CHAPTER 11

REGRESSION ANALYSIS: SOME ADDITIONAL TECHNIQUES

CHAPTER OVERVIEW

This chapter provides an introduction to some additional tools and concepts that are useful in regression analysis. The presentation includes expansions of the basic ideas and techniques of regression analysis that were introduced in Chapters 9 and 10.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand how to include qualitative variables in a regression analysis.
2. understand how to use automated variable selection procedures to develop regression models.
3. be able to perform logistic regression for dichotomous and polytomous dependent variables.

11.1 INTRODUCTION

The basic concepts and methodology of linear regression analysis are covered in Chapters 9 and 10. In Chapter 9 we discuss the situation in which the objective is to obtain an equation that can be used to make predictions and estimates about some dependent variable from knowledge of some other single variable that we call the independent, predictor, or explanatory variable. In Chapter 10 the ideas and techniques learned in Chapter 9 are expanded to cover the situation in which it is believed that the inclusion of information on two or more independent variables will yield a better equation for use in making predictions and estimations. Regression analysis is a complex and powerful statistical tool that is widely employed in health sciences research. To do the subject justice requires more space than is available in an introductory statistics textbook. However, for the benefit of those who wish additional coverage of regression analysis, we present in this chapter some additional topics that should prove helpful to the student and practitioner of statistics.

Regression Assumptions Revisited As we learned in Chapters 9 and 10, there are several assumptions underlying the appropriate use of regression procedures. Often there are certain measurements that strongly influence the shape of a distribution or impact the magnitude of the variance of a measured variable. Other times, certain independent variables that are being used to develop a model are highly correlated, leading to the development of a model that may not be unique or correct.

Non-Normal Data Many times the data that are used to build a regression model are not normally distributed. One may wish to explore the possibility that some of the observed data points are outliers or that they disproportionately affect the distribution of the data. Such an investigation may be accomplished informally by constructing a scatter plot and looking for observations that do not seem to fit with the others. Alternatively, many computer packages produce formal tests to evaluate potential outlying observations in either the dependent variable or the independent variables. It is always up to the researcher, however, to justify which observations are to be removed from the data set prior to analysis.

Often one may wish to attempt a transformation of the data. Mathematical transformations are useful because they do not affect the underlying relationships among variables. Since hypothesis tests for the regression coefficients are based on normal distribution statistics, data transformations can sometimes normalize the data to the extent necessary to perform such tests. Simple transformations, such as taking the square root of measurements or taking the logarithm of measurements, are quite common.

EXAMPLE 11.1.1

Researchers were interested in blood concentrations of delta-9-tetrahydrocannabinol (Δ-9-THC), the active psychotropic component in marijuana, from 25 research subjects. These data are presented in Table 11.1.1, as are these same data after using a $\log_{10}$ transformation.

TABLE 11.1.1 Data from a Random Sample of 25 Research Subjects Tested for Δ-9-THC, Example 11.1.1

Case No.	Concentration (μg/ml)	$\log_{10}$ Concentration (μg/ml)
1	.30	−.52
2	2.75	.44
3	2.27	.36
4	2.37	.37
5	1.12	.05
6	.60	−.22
7	.61	−.21
8	.89	−.05
9	.33	−.48
10	.85	−.07
11	2.18	.34
12	3.59	.56
13	.28	−.55
14	1.90	.28
15	1.71	.23
16	.85	−.07
17	1.53	.18
18	2.25	.35
19	.88	−.05
20	.49	−.31
21	4.35	.64
22	.67	−.17
23	2.74	.44
24	.79	−.10
25	6.94	.84

Box-and-whisker plots from SPSS software for these data are shown in Figure 11.1.1. The raw data are clearly skewed, and an outlier is identified (observation 25). A $\log_{10}$ transformation, which is often useful for such skewed data, removes the magnitude of the outlier and results in a distribution that is much more nearly symmetric about the median. Therefore, the transformed data could be used in lieu of the raw data for constructing the regression model. Though symmetric data do not, necessarily, imply that the data are normal, they do result in a more appropriate model. Formal tests of normality, as previously mentioned, should always be carried out prior to analysis. ■

Unequal Error Variances When the variances of the error terms are not equal, we may obtain a satisfactory equation for the model, but, because the assumption that the error variances are equal is violated, we will not be able to perform appropriate hypothesis tests on the model coefficients. Just as was the case in overcoming the non-normality problem, transformations of the regression variables may reduce the impact of unequal error variances.

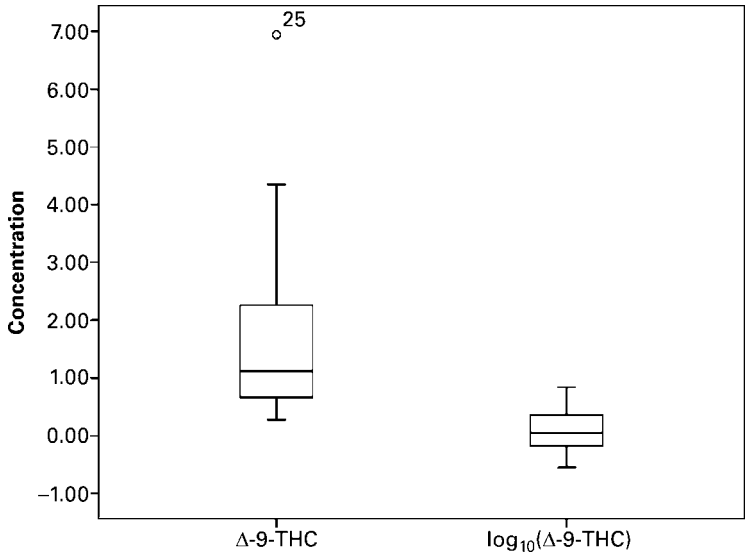


FIGURE 11.1.1 Box-and-whisker plots of data from Example 11.1.1.

Correlated Independent Variables *Multicollinearity* is a common problem that arises when one attempts to build a model using many independent variables. Multicollinearity occurs when there is a high degree of correlation among the independent variables. For example, imagine that we want to find an equation relating height and weight to blood pressure. A common variable that is derived from height and weight is called the body mass index (BMI). If we attempt to find an equation relating height, weight, and BMI to blood pressure, we can expect to run into analytical problems because BMI, by definition, is highly correlated with both height and weight.

The problem arises mathematically when the solutions for the regression coefficients are derived. Since the data are correlated, solutions may not be found that are unique to a given model. The least complex solution to multicollinearity is to calculate correlations among all of the independent variables and to retain only those variables that are not highly correlated. A conservative rule of thumb to remove redundancy in the data set is to eliminate variables that are related to others with a significant correlation coefficient above 0.7.

EXAMPLE 11.1.2

A study of obesity and metabolic syndrome used data collected from 15 students, and included systolic blood pressure (SBP), weight, and BMI. These data are presented in Table 11.1.2.

Correlations for the three variables are shown in Figure 11.1.2. The very large and significant correlation between the variables weight and BMI suggests that including both of these variables in the model is inappropriate because of the high level of redundancy in the information provided by these variables. This makes logical sense since BMI is a function of weight. The researcher is now faced with the task of deciding which of the variables to retain for constructing the regression model.

TABLE 11.1.2 Data from 8 Random Sample of 15 Students

Case No.	SBP	Weight (lbs.)	BMI
1	126	125	24.41
2	129	130	23.77
3	126	132	20.07
4	123	200	27.12
5	124	321	39.07
6	125	100	20.90
7	127	138	22.96
8	125	138	24.44
9	123	149	23.33
10	119	180	25.82
11	127	184	26.40
12	126	251	31.37
13	122	197	26.72
14	126	107	20.22
15	125	125	23.62

Correlations: SBP, Weight, BMI

	SBP	Weight
Weight	−0.289	
p-value	0.296	
BMI	−0.213	0.962
p-value	0.447	0.000

FIGURE 11.1.2 Correlations calculated in MINITAB software for the data in Example 11.1.2.

■

11.2 QUALITATIVE INDEPENDENT VARIABLES

The independent variables considered in the discussion in Chapter 10 were all quantitative; that is, they yielded numerical values that were either counts or measurements in the usual sense of the word. For example, some of the independent variables used in our examples and exercises were age, education level, collagen porosity, and collagen tensile strength. Frequently, however, it is desirable to use one or more qualitative variables as independent variables in the regression model. Qualitative variables, it will be recalled, are those variables whose "values" are categories and that convey the concept of attribute rather than amount or quantity. The variable marital status, for example, is a qualitative variable whose categories are "single," "married," "widowed," and "divorced." Other examples of qualitative variables include sex (male or female), diagnosis, race, occupation, and

immunity status to some disease. In certain situations an investigator may suspect that including one or more variables such as these in the regression equation would contribute significantly to the reduction of the error sum of squares and thereby provide more precise estimates of the parameters of interest.

Suppose, for example, that we are studying the relationship between the dependent variable systolic blood pressure and the independent variables weight and age. We might also want to include the qualitative variable sex as one of the independent variables. Or suppose we wish to gain insight into the nature of the relationship between lung capacity and other relevant variables. Candidates for inclusion in the model might consist of such quantitative variables as height, weight, and age, as well as qualitative variables such as sex, area of residence (urban, suburban, rural), and smoking status (current smoker, ex-smoker, never smoked).

Dummy Variables In order to incorporate a qualitative independent variable in the multiple regression model, it must be quantified in some manner. This may be accomplished through the use of what are known as *dummy variables*.

