 weeks, less than 160 weeks).

	Percent change in antibody response	
Survival time	Greater than 40%	Less than or equal to 40%
At least 160 weeks	4	2
Less than 160 weeks	1	6

Use Fisher's exact test to test the null hypothesis that there is no association between these two qualitative variables in patients similar to those in the study.

EXERCISE 16-8

In Exercise 4-27 we considered a study of the response of beetles to an airborne sex pheromone. Scientists exposed 30 beetles at each of four dose rates (units not given) and recorded the number of beetles responding within 60 seconds (Nordheim, Tsiatis, and Shapas, 1983):

Dose rate	Number of beetles responding within 60 seconds
10^{-6}	2
10^{-5}	10
10^{-4}	17
10^{-3}	25

Test the null hypothesis that there is no difference in proportion of beetles responding within 60 seconds across the four dose rates. Discuss your findings.

EXERCISE 16-9 In a study in Western Australia, investigators classified adults by obesity and hypertension (Knuiman and Speed, 1988):

Hyper-tension	Obesity category		
	Low	Average	High
Yes	32	40	59
No	133	121	106

Test the null hypothesis that obesity classification and hypertension are independent in the population sampled. Discuss your findings.

EXERCISE 16-10 In a study of randomly selected Wisconsin high school seniors, investigators classified students by their college plans and by level of parental encouragement to attend college (Bonney, 1987; from Fienberg, 1977, page 101; originally from Sewell and Shah, 1968). Results are shown below separately for male and female students, all with high IQ scores and high socioeconomic status.

Males:	Plans to attend college	Parental encouragement	
		High	Low
	Yes	414	8
	No	54	17

Females:	Plans to attend college	Parental encouragement	
		High	Low
	Yes	360	13
	No	98	49

a. Test the null hypothesis that plans to attend college are independent of level of parental encouragement among high-IQ, high-socioeconomic-status male high school seniors.

b. Test the null hypothesis that plans to attend college are independent of level of parental encouragement among high-IQ, high-socioeconomic-status female high school seniors.

c. Does the relationship between plans to attend college and level of parental encouragement seem to be the same for males and females in the high-IQ, high-socioeconomic-status group?

EXERCISE 16-11 Women under 50 years of age who had breast cancer with minimal inflammation were classified by appearance of the tumor and 3-year survival (Bonney, 1987; from Morrison et al., 1973). The results for a group of patients diagnosed in Boston, Massachusetts, are shown below.

	Appearance	
3-year survival	Malig-nant	Benign
Yes	11	24
No	6	7

Is 3-year survival independent of tumor appearance among breast cancer patients under 50 years of age? State and test appropriate hypotheses.

EXERCISE 16-12 Researchers studied minor psychiatric disorders among patients receiving primary medical care (Grayson, 1987; from Goldberg et al., 1987). Two symptoms checked for 283 such patients were irritability and poor sleeping. The observed results are shown below.

	Irritability	
Poor sleeping	Yes	No
Yes	77	28
No	60	118

Does there appear to be an association between irritability and poor sleeping in patients receiving primary medical care? State and test appropriate hypotheses. Discuss the experimental results.

EXERCISE 16-13 Researchers painted a tobacco condensate at one of two dose levels onto the backs of mice. They treated 100 mice at each dose level. After 546 days, researchers classified each mouse into one of three categories: developed a skin tumor during the experimental period, died without a tumor before the end of the experiment, alive and no tumor at the end of the experiment. The results are shown below (Gart and Tarone, 1987; from Gart, 1976).

Termination category	Low dose	High dose
Developed tumor before end of experiment	34	53
Died without tumor before end of experiment	17	17
Alive and no tumor at end of experiment	49	30

Is the distribution across the three termination categories the same for the two dose levels? State and test appropriate hypotheses.