DEFINITION

A ***dummy variable*** **is a variable that assumes only a finite number of values (such as 0 or 1) for the purpose of identifying the different categories of a qualitative variable.**

The term "dummy" is used to indicate the fact that the numerical values (such as 0 and 1) assumed by the variable have no quantitative meaning but are used merely to identify different categories of the qualitative variable under consideration. Qualitative variables are sometimes called *indicator* variables, and when there are only two categories, they are sometimes called *dichotomous* variables.

The following are some examples of qualitative variables and the dummy variables used to quantify them:

Qualitative Variable	Dummy Variable
Sex (male, female):	$x_1 = \begin{cases} 1 \text{ for male} \\ 0 \text{ for female} \end{cases}$
Place of residence (urban, rural, suburban):	$x_1 = \begin{cases} 1 \text{ for urban} \\ 0 \text{ for rural and suburban} \end{cases}$
	$x_2 = \begin{cases} 1 \text{ for rural} \\ 0 \text{ for urban and suburban} \end{cases}$
Smoking status [current smoker, ex-smoker (has not smoked for 5 years or less), ex-smoker (has not smoked for more than 5 years), never smoked]:	$x_1 = \begin{cases} 1 \text{ for current smoker} \\ 0 \text{ for otherwise} \end{cases}$
	$x_2 = \begin{cases} 1 \text{ for ex-smoker}(\leq 5 \text{ years}) \\ 0 \text{ otherwise} \end{cases}$
	$x_3 = \begin{cases} 1 \text{ for ex-smoker}(> 5 \text{ years}) \\ 0 \text{ otherwise} \end{cases}$

Note in these examples that when the qualitative variable has k categories, $k - 1$ dummy variables must be defined for all the categories to be properly coded. This rule is applicable for any multiple regression containing an intercept constant. The variable sex, with two categories, can be quantified by the use of only one dummy variable, while three dummy variables are required to quantify the variable smoking status, which has four categories.

The following examples illustrate some of the uses of qualitative variables in multiple regression. In the first example we assume that there is no interaction between the independent variables. Since the assumption of no interaction is not realistic in many instances, we illustrate, in the second example, the analysis that is appropriate when interaction between variables is accounted for.

EXAMPLE 11.2.1

In a study of factors thought to be associated with birth weight, a simple random sample of 100 birth records was selected from the North Carolina 2001 Birth Registry (A-1). Table 11.2.1 shows, for three variables, the data extracted from each record. There are two independent variables: length of gestation (weeks), which is quantitative, and smoking status of mother (smoke), a qualitative variable. The dependent variable is birth weight (grams).

TABLE 11.2.1 Data from a Simple Random Sample of 100 Births from the North Carolina Birth Registry, Example 11.2.1

Case No.	Grams	Weeks	Smoke	Case No.	Grams	Weeks	Smoke
1	3147	40	0	51	3232	38	0
2	2977	41	0	52	3317	40	0
3	3119	38	0	53	2863	37	0
4	3487	38	0	54	3175	37	0
5	4111	39	0	55	3317	40	0
6	3572	41	0	56	3714	34	0
7	3487	40	0	57	2240	36	0
8	3147	41	0	58	3345	39	0
9	3345	38	1	59	3119	39	0
10	2665	34	0	60	2920	37	0
11	1559	34	0	61	3430	41	0
12	3799	38	0	62	3232	35	0
13	2750	38	0	63	3430	38	0
14	3487	40	0	64	4139	39	0
15	3317	38	0	65	3714	39	0
16	3544	43	1	66	1446	28	1
17	3459	45	0	67	3147	39	1
18	2807	37	0	68	2580	31	0
19	3856	40	0	69	3374	37	0
20	3260	40	0	70	3941	40	0

(*Continued*)

Case No.	Grams	Weeks	Smoke	Case No.	Grams	Weeks	Smoke
21	2183	42	1	71	2070	37	0
22	3204	38	0	72	3345	40	0
23	3005	36	0	73	3600	40	0
24	3090	40	1	74	3232	41	0
25	3430	39	0	75	3657	38	1
26	3119	40	0	76	3487	39	0
27	3912	39	0	77	2948	38	0
28	3572	40	0	78	2722	40	0
29	3884	41	0	79	3771	40	0
30	3090	38	0	80	3799	45	0
31	2977	42	0	81	1871	33	0
32	3799	37	0	82	3260	39	0
33	4054	40	0	83	3969	38	0
34	3430	38	1	84	3771	40	0
35	3459	41	0	85	3600	40	0
36	3827	39	0	86	2693	35	1
37	3147	44	1	87	3062	45	0
38	3289	38	0	88	2693	36	0
39	3629	36	0	89	3033	41	0
40	3657	36	0	90	3856	42	0
41	3175	41	1	91	4111	40	0
42	3232	43	1	92	3799	39	0
43	3175	36	0	93	3147	38	0
44	3657	40	1	94	2920	36	0
45	3600	39	0	95	4054	40	0
46	3572	40	0	96	2296	36	0
47	709	25	0	97	3402	38	0
48	624	25	0	98	1871	33	1
49	2778	36	0	99	4167	41	0
50	3572	35	0	100	3402	37	1

Source: John P. Holcomb, sampled and coded from North Carolina Birth Registry data found at *www.irss.unc.edu/ncvital/bfd1down.html*.

Solution: For the analysis, we quantify smoking status by means of a dummy variable that is coded 1 if the mother is a smoker and 0 if she is a nonsmoker. The data in Table 11.2.1 are plotted as a scatter diagram in Figure 11.2.1. The scatter diagram suggests that, in general, longer periods of gestation are associated with larger birth weights.

To obtain additional insight into the nature of these data, we may enter them into a computer and employ an appropriate program to perform further analyses. For example, we enter the observations $y_1 = 3147, x_{11} = 40, x_{21} = 0$, for the first case; $Y_2 = 2977, x_{12} = 41, x_{22} = 0$ for the second case; and so on. Figure 11.2.2 shows the computer output obtained with the use of the MINITAB multiple regression program.

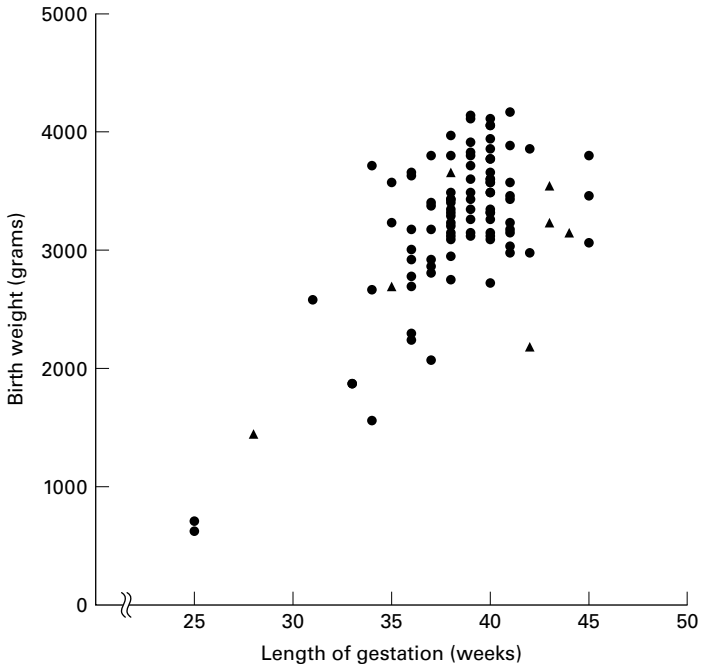


FIGURE 11.2.1 Birth weights and lengths of gestation for 100 births: (▲) smoking and (●) nonsmoking mothers.

The regression equation is

grams = −1724 + 130 x1 − 294 x2

Predictor	Coef	SE Coef	T	P
Constant	−1724.4	558.8	−3.09	0.003
weeks (x1)	130.05	14.52	8.96	0.000
smoke (x2)	−294.4	135.8	−2.17	0.033

S = 484.6 R-Sq = 46.4% R-Sq(adj) = 45.3%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	2	19689185	9844593	41.92	0.000
Residual Error	97	22781681	234863		
Total	99	42470867			

SOURCE	DF	Seq SS
x1	1	18585166
x2	1	1104020

FIGURE 11.2.2 Partial computer printout, MINITAB multiple regression analysis. Example 11.2.1.

We see in the printout that the multiple regression equation is

$$\begin{aligned}\hat{y}_j &= \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} \\ \hat{y}_j &= -1724.4 + 130.05 x_{1j} - 294.4 x_{2j}\end{aligned} \tag{11.2.1}$$

To observe the effect on this equation when we wish to consider only the births to smoking mothers, we let $x_{2j} = 1$. The equation then becomes

$$\begin{aligned}\hat{y}_j &= -1724.4 + 130.05 x_{1j} - 294.4(1) \\ &= -2018.8 + 130.05 x_{1j}\end{aligned} \tag{11.2.2}$$

which has a y-intercept of -2018.8 and a slope of 130. Note that the y-intercept for the new equation is equal to $(\hat{\beta}_0 + \hat{\beta}_1) = [-1724.4 + (-294.4)] = -2018$.

Now let us consider only births to nonsmoking mothers. When we let $x_2 = 0$, our regression equation reduces to

$$\begin{aligned}\hat{y}_j &= -1724.4 + 130.05 x_{1j} - 294(0) \\ &= -1724.4 + 130.05 x_{1j}\end{aligned} \tag{11.2.3}$$

The slope of this equation is the same as the slope of the equation for smoking mothers, but the y-intercepts are different. The y-intercept for the equation associated with nonsmoking mothers is larger than the one for the smoking mothers. These results show that for this sample, babies born to

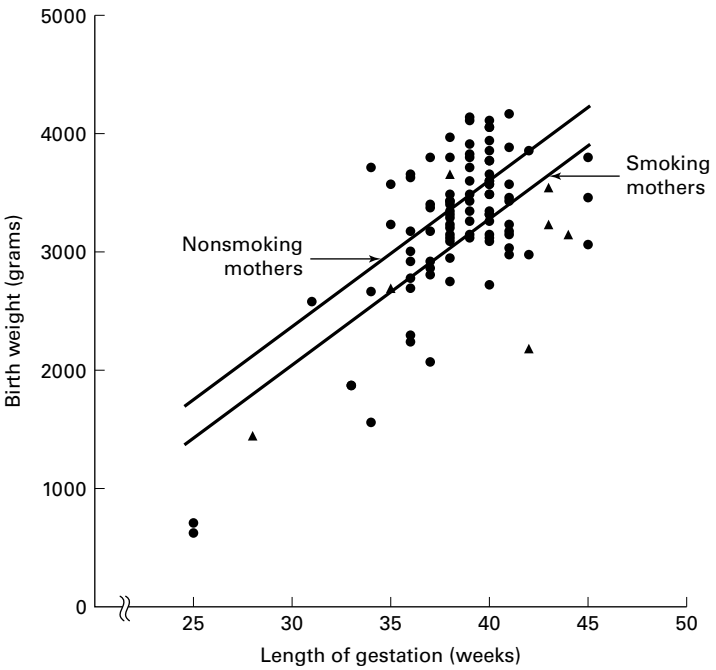


FIGURE 11.2.3 Birth weights and lengths of gestation for 100 births and the fitted regression lines: (▲) smoking and (●) nonsmoking mothers.

mothers who do not smoke weighed, on the average, more than babies born to mothers who do smoke, when length of gestation is taken into account. The amount of the difference, on the average, is 294 grams. Stated another way, we can say that for this sample, babies born to mothers who smoke weighed, on the average, 294 grams less than the babies born to mothers who do not smoke, when length of gestation is taken into account. Figure 11.2.3 shows the scatter diagram of the original data along with a plot of the two regression lines (Equations 11.2.2 and 11.2.3). ■

EXAMPLE 11.2.2

At this point a question arises regarding what inferences we can make about the sampled population on the basis of the sample results obtained in Example 11.2.1. First of all, we wish to know if the sample difference of 294 grams is significant. In other words, does smoking have an effect on birth weight? We may answer this question through the following hypothesis testing procedure.

Solution:

1. **Data.** The data are as given in Example 11.2.1.
2. **Assumptions.** We presume that the assumptions underlying multiple regression analysis are met.
3. **Hypotheses.** H_0: $\beta_2 = 0$; H_A: $\beta_2 \neq 0$. Suppose we let $\alpha = .05$.
4. **Test statistic.** The test statistic is $t = (\hat{\beta}_2 - 0)/s\hat{\beta}_2$.
5. **Distribution of test statistic.** When the assumptions are met and H_0 is true the test statistic is distributed as Student's t with 97 degrees of freedom.
6. **Decision rule.** We reject H_0 if the computed t is either greater than or equal to 1.9848 or less than or equal to -1.9848 (obtained by interpolation).
7. **Calculation of test statistic.** The calculated value of the test statistic appears in Figure 11.2.2 as the t ratio for the coefficient associated with the variable appearing in Column 4 of Table 11.2.1. This coefficient, of course, is $\hat{\beta}_2$. We see that the computed t is -2.17.
8. **Statistical decision.** Since $-2.17 < -1.9848$, we reject H_0.
9. **Conclusion.** We conclude that, in the sampled population, whether the mothers smoke is associated with a reduction in the birth weights of their babies.
10. ***p* value.** For this test we have $p = .033$ from Figure 11.2.2. ■

A Confidence Interval for β_2 Given that we are able to conclude that in the sampled population the smoking status of the mothers does have an effect on the birth weights of their babies, we may now inquire as to the magnitude of the effect. Our best

point estimate of the average difference in birth weights, when length of gestation is taken into account, is 294 grams in favor of babies born to mothers who do not smoke. We may obtain an interval estimate of the mean amount of the difference by using information from the computer printout by means of the following expression:

$$\hat{\beta}_2 \pm ts_{\hat{\beta}_2}$$

For a 95% confidence interval, we have

$$-294.4 \pm 1.9848(135.8)$$
$$(-563.9, -24.9)$$

Thus, we are 95% confident that the difference is somewhere between about 564 grams and 25 grams.

Advantages of Dummy Variables The reader may have correctly surmised that an alternative analysis of the data of Example 11.2.1 would consist of fitting two separate regression equations: one to the subsample of mothers who smoke and another to the subsample of those who do not. Such an approach, however, lacks some of the advantages of the dummy variable technique and is a less desirable procedure when the latter procedure is valid. If we can justify the assumption that the two separate regression lines have the same slope, we can get a better estimate of this common slope through the use of dummy variables, which entails pooling the data from the two subsamples. In Example 11.2.1 the estimate using a dummy variable is based on a total sample size of 100 observations, whereas separate estimates would be based on a sample of 85 smokers and only 15 nonsmokers. The dummy variables approach also yields more precise inferences regarding other parameters since more degrees of freedom are available for the calculation of the error mean square.

Use of Dummy Variables: Interaction Present Now let us consider the situation in which interaction between the variables is assumed to be present. Suppose, for example, that we have two independent variables: one quantitative variable X_1 and one qualitative variable with three response levels yielding the two dummy variables X_2 and X_3. The model, then, would be

$$y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{1j} X_{2j} + \beta_5 X_{1j} X_{3j} + \epsilon_j \tag{11.2.4}$$

in which $\beta_4 X_{1j} X_{2j}$ and $\beta_5 X_{1j} X_{3j}$ are called *interaction terms* and represent the interaction between the quantitative and the qualitative independent variables. Note that there is no need to include in the model the term containing $X_{2j} X_{3j}$; it will always be zero because when $X_2 = 1$, $X_3 = 0$, and when $X_3 = 1$, $X_2 = 0$. The model of Equation 11.2.4 allows for a different slope and Y-intercept for each level of the qualitative variable.

Suppose we use dummy variable coding to quantify the qualitative variable as follows:

$$X_2 = \begin{cases} 1 \text{ for level } 1 \\ 0 \text{ otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 \text{ for level } 2 \\ 0 \text{ otherwise} \end{cases}$$

The three sample regression equations for the three levels of the qualitative variable, then, are as follows:

Level 1 $(\boldsymbol{X_2 = 1, X_3 = 0})$

$$\begin{aligned}\hat{y}_j &= \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2(1) + \hat{\beta}_3(0) + \hat{\beta}_4 x_{1j}(1) + \hat{\beta}_5 x_{1j}(0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 + \hat{\beta}_4 x_{1j} \\ &= (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_4) x_{1j}\end{aligned} \tag{11.2.5}$$

Level 2 $(\boldsymbol{X_2 = 0, X_3 = 1})$

$$\begin{aligned}\hat{y}_j &= \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2(0) + \hat{\beta}_3(1) + \hat{\beta}_4 x_{1j}(0) + \hat{\beta}_5 x_{1j}(1) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_3 + \hat{\beta}_5 x_{1j} \\ &= (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_5) x_{1j}\end{aligned} \tag{11.2.6}$$

Level 3 $(\boldsymbol{X_2 = 0, X_3 = 0})$

$$\begin{aligned}\hat{y}_j &= \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2(0) + \hat{\beta}_3(0) + \hat{\beta}_4 x_{1j}(0) + \hat{\beta}_5 x_{1j}(0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{1j}\end{aligned} \tag{11.2.7}$$

Let us illustrate these results by means of an example.

EXAMPLE 11.2.3

A team of mental health researchers wishes to compare three methods (A, B, and C) of treating severe depression. They would also like to study the relationship between age and treatment effectiveness as well as the interaction (if any) between age and treatment. Each member of a simple random sample of 36 patients, comparable with respect to diagnosis and severity of depression, was randomly assigned to receive treatment A, B, or C. The results are shown in Table 11.2.2. The dependent variable Y is treatment effectiveness, the quantitative independent variable X_1 is patient's age at nearest birthday, and the independent variable type of treatment is a qualitative variable that occurs at three levels. The following dummy variable coding is used to quantify the qualitative variable:

$$X_2 = \begin{cases} 1 \text{ for treatment A} \\ 0 \text{ otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 \text{ for treatment B} \\ 0 \text{ otherwise} \end{cases}$$

The scatter diagram for these data is shown in Figure 11.2.4. Table 11.2.3 shows the data as they were entered into a computer for analysis. Figure 11.2.5 contains the printout of the analysis using the MINITAB multiple regression program.

Solution: Now let us examine the printout to see what it provides in the way of insight into the nature of the relationships among the variables. The least-squares equation is

$$\hat{y}_j = 6.21 + 1.03x_{1j} + 41.3x_{2j} + 22.7x_{3j} - .703x_{1j}x_{2j} - .510x_{1j}x_{3j}$$

TABLE 11.2.2 Data for Example 11.2.3

Measure of Effectiveness	Age	Method of Treatment
56	21	A
41	23	B
40	30	B
28	19	C
55	28	A
25	23	C
46	33	B
71	67	C
48	42	B
63	33	A
52	33	A
62	56	C
50	45	C
45	43	B
58	38	A
46	37	C
58	43	B
34	27	C
65	43	A
55	45	B
57	48	B
59	47	C
64	48	A
61	53	A
62	58	B
36	29	C
69	53	A
47	29	B
73	58	A
64	66	B
60	67	B
62	63	A
71	59	C
62	51	C
70	67	A
71	63	C

The three regression equations for the three treatments are as follows:

Treatment A (Equation 11.2.5)

$$\begin{aligned}\hat{y}_j &= (6.21 + 41.3) + (1.03 - .703)x_{1j} \\ &= 47.51 + .327x_{1j}\end{aligned}$$

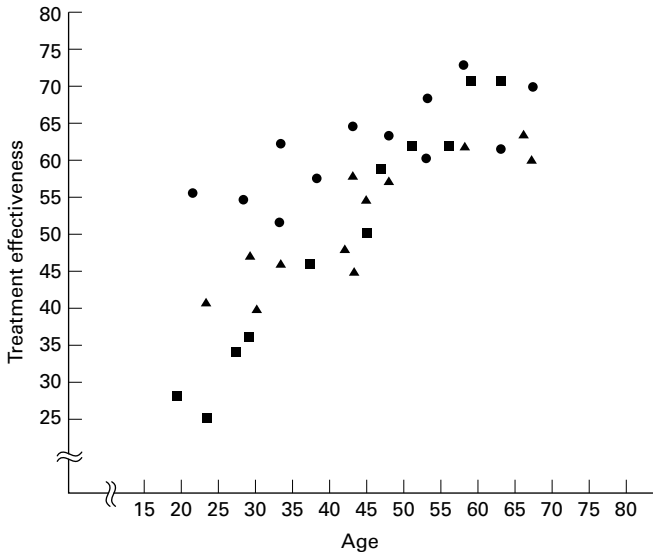


FIGURE 11.2.4 Scatter diagram of data for Example 11.2.3: (●) treatment A, (▲) treatment B, (■) treatment C.

Treatment B (Equation 11.2.6)

$$\begin{aligned}\hat{y}_j &= (6.21 + 22.7) + (1.03 - .510)x_{1j} \\ &= 28.91 + .520x_{1j}\end{aligned}$$

Treatment C (Equation 11.2.7)

$$\hat{y}_j = 6.21 + 1.03x_{1j}$$

Figure 11.2.6 contains the scatter diagram of the original data along with the regression lines for the three treatments. Visual inspection of Figure 11.2.6 suggests that treatments A and B do not differ greatly with respect to their slopes, but their y-intercepts are considerably different. The graph suggests that treatment A is better than treatment B for younger patients, but the difference is less dramatic with older patients. Treatment C appears to be decidedly less desirable than both treatments A and B for younger patients but is about as effective as treatment B for older patients. These subjective impressions are compatible with the contention that there is interaction between treatments and age.

Inference Procedures

The relationships we see in Figure 11.2.6, however, are sample results. What can we conclude about the population from which the sample was drawn?

For an answer let us look at the t ratios on the computer printout in Figure 11.2.5. Each of these is the test statistic

$$t = \frac{\hat{\beta}_i - 0}{s_{\hat{\beta}_i}}$$

TABLE 11.2.3 Data for Example 11.2.3 Coded for Computer Analysis

Y	X_1	X_2	X_3	X_1X_2	X_1X_3
56	21	1	0	21	0
55	28	1	0	28	0
63	33	1	0	33	0
52	33	1	0	33	0
58	38	1	0	38	0
65	43	1	0	43	0
64	48	1	0	48	0
61	53	1	0	53	0
69	53	1	0	53	0
73	58	1	0	58	0
62	63	1	0	63	0
70	67	1	0	67	0
41	23	0	1	0	23
40	30	0	1	0	30
46	33	0	1	0	33
48	42	0	1	0	42
45	43	0	1	0	43
58	43	0	1	0	43
55	45	0	1	0	45
57	48	0	1	0	48
62	58	0	1	0	58
47	29	0	1	0	29
64	66	0	1	0	66
60	67	0	1	0	67
28	19	0	0	0	0
25	23	0	0	0	0
71	67	0	0	0	0
62	56	0	0	0	0
50	45	0	0	0	0
46	37	0	0	0	0
34	27	0	0	0	0
59	47	0	0	0	0
36	29	0	0	0	0
71	59	0	0	0	0
62	51	0	0	0	0
71	63	0	0	0	0

for testing H_0: $\beta_i = 0$. We see by Equation 11.2.5 that the y-intercept of the regression line for treatment A is equal to $\hat{\beta}_0 + \hat{\beta}_2$. Since the t ratio of 8.12 for testing H_0: $\beta_2 = 0$ is greater than the critical t of 2.0423 (for $\alpha = .05$), we can reject H_0 that $\beta_2 = 0$ and conclude that the y-intercept of the population regression line for treatment A is different from the y-intercept of the population regression line for treatment C, which has a y-intercept of β_0. Similarly,

The regression equation is

$y = 6.21 + 1.03\,x1 + 41.3\,x2 + 22.7\,x3 - 0.703\,x4 - 0.510\,x5$

Predictor	Coef	Stdev	t-ratio	p
Constant	6.211	3.350	1.85	0.074
x1	1.03339	0.07233	14.29	0.000
x2	41.304	5.085	8.12	0.000
x3	22.707	5.091	4.46	0.000
x4	−0.7029	0.1090	−6.45	0.000
x5	−0.5097	0.1104	−4.62	0.000

s = 3.925 R-sq = 91.4% R-sq(adj) = 90.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	5	4932.85	986.57	64.04	0.000
Error	30	462.15	15.40		
Total	35	5395.00			

SOURCE	DF	SEQ SS
x1	1	3424.43
x2	1	803.80
x3	1	1.19
x4	1	375.00
x5	1	328.42

FIGURE 11.2.5 Computer printout, MINITAB multiple regression analysis, Example 11.2.3.

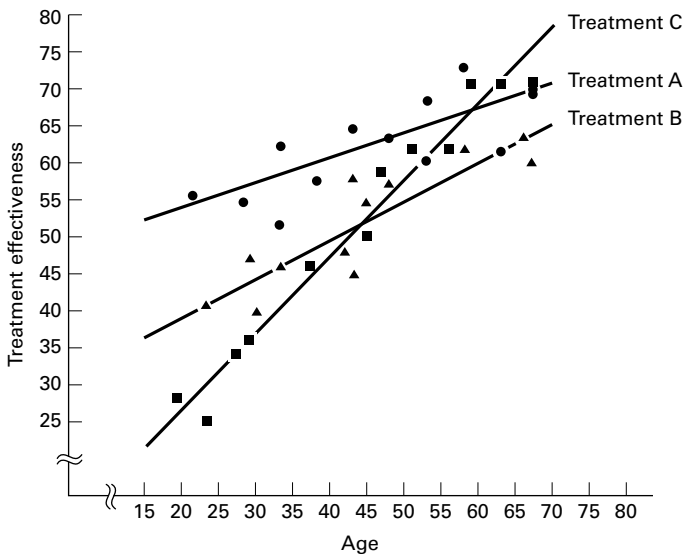


FIGURE 11.2.6 Scatter diagram of data for Example 11.2.3 with the fitted regression lines: (●) treatment A, (▲) treatment B, (■) treatment C.

since the t ratio of 4.46 for testing H_0: $\beta_3 = 0$ is also greater than the critical t of 2.0423, we can conclude (at the .05 level of significance) that the y-intercept of the population regression line for treatment B is also different from the y-intercept of the population regression line for treatment C. (See the y-intercept of Equation 11.2.6.)

Now let us consider the slopes. We see by Equation 11.2.5 that the slope of the regression line for treatment A is equal to $\hat{\beta}_1$ (the slope of the line for treatment C) $+\hat{\beta}_4$. Since the t ratio of -6.45 for testing H_0: $\beta_4 = 0$ is less than the critical t of -2.0423, we can conclude (for $\alpha = .05$) that the slopes of the population regression lines for treatments A and C are different. Similarly, since the computed t ratio for testing H_0: $\beta_5 = 0$ is also less than -2.0423, we conclude (for $\alpha = .05$) that the population regression lines for treatments B and C have different slopes (see the slope of Equation 11.2.6). Thus, we conclude that there is interaction between age and type of treatment. This is reflected by a lack of parallelism among the regression lines in Figure 11.2.6. ■

Another question of interest is this: Is the slope of the population regression line for treatment A different from the slope of the population regression line for treatment B? To answer this question requires computational techniques beyond the scope of this text. The interested reader is referred to books devoted specifically to regression analysis.

In Section 10.4 the reader was warned that there are problems involved in making multiple inferences from the same sample data. Again, books on regression analysis are available that may be consulted for procedures to be followed when multiple inferences, such as those discussed in this section, are desired.

We have discussed only two situations in which the use of dummy variables is appropriate. More complex models involving the use of one or more qualitative independent variables in the presence of two or more quantitative variables may be appropriate in certain circumstances. More complex models are discussed in the many books devoted to the subject of multiple regression analysis.

At this point it may be evident that there are many similarities between the use of a linear regression model using dummy variables and the basic ANOVA approach. In both cases, one is attempting to model the relationship between predictor variables and an outcome variable. In the case of linear regression, we are generally most interested in prediction, and in ANOVA, we are generally most interested in comparing means. If the desire is to compare means using regression, one could develop a model to predict mean response, say μ_i, instead of an outcome, y_i. Modeling the mean response using regression with dummy variables is equivalent to ANOVA. For the interested student, we suggest the book by Bowerman and O'Connell (1), who provide an example of using both approaches for the same data.

EXERCISES

For each exercise do the following:

(a) Draw a scatter diagram of the data using different symbols for the different categorical variables.

(b) Use dummy variable coding and regression to analyze the data.

(c) Perform appropriate hypothesis tests and construct appropriate confidence intervals using your choice of significance and confidence levels.

(d) Find the p value for each test that you perform.

11.2.1 For subjects undergoing stem cell transplants, dendritic cells (DCs) are antigen-presenting cells that are critical to the generation of immunologic tumor responses. Bolwell et al. (A-2) studied lymphoid DCs in 44 subjects who underwent autologous stem cell transplantation. The outcome variable is the concentration of DC2 cells as measured by flow cytometry. One of the independent variables is the age of the subject (years), and the second independent variable is the mobilization method. During chemotherapy, 11 subjects received granulocyte colony-stimulating factor (G-CSF) mobilizer (μg/kg/day) and 33 received etoposide (2 g/m^2). The mobilizer is a kind of blood progenitor cell that triggers the formation of the DC cells. The results were as follows:

G-CSF		Etoposide					
DC	**Age**	**DC**	**Age**	**DC**	**Age**	**DC**	**Age**
6.16	65	3.18	70	4.24	60	4.09	36
6.14	55	2.58	64	4.86	40	2.86	51
5.66	57	1.69	65	4.05	48	2.25	54
8.28	47	2.16	55	5.07	50	0.70	50
2.99	66	3.26	51	4.26	23	0.23	62
8.99	24	1.61	53	11.95	26	1.31	56
4.04	59	6.34	24	1.88	59	1.06	31
6.02	60	2.43	53	6.10	24	3.14	48
10.14	66	2.86	37	0.64	52	1.87	69
27.25	63	7.74	65	2.21	54	8.21	62
8.86	69	11.33	19	6.26	43	1.44	60

Source: Data provided courtesy of Lisa Rybicki, M.S.

11.2.2 According to Pandey et al. (A-3) carcinoma of the gallbladder is not infrequent. One of the primary risk factors for gallbladder cancer is cholelithiasis, the asymptomatic presence of stones in the gallbladder. The researchers performed a case-control study of 50 subjects with gallbladder cancer and 50 subjects with cholelithiasis. Of interest was the concentration of lipid peroxidation products in gallbladder bile, a condition that may give rise to gallbladder cancer. The lipid peroxidation product melonaldehyde (MDA, μg/mg) was used to measure lipid peroxidation. One of the independent variables considered was the cytochrome P-450 concentration (CYTO, nmol/mg). Researchers used disease status (gallbladder cancer vs. cholelithiasis) and cytochrome P-450 concentration to predict MDA. The following data were collected.

Cholelithiasis				Gallbladder Cancer			
MDA	**CYTO**	**MDA**	**CYTO**	**MDA**	**CYTO**	**MDA**	**CYTO**
0.68	12.60	11.62	4.83	1.60	22.74	9.20	8.99
0.16	4.72	2.71	3.25	4.00	4.63	0.69	5.86
0.34	3.08	3.39	7.03	4.50	9.83	10.20	28.32
3.86	5.23	6.10	9.64	0.77	8.03	3.80	4.76
0.98	4.29	1.95	9.02	2.79	9.11	1.90	8.09
3.31	21.46	3.80	7.76	8.78	7.50	2.00	21.05
1.11	10.07	1.72	3.68	2.69	18.05	7.80	20.22
4.46	5.03	9.31	11.56	0.80	3.92	16.10	9.06
1.16	11.60	3.25	10.33	3.43	22.20	0.98	35.07

(*Continued*)

Cholelithiasis				Gallbladder Cancer			
MDA	CYTO	MDA	CYTO	MDA	CYTO	MDA	CYTO
1.27	9.00	0.62	5.72	2.73	11.68	2.85	29.50
1.38	6.13	2.46	4.01	1.41	19.10	3.50	45.06
3.83	6.06	7.63	6.09	6.08	36.70	4.80	8.99
0.16	6.45	4.60	4.53	5.44	48.30	1.89	48.15
0.56	4.78	12.21	19.01	4.25	4.47	2.90	10.12
1.95	34.76	1.03	9.62	1.76	8.83	0.87	17.98
0.08	15.53	1.25	7.59	8.39	5.49	4.25	37.18
2.17	12.23	2.13	12.33	2.82	3.48	1.43	19.09
0.00	0.93	0.98	5.26	5.03	7.98	6.75	6.05
1.35	3.81	1.53	5.69	7.30	27.04	4.30	17.05
3.22	6.39	3.91	7.72	4.97	16.02	0.59	7.79
1.69	14.15	2.25	7.61	1.11	6.14	5.30	6.78
4.90	5.67	1.67	4.32	13.27	13.31	1.80	16.03
1.33	8.49	5.23	17.79	7.73	10.03	3.50	5.07
0.64	2.27	2.79	15.51	3.69	17.23	4.98	16.60
5.21	12.35	1.43	12.43	9.26	9.29	6.98	19.89

Source: Data provided courtesy of Manoj Pandey, M.D.

11.2.3 The purpose of a study by Krantz et al. (A-4) was to investigate dose-related effects of methadone in subjects with *torsades de pointes*, a polymorphic ventricular tachycardia. In the study of 17 subjects, 10 were men (sex = 0) and seven were women (sex = 1). The outcome variable, is the QTc interval, a measure of arrhythmia risk. The other independent variable, in addition to sex, was methadone dose (mg/day). Measurements on these variables for the 17 subjects were as follows.

Sex	Dose (mg/day)	QTc (msec)
0	1000	600
0	550	625
0	97	560
1	90	585
1	85	590
1	126	500
0	300	700
0	110	570
1	65	540
1	650	785
1	600	765
1	660	611
1	270	600
1	680	625
0	540	650
0	600	635
1	330	522

Source: Data provided courtesy of Mori J. Krantz, M.D.

11.2.4 Refer to Exercise 9.7.2, which describes research by Reiss et al. (A-5), who collected samples from 90 patients and measured partial thromboplastin time (aPTT) using two different methods: the CoaguChek point-of-care assay and standard laboratory hospital assay. The subjects were also classified by their medication status: 30 receiving heparin alone, 30 receiving heparin with warfarin, and 30 receiving warfarin and enoxaparin. The data are as follows.

Heparin		Heparin and Warfarin		Warfarin and Enoxaparin	
CoaguChek aPTT	**Hospital aPTT**	**CoaguChek aPTT**	**Hospital aPTT**	**CoaguChek aPTT**	**Hospital aPTT**
49.3	71.4	18.0	77.0	56.5	46.5
57.9	86.4	31.2	62.2	50.7	34.9
59.0	75.6	58.7	53.2	37.3	28.0
77.3	54.5	75.2	53.0	64.8	52.3
42.3	57.7	18.0	45.7	41.2	37.5
44.3	59.5	82.6	81.1	90.1	47.1
90.0	77.2	29.6	40.9	23.1	27.1
55.4	63.3	82.9	75.4	53.2	40.6
20.3	27.6	58.7	55.7	27.3	37.8
28.7	52.6	64.8	54.0	67.5	50.4
64.3	101.6	37.9	79.4	33.6	34.2
90.4	89.4	81.2	62.5	45.1	34.8
64.3	66.2	18.0	36.5	56.2	44.2
89.8	69.8	38.8	32.8	26.0	28.2
74.7	91.3	95.4	68.9	67.8	46.3
150.0	118.8	53.7	71.3	40.7	41.0
32.4	30.9	128.3	111.1	36.2	35.7
20.9	65.2	60.5	80.5	60.8	47.2
89.5	77.9	150.0	150.0	30.2	39.7
44.7	91.5	38.5	46.5	18.0	31.3
61.0	90.5	58.9	89.1	55.6	53.0
36.4	33.6	112.8	66.7	18.0	27.4
52.9	88.0	26.7	29.5	18.0	35.7
57.5	69.9	49.7	47.8	78.3	62.0
39.1	41.0	85.6	63.3	75.3	36.7
74.8	81.7	68.8	43.5	73.2	85.3
32.5	33.3	18.0	54.0	42.0	38.3
125.7	142.9	92.6	100.5	49.3	39.8
77.1	98.2	46.2	52.4	22.8	42.3
143.8	108.3	60.5	93.7	35.8	36.0

Source: Data provided courtesy of Curtis E. Haas, Pharm.D.

Use the multiple regression to predict the hospital aPTT from the CoaguCheck aPTT level as well as the medication received. Is knowledge of medication useful in the prediction? Let $\alpha = .05$ for all tests.

11.3 VARIABLE SELECTION PROCEDURES

Health sciences researchers contemplating the use of multiple regression analysis to investigate a question usually find that they have a large number of variables from which to select the independent variables to be employed as predictors of the dependent variable. Such investigators will want to include in their model as many variables as possible in order to maximize the model's predictive ability. The investigator must realize, however, that adding another independent variable to a set of independent variables always increases the coefficient of determination R^2. Therefore, independent variables should not be added to the model indiscriminately, but only for good reason. In most situations, for example, some potential predictor variables are more expensive than others in terms of data-collection costs. The cost-conscious investigator, therefore, will not want to include an expensive variable in a model unless there is evidence that it makes a worthwhile contribution to the predictive ability of the model.

The investigator who wishes to use multiple regression analysis most effectively must be able to employ some strategy for making intelligent selections from among those potential predictor variables that are available. Many such strategies are in current use, and each has its proponents. The strategies vary in terms of complexity and the tedium involved in their employment. Unfortunately, the strategies do not always lead to the same solution when applied to the same problem.

Stepwise Regression Perhaps the most widely used strategy for selecting independent variables for a multiple regression model is the stepwise procedure. The procedure consists of a series of steps. At each step of the procedure each variable then in the model is evaluated to see if, according to specified criteria, it should remain in the model.

Suppose, for example, that we wish to perform stepwise regression for a model containing k predictor variables. The criterion measure is computed for each variable. Of all the variables that do not satisfy the criterion for inclusion in the model, the one that least satisfies the criterion is removed from the model. If a variable is removed in this step, the regression equation for the smaller model is calculated and the criterion measure is computed for each variable now in the model. If any of these variables fail to satisfy the criterion for inclusion in the model, the one that least satisfies the criterion is removed. If a variable is removed at this step, the variable that was removed in the first step is reentered into the model, and the evaluation procedure is continued. This process continues until no more variables can be entered or removed.

The nature of the stepwise procedure is such that, although a variable may be deleted from the model in one step, it is evaluated for possible reentry into the model in subsequent steps.

MINITAB's STEPWISE procedure, for example, uses the associated F statistic as the evaluative criterion for deciding whether a variable should be deleted or added to the model. Unless otherwise specified, the cutoff value is $F = 4$. The printout of the STEPWISE results contains t statistics (the square root of F) rather than F statistics. At each step MINITAB calculates an F statistic for each variable then in the model. If the F statistic for any of these variables is less than the specified cutoff value (4 if some other value is not specified), the variable with the smallest F is removed from the model. The regression equation is refitted for the reduced model, the results are printed, and the

procedure goes to the next step. If no variable can be removed, the procedure tries to add a variable. An F statistic is calculated for each variable not then in the model. Of these variables, the one with the largest associated F statistic is added, provided its F statistic is larger than the specified cutoff value (4 if some other value is not specified). The regression equation is refitted for the new model, the results are printed, and the procedure goes on to the next step. The procedure stops when no variable can be added or deleted.

The following example illustrates the use of the stepwise procedure for selecting variables for a multiple regression model.

EXAMPLE. 11.3.1

A nursing director would like to use nurses' personal characteristics to develop a regression model for predicting the job performance (JOBPER). The following variables are available from which to choose the independent variables to include in the model:

$X_1 =$ assertiveness (ASRV)
$X_2 =$ enthusiasm (ENTH)
$X_3 =$ ambition (AMB)
$X_4 =$ communication skills (COMM)
$X_5 =$ problem-solving skills (PROB)
$X_6 =$ initiative (INIT)

We wish to use the stepwise procedure for selecting independent variables from those available in the table to construct a multiple regression model for predicting job performance.

Solution: Table 11.3.1 shows the measurements taken on the dependent variable, JOBPER, and each of the six independent variables for a sample of 30 nurses.

TABLE 11.3.1 Measurements on Seven Variables for Examples 11.3.1

Y	X_1	X_2	X_3	X_4	X_5	X_6
45	74	29	40	66	93	47
65	65	50	64	68	74	49
73	71	67	79	81	87	33
63	64	44	57	59	85	37
83	79	55	76	76	84	33
45	56	48	54	59	50	42
60	68	41	66	71	69	37
73	76	49	65	75	67	43
74	83	71	77	76	84	33

(*Continued*)

Y	X_1	X_2	X_3	X_4	X_5	X_6
69	62	44	57	67	81	43
66	54	52	67	63	68	36
69	61	46	66	84	75	43
71	63	56	67	60	64	35
70	84	82	68	84	78	37
79	78	53	82	84	78	39
83	65	49	82	65	55	38
75	86	63	79	84	80	41
67	61	64	75	60	81	45
67	71	45	67	80	86	48
52	59	67	64	69	79	54
52	71	32	44	48	65	43
66	62	51	72	71	81	43
55	67	51	60	68	81	39
42	65	41	45	55	58	51
65	55	41	58	71	76	35
68	78	65	73	93	77	42
80	76	57	84	85	79	35
50	58	43	55	56	84	40
87	86	70	81	82	75	30
84	83	38	83	69	79	41

We use MINITAB to obtain a useful model by the stepwise procedure. Observations on the dependent variable job performance (JOBPER) and the six candidate independent variables are stored in MINITAB Columns 1 through 7, respectively. Figure 11.3.1 shows the appropriate MINITAB procedure and the printout of the results.

To obtain the results in Figure 11.3.1, the values of F to enter and F to remove both were set automatically at 4. In step 1 there are no variables to be considered for deletion from the model. The variable AMB (Column 4) has the largest associated F statistic, which is $F = (9.74)^2 = 94.8676$. Since 94.8676 is greater than 4, AMB is added to the model. In step 2 the variable INIT (Column 7) qualifies for addition to the model since its associated F of $(-2.2)^2 = 4.84$ is greater than 4 and it is the variable with the largest associated F statistic. It is added to the model. After step 2 no other variable could be added or deleted, and the procedure stopped. We see, then, that the model chosen by the stepwise procedure is a two-independent-variable model with AMB and INIT as the independent variables. The estimated regression equation is

$$\hat{y} = 31.96 + .787x_3 - .45x_6$$

■

Dialog box:

Stat ➤ Regression ➤ Stepwise

Type *C1* in **Response** and *C2–C7* in **Predictors.**

Session command:

```
MTB > Stepwise  C1  C2-C7;
SUBC>   FEnter  4.0;
SUBC>   FRemove  4.0.
```

Stepwise Regression

```
F-to-Enter:    4.00  F-to-Remove:    4.00

Response  is  C1  on  6  predictors,  with  N = 30

Step        1         2
Constant  7.226    31.955

C4         0.888     0.787
T-Ratio    9.74      8.13

C7                  -0.45
T-Ratio             -2.20

S          5.90      5.53
R-Sq      77.21     80.68
```

FIGURE 11.3.1 MINITAB stepwise procedure and output for the data of Table 11.3.1.

To change the criterion for allowing a variable to enter the model from 4 to some other value K, click on Options, then type the desired value of K in the Enter box. The new criterion F statistic, then, is K rather than 4. To change the criterion for deleting a variable from the model from 4 to some other value K, click on Options, then type the desired value of K in the Remove box. We must choose K to enter to be greater than or equal to K to remove.

Though the stepwise selection procedure is a common technique employed by researchers, other methods are available. Following is a brief discussion of two such tools. The final model obtained by each of these procedures is the same model that was found by using the stepwise procedure in Example 11.3.1.

Forward Selection This strategy is closely related to the stepwise regression procedure. This method builds a model using correlations. Variables are retained that meet the criteria for inclusion, as in stepwise selection. The first variable entered into the model is the one with the highest correlation with the dependent variable. If this variable meets the inclusion criterion, it is retained. The next variable to be considered for inclusion is the one with the highest partial correlation with the dependent variable. If it meets the inclusion criteria, it is retained. This procedure continues until all of the independent variables have

been considered. The final model contains all of the independent variables that meet the inclusion criteria.

Backward Elimination This model-building procedure begins with all of the variables in the model. This strategy also builds a model using correlations and a predetermined inclusion criterion based on the F statistic. The first variable considered for removal from the model is the one with the smallest partial correlation coefficient. If this variable does not meet the criterion for inclusion, it is eliminated from the model. The next variable to be considered for elimination is the one with the next lowest partial correlation. It will be eliminated if it fails to meet the criterion for inclusion. This procedure continues until all variables have been considered for elimination. The final model contains all of the independent variables that meet the inclusion criteria.

EXERCISES

11.3.1 Refer to the data of Exercise 10.3.2 reported by Son et al. (A-6), who studied family caregiving in Korea of older adults with dementia. The outcome variable, caregiver burden (BURDEN), was measured by the Korean Burden Inventory (KBI) where scores ranged from 28 to 140 with higher scores indicating higher burden. Perform a stepwise regression analysis on the following independent variables reported by the researchers:

CGAGE: caregiver age (years)

CGINCOME: caregiver income (Won-Korean currency)

CGDUR: caregiver-duration of caregiving (month)

ADL: total activities of daily living where low scores indicate the elderly perform activities independently.

MEM: memory and behavioral problems with higher scores indicating more problems.

COG: cognitive impairment with lower scores indicating a greater degree of cognitive impairment.

SOCIALSU: total score of perceived social support (25–175, higher values indicating more support). The reported data are as follows.

CGAGE	CGINCOME	CGDUR	ADL	MEM	COG	SOCIALSU	BURDEN
41	200	12	39	4	18	119	28
30	120	36	52	33	9	131	68
41	300	60	89	17	3	141	59
35	350	2	57	31	7	150	91
37	600	48	28	35	19	142	70
42	90	4	34	3	25	148	38
49	300	26	42	16	17	172	46
39	500	16	52	6	26	147	57
49	309	30	88	41	13	98	89
40	250	60	90	24	3	147	48

(*Continued*)

CGAGE	CGINCOME	CGDUR	ADL	MEM	COG	SOCIALSU	BURDEN
40	300	36	38	22	13	146	74
70	60	10	83	41	11	97	78
49	450	24	30	9	24	139	43
55	300	18	45	33	14	127	76
27	309	30	47	36	18	132	72
39	250	10	90	17	0	142	61
39	260	12	63	14	16	131	63
44	250	32	34	35	22	141	77
33	200	48	76	33	23	106	85
42	200	12	26	13	18	144	31
52	200	24	68	34	26	119	79
48	300	36	85	28	10	122	92
53	300	12	22	12	16	110	76
40	300	11	82	57	3	121	91
35	200	8	80	51	3	142	78
47	150	60	80	20	18	101	103
33	180	19	81	20	1	117	99
41	200	48	30	7	17	129	73
43	300	36	27	27	27	142	88
25	309	24	72	9	0	137	64
35	250	12	46	15	22	148	52
35	200	6	63	52	13	135	71
45	200	7	45	26	18	144	41
36	300	24	77	57	0	128	85
52	600	60	42	10	19	148	52
41	230	6	60	34	11	141	68
40	200	36	33	14	14	151	57
45	400	96	49	30	15	124	84
48	75	6	89	64	0	105	91
50	200	30	72	31	3	117	83
31	250	30	45	24	19	111	73
33	300	2	73	13	3	146	57
30	200	30	58	16	15	99	69
36	250	6	33	17	21	115	81
45	500	12	34	13	18	119	71
32	300	60	90	42	6	134	91
55	200	24	48	7	23	165	48
50	309	20	47	17	18	101	94
37	250	30	32	13	15	148	57
40	1000	21	63	32	15	132	49
40	300	12	76	50	5	120	88
49	300	18	79	44	11	129	54
37	309	18	48	57	9	133	73
47	250	38	90	33	6	121	87
41	200	60	55	11	20	117	47
33	1000	18	83	24	11	140	60
28	309	12	50	21	25	117	65

(*Continued*)

CGAGE	CGINCOME	CGDUR	ADL	MEM	COG	SOCIALSU	BURDEN
33	400	120	44	31	18	138	57
34	330	18	79	30	20	163	85
40	200	18	24	5	22	157	28
54	200	12	40	20	17	143	40
32	300	32	35	15	27	125	87
44	280	66	55	9	21	161	80
44	350	40	45	28	17	142	49
42	280	24	46	19	17	135	57
44	500	14	37	4	21	137	32
25	600	24	47	29	3	133	52
41	250	84	28	23	21	131	42
28	1000	30	61	8	7	144	49
24	200	12	35	31	26	136	63
65	450	120	68	65	6	169	89
50	200	12	80	29	10	127	67
40	309	12	43	8	13	110	43
47	1000	12	53	14	18	120	47
44	300	24	60	30	16	115	70
37	309	54	63	22	18	101	99
36	300	12	28	9	27	139	53
55	200	12	35	18	14	153	78
45	2000	12	37	33	17	111	112
45	400	14	82	25	13	131	52
23	200	36	88	16	0	139	68
42	1000	12	52	15	0	132	63
38	200	36	30	16	18	147	49
41	230	36	69	49	12	171	42
25	200	30	52	17	20	145	56
47	200	12	59	38	17	140	46
35	100	12	53	22	21	139	72
59	150	60	65	56	2	133	95
49	300	60	90	12	0	145	57
51	200	48	88	42	6	122	88
54	250	6	66	12	23	133	81
53	30	24	60	21	7	107	104
49	100	36	48	14	13	118	88
44	300	48	82	41	13	95	115
36	200	18	88	24	14	100	66
64	200	48	63	49	5	125	92
51	120	2	79	34	3	116	97
43	200	66	71	38	17	124	69
54	150	96	66	48	13	132	112
29	309	19	81	66	1	152	88

Source: Data provided courtesy of Gwi-Ryung Son, R.N., Ph.D.

11.3.2 Machiel Naeije (A-7) identifies variables useful in predicting maximum mouth opening (MMO, millimeters) for 35 healthy volunteers. The variables examined were:

AGE:	years
DOWN_CON:	downward condylar translation, mm
FORW_CON:	forward condylar translation, mm
Gender:	0 = Female, 1 = Male
MAN_LENG:	mandibular length, mm
MAN_WIDT:	mandibular width, mm

Use the following reported measurements to perform a stepwise regression.

AGE	DOWN_CON	FORW_CON	GENDER	MAN_LENG	MAN_WIDT	MMO
21.00	4.39	14.18	1	100.86	121.00	52.34
26.00	1.39	20.23	0	93.08	118.29	51.90
30.00	2.42	13.45	1	98.43	130.56	52.80
28.00	−.18	19.66	1	102.95	125.34	50.29
21.00	4.10	22.71	1	108.24	125.19	57.79
20.00	4.49	13.94	0	98.34	113.84	49.41
21.00	2.07	19.35	0	95.57	115.41	53.28
19.00	−.77	25.65	1	98.86	118.30	59.71
24.00	7.88	18.51	1	98.32	119.20	53.32
18.00	6.06	21.72	0	92.70	111.21	48.53
22.00	9.37	23.21	0	88.89	119.07	51.59
21.00	3.77	23.02	1	104.06	127.34	58.52
20.00	1.10	19.59	0	98.18	111.24	62.93
22.00	2.52	16.64	0	91.01	113.81	57.62
24.00	5.99	17.38	1	96.98	114.94	65.64
22.00	5.28	22.57	0	97.86	111.58	52.85
22.00	1.25	20.89	0	96.89	115.16	64.43
22.00	6.02	20.38	1	98.35	122.52	57.25
19.00	1.59	21.63	0	90.65	118.71	50.82
26.00	6.05	10.59	0	92.99	119.10	40.48
22.00	−1.51	20.03	1	108.97	129.00	59.68
24.00	−.41	24.55	0	91.85	100.77	54.35
21.00	6.75	14.67	1	104.30	127.15	47.00
22.00	4.87	17.91	1	93.16	123.10	47.23
22.00	.64	17.60	1	94.18	113.86	41.19
29.00	7.18	15.19	0	89.56	110.56	42.76
25.00	6.57	17.25	1	105.85	140.03	51.88
20.00	1.51	18.01	0	89.29	121.70	42.77
27.00	4.64	19.36	0	92.58	128.01	52.34
26.00	3.58	16.57	1	98.64	129.00	50.45
23.00	6.64	12.47	0	83.70	130.98	43.18
25.00	7.61	18.52	0	88.46	124.97	41.99
22.00	5.39	11.66	1	94.93	129.99	39.45
31.00	5.47	12.85	1	96.81	132.97	38.91
23.00	2.60	19.29	0	93.13	121.03	49.10

Source: Data provided courtesy of Machiel Naeije, D.D.S.

11.3.3 One purpose of a study by Connor et al. (A-8) was to examine reactive aggression among children and adolescents referred to a residential treatment center. The researchers used the Proactive/Reactive Rating Scale, obtained by presenting three statements to clinicians who examined the subjects. The respondents answered, using a scale from 1 to 5, with 5 indicating that the statement almost always applied to the child. An example of a reactive aggression statement is, "When this child has been teased or threatened, he or she gets angry easily and strikes back." The reactive score was the average response to three statements of this type. With this variable as the outcome variable, researchers also examined the following: AGE (years), VERBALIQ (verbal IQ), STIM (stimulant use), AGEABUSE (age when first abused), CTQ (a measure of hyperactivity in which higher scores indicate higher hyperactivity), TOTALHOS (total hostility as measured by an evaluator, with higher numbers indicating higher hostility). Perform stepwise regression to find the variables most useful in predicting reactive aggression in the following sample of 68 subjects.

REACTIVE	AGE	VERBALIQ	STIM	AGEABUSE	CTQ	TOTALHOS
4.0	17	91	0	0	0	8
3.7	12	94	0	1	29	10
2.3	14	105	0	1	12	10
5.0	16	97	0	1	9	11
2.0	15	97	0	2	17	10
2.7	8	91	0	0	6	4
2.0	10	111	0	0	6	6
3.3	12	105	0	0	28	7
2.0	17	101	1	0	12	9
4.3	13	102	1	1	8	11
4.7	15	83	0	0	9	9
4.3	15	66	0	1	5	8
2.0	15	90	0	2	3	8
4.0	13	88	0	1	28	8
2.7	13	98	0	1	17	10
2.7	9	135	0	0	30	11
2.7	18	72	0	0	10	9
2.0	13	93	0	2	20	8
3.0	14	94	0	2	10	11
2.7	13	93	0	1	4	8
3.7	16	73	0	0	11	11
2.7	12	74	0	1	10	7
2.3	14	97	0	2	3	11
4.0	13	91	1	1	21	11
4.0	12	88	0	1	14	9
4.3	13	90	0	0	15	2
3.7	14	104	1	1	10	10
3.0	18	82	0	0	1	7
4.3	14	79	1	3	6	7
1.0	16	93	0	0	5	8
4.3	16	99	0	1	21	11

(*Continued*)

REACTIVE	AGE	VERBALIQ	STIM	AGEABUSE	CTQ	TOTALHOS
2.3	14	73	0	2	8	9
3.0	12	112	0	0	15	9
1.3	15	102	0	1	1	5
3.0	16	78	1	1	26	8
2.3	9	95	1	0	23	10
1.0	15	124	0	3	0	11
3.0	17	73	0	1	1	10
3.3	11	105	0	0	23	5
4.0	11	89	0	0	27	8
1.7	9	88	0	1	2	8
2.3	16	96	0	1	5	7
4.7	15	76	1	1	17	9
1.7	16	87	0	2	0	4
1.7	15	90	0	1	10	12
4.0	12	76	0	0	22	10
5.0	12	83	1	1	19	7
4.3	10	88	1	0	10	5
5.0	9	98	1	0	8	9
3.7	12	100	0	0	6	4
3.3	14	80	0	1	3	10
2.3	16	84	0	1	3	9
1.0	17	117	0	2	1	9
1.7	12	145	1	0	0	5
3.7	12	123	0	0	1	3
2.0	16	94	0	2	6	6
3.7	17	70	0	1	11	13
4.3	14	113	0	0	8	8
2.0	12	123	1	0	2	8
3.0	7	107	0	0	11	9
3.7	12	78	1	0	15	11
4.3	14	73	0	1	2	8
2.3	18	91	0	3	8	10
4.7	12	91	0	0	6	9
3.7	15	111	0	0	2	9
1.3	15	71	0	1	20	10
3.7	7	102	0	0	14	9
1.7	9	89	0	0	24	6

Source: Data provided courtesy of Daniel F. Connor, M.D. and Lang Lin.

11.4 LOGISTIC REGRESSION

Up to now our discussion of regression analysis has been limited to those situations in which the dependent variable is a continuous variable such as weight, blood pressure, or plasma levels of some hormone. Much research in the health sciences field is

motivated by a desire to describe, understand, and make use of the relationship between independent variables and a dependent (or outcome) variable that is discrete. Particularly plentiful are circumstances in which the outcome variable is dichotomous. A dichotomous variable, we recall, is a variable that can assume only one of two mutually exclusive values. These values are usually coded $Y = 1$ for a success and $Y = 0$ for a nonsuccess, or failure. Dichotomous variables include those whose two possible values are such categories as died–did not die; cured–not cured; disease occurred–disease did not occur; and smoker–nonsmoker. The health sciences professional who either engages in research or needs to understand the results of research conducted by others will find it advantageous to have, at least, a basic understanding of *logistic regression*, the type of regression analysis that is usually employed when the dependent variable is dichotomous. The purpose of the present discussion is to provide the reader with this level of understanding. We shall limit our presentation to the case in which there is only one independent variable that may be either continuous or dichotomous.

The Logistic Regression Model Recall that in Chapter 9 we referred to regression analysis involving only two variables as simple linear regression analysis. The simple linear regression model was expressed by the equation

$$y = \beta_0 + \beta_1 x + \epsilon \tag{11.4.1}$$

in which y is an arbitrary observed value of the continuous dependent variable. When the observed value of Y is $\mu_{y|x}$, the mean of a subpopulation of Y values for a given value of X, the quantity ϵ, the difference between the observed Y and the regression line (see Figure 9.2.1) is zero, and we may write Equation 11.4.1 as

$$\mu_{y|x} = \beta_0 + \beta_1 x \tag{11.4.2}$$

which may also be written as

$$E(y|x) = \beta_0 + \beta_1 x \tag{11.4.3}$$

Generally, the right-hand side of Equations (11.4.1) through (11.4.3) may assume any value between minus infinity and plus infinity.

Even though only two variables are involved, the simple linear regression model is not appropriate when Y is a dichotomous variable because the expected value (or mean) of Y is the probability that $Y = 1$ and, therefore, is limited to the range 0 through 1, inclusive. Equations (11.4.1) through (11.4.3), then, are incompatible with the reality of the situation.

If we let $p = P(Y = 1)$, then the ratio $p/(1 - p)$ can take on values between 0 and plus infinity. Furthermore, the natural logarithm (ln) of $p/(1 - p)$ can take on values

between minus infinity and plus infinity just as can the right-hand side of Equations 11.4.1 through (11.4.3). Therefore, we may write

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x \tag{11.4.4}$$

Equation 11.4.4 is called the *logistic regression model* and the transformation of $\mu_{y|x}$ (that is, p) to $\ln[p/(1-p)]$ is called the *logit transformation*. Equation 11.4.4 may also be written as

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{11.4.5}$$

in which exp is the inverse of the natural logarithm.

The logistic regression model is widely used in health sciences research. For example, the model is frequently used by epidemiologists as a model for the probability (interpreted as the risk) that an individual will acquire a disease during some specified time period during which he or she is exposed to a condition (called a *risk factor*) known to be or suspected of being associated with the disease.

Logistic Regression: Dichotomous Independent Variable The simplest situation in which logistic regression is applicable is one in which both the dependent and the independent variables are dichotomous. The values of the dependent (or outcome) variable usually indicate whether or not a subject acquired a disease or whether or not the subject died. The values of the independent variable indicate the status of the subject relative to the presence or absence of some risk factor. In the discussion that follows we assume that the dichotomies of the two variables are coded 0 and 1. When this is the case the variables may be cross-classified in a table, such as Table 11.4.1, that contains two rows and two columns. The cells of the table contain the frequencies of occurrence of all possible pairs of values of the two variables: (1, 1), (1, 0), (0, 1), and (0, 0).

An objective of the analysis of data that meet these criteria is a statistic known as the *odds ratio*. To understand the concept of the odds ratio, we must understand the term *odds*,

TABLE 11.4.1 Two Cross-Classified Dichotomous Variables Whose Values Are Coded 1 and 0

	Independent Variable (X)	
Dependent Variable (Y)	1	0
1	$n_{1,1}$	$n_{1,0}$
2	$n_{0,1}$	$n_{0,0}$

which is frequently used by those who place bets on the outcomes of sporting events or participate in other types of gambling activities. Using probability terminology, we may define odds as follows.

DEFINITION

The odds for success is the ratio of the probability of success to the probability of failure.

The odds ratio is a measure of how much greater (or less) the odds are for subjects possessing the risk factor to experience a particular outcome. This conclusion assumes that the outcome is a rare event. For example, when the outcome is the contracting of a disease, the interpretation of the odds ratio assumes that the disease is rare.

Suppose, for example, that the outcome variable is the acquisition or nonacquisition of skin cancer and the independent variable (or risk factor) is high levels of exposure to the sun. Analysis of such data collected on a sample of subjects might yield an odds ratio of 2, indicating that the odds of skin cancer are two times higher among subjects with high levels of exposure to the sun than among subjects without high levels of exposure.

Computer software packages that perform logistic regression frequently provide as part of their output estimates of β_0 and β_1 and the numerical value of the odds ratio. As it turns out the odds ratio is equal to $\exp(\beta_1)$.

EXAMPLE 11.4.1

LaMont et al. (A-9) tested for obstructive coronary artery disease (OCAD) among 113 men and 35 women who complained of chest pain or possible equivalent to their primary care physician. Table 11.4.2 shows the cross-classification of OCAD with gender. We wish to use logistic regression analysis to determine how much greater the odds are of finding OCAD among men than among women.

Solution: We may use the SAS® software package to analyze these data. The independent variable is gender and the dependent variable is status with respect to having obstructive coronary artery disease (OCAD). Use of the SAS® command PROC LOGIST yields. as part of the resulting output, the statistics shown in Figure 11.4.1.

TABLE 11.4.2 Cases of Obstructive Coronary Artery Disease (OCAD) Classified by Sex

Disease	Males	Females	Total
OCAD present	92	15	107
OCAD not present	21	20	41
Total	113	35	148

Source: Data provided courtesy of Matthew J. Budoff, M.D.

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.2877	0.3416	0.7090	0.3997
sex	1	1.7649	0.4185	17.7844	<.0001

FIGURE 11.4.1 Partial output from use of SAS® command PROC LOGISTIC with the data of Table 11.4.2.

We see that the estimate of α is -0.2877 and the estimate of β_1 is 1.7649. The estimated odds ratio, then, is $\widehat{OR} = \exp(1.7649) = 5.84$. Thus, we estimate that the odds of finding a case of obstructive coronary artery disease to be almost six times higher among men than women. ■

Logistic Regression: Continuous Independent Variable Now let us consider the situation in which we have a dichotomous dependent variable and a continuous independent variable. We shall assume that a computer is available to perform the calculations. Our discussion, consequently, will focus on an evaluation of the adequacy of the model as a representation of the data at hand, interpretation of key elements of the computer printout, and the use of the results to answer relevant questions about the relationship between the two variables.

EXAMPLE 11.4.2

According to Gallagher et al. (A-10), cardiac rehabilitation programs offer "information, support, and monitoring for return to activities, symptom management, and risk factor modification." The researchers conducted a study to identify among women factors that are associated with participation in such programs. The data in Table 11.4.3 are the ages of 185 women discharged from a hospital in Australia who met eligibility criteria involving discharge for myocardial infarction, artery bypass surgery, angioplasty, or stent. We wish to use these data to obtain information regarding the relationship between age (years) and participation in a cardiac rehabilitation program (ATT = 1, if participated, and ATT = 0, if not). We wish also to know if we may use the results of our analysis to predict the likelihood of participation by a woman if we know her age.

Solution: The independent variable is the continuous variable age (AGE), and the dependent or response variable is status with respect to attendance in a cardiac rehabilitation program. The dependent variable is a dichotomous variable that can assume one of two values: 0 = did not attend, and 1 = did

TABLE 11.4.3 Ages of Women Participating and Not Participating in a Cardiac Rehabilitation Program

Nonparticipating (ATT = 0)				Participating (ATT = 1)	
50	73	46	74	74	62
59	75	57	59	50	74
42	71	53	81	55	61
50	69	40	74	66	69
34	78	73	77	49	76
49	69	68	59	55	71
67	74	72	75	73	61
44	86	59	68	41	46
53	49	64	81	64	69
45	63	78	74	46	66
79	63	68	65	65	57
46	72	67	81	50	60
62	64	55	62	61	63
58	72	71	85	64	63
70	79	80	84	59	56
60	75	75	39	73	70
67	70	69	52	73	70
64	73	80	67	65	63
62	66	79	82	67	63
50	75	71	84	60	65
61	73	69	79	69	67
69	71	78	81	61	68
74	72	75	74	79	84
65	69	71	85	66	69
80	76	69	92	68	78
69	60	77	69	61	69
77	79	81	83	63	79
61	78	78	82	70	83
72	62	76	85	68	67
67	73	84	82	59	47
			80	64	57
					66

Source: Data provided courtesy of Robyn Gallagher, R.N., Ph.D.

attend. We use the SAS® software package to analyze the data. The SAS® command is PROC LOGISTIC, but if we wish to predict attendance in the cardiac program, we need to use the "descending" option with PROC LOGISTIC. (When you wish to predict the outcome labeled "1" of the dependent variable, use the "descending option" in SAS®. Consult

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.8744	0.9809	3.6518	0.0560
age	1	−0.0379	0.0146	6.7083	0.0096

FIGURE 11.4.2 Partial SAS® printout of the logistic regression analysis of the data in Table 11.4.3.

SAS®documentation for further details.) A partial printout of the analysis is shown in Figure 11.4.2.

The slope of our regression is −.0379, and the intercept is 1.8744. The regression equation, then, is

$$\hat{y}_i = 1.8744 - .0379x_i$$

where $\hat{y}_i = \ln[\hat{p}_i(1 - \hat{p}_i)]$ and $\hat{p}_i$ is the predicted probability of attending cardiac rehabilitation