EXERCISE 16-14 If a child with sickle cell disease develops an overwhelming infection (called sepsis), he or she has a 30% chance of dying from it. In this study, researchers sought to prevent sepsis in children with sickle cell disease (Kolata, 1987). Of 215 children with sickle cell disease, researchers treated 110 with oral penicillin. They gave the other 105 children a placebo. The researchers followed the children for 15 months, on average. Over that period, 13 children suffered an overwhelming infection in the placebo group and 3 died. In the penicillin group, 2 children suffered an overwhelming infection and none died.

a. Test the null hypothesis that the probability of an overwhelming infection is the same for the two treatments.

b. Use Fisher's exact test to test the null hypothesis that the probability of death is the same for the two treatments.

c. Discuss your findings.

EXERCISE 16-15 In a study of the relationship between depression and coronary artery disease, researchers diagnosed 9 cases of major depression among 52 patients with newly diagnosed coronary artery disease. One year later, 14 of the 43 nondepressed patients had had at least one serious cardiac complication, compared with 7 of the 9 depressed patients (*Science News,* January 7, 1989, volume 135, page 13). Based on this sample, does there appear to be an association between cardiac complications and depression in patients with coronary artery disease?

a. Use a large-sample test to test the null hypothesis that depression and cardiac complications are independent in patients with coronary artery disease.

b. Use Fisher's exact test to test the null hypothesis that depression and cardiac complications are independent in patients with coronary artery disease.

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

EXERCISE 16-16 Early research suggested an association between chronic fatigue syndrome and the Epstein–Barr virus. To study this possibility, some scientists designed an experiment to test the effects of a drug known to stop replication of the Epstein–Barr virus. The experiment included 24 patients with Epstein–Barr

antibodies and a history of chronic debilitating fatigue (*Science News,* January 7, 1989, volume 135, page 4). Half the patients received the drug and half a placebo, by intravenous injection every 8 hours for 7 days, followed by 30 days of oral doses. Eleven of the 12 patients on the experimental drug reported improvement at the end of the study, compared with 10 of the 12 placebo patients. How do these experimental findings contribute to an assessment of the link between the Epstein–Barr virus and chronic fatigue syndrome? State and test appropriate hypotheses, using Fisher's exact test. Discuss your findings.

EXERCISE 16-17 Can the nitrous oxide used as an anesthetic by dental workers contribute to infertility? Researchers divided 24 female rats into two groups. They exposed 12 rats to doses of nitrous oxide comparable to what a dentist might breathe, 8 hours a day for 35 days. The other 12 rats served as controls. When mated, 6 of the 12 exposed rats and all 12 of the control rats conceived (*Science News,* March 25, 1989, volume 135, page 182). What does this study suggest about the relationship between nitrous oxide exposure and infertility? State and test appropriate hypotheses, using Fisher's exact test. Discuss your findings.

EXERCISE 16-18 In Exercise 10-4, we considered a study of divers with a history of decompression sickness, or the bends (*Science News,* March 25, 1989, volume 135, page 188). Eleven of 30 such divers showed evidence of a heart defect known as patent foramen ovale. About 5% of the general population has this heart defect.

a. Use a small-sample test to test the null hypothesis that among divers with a history of decompression sickness, the proportion having this heart defect equals .05.

b. Compare the result in part (a) with the result of the large-sample test in Exercise 10-4.

EXERCISE 16-19 In the same investigation discussed in Example 16-4, investigators surveyed 516 20- to 24-year-old men in Britain, 819 in Canada, and 581 in the United States (Millar and Stephens, 1987). The men in the United States survey were all white. (Why?)

Define a variable body weight with two categories. The excessive category corresponds to a Quetelet index greater than 25 kg/m^2, the not excessive category to a Quetelet index less than or equal to 25 kg/m^2. The results of the surveys are summarized below:

Body weight	Britain	Canada	United States
Not excessive	402	614	395
Excessive	114	205	186

Compare the proportion of 20- to 24-year-old men in the excessive body weight category across the three countries. State and test appropriate hypotheses. Discuss your findings.

EXERCISE 16-20 Carry out the chi-square test of independence in Example 16-5. Compare your results with the results of Fisher's exact test in Example 16-5.

EXERCISE 16-21 Carry out the chi-square test of homogeneity in Example 16-6. Compare your results with the results of Fisher's exact test in Example 16-6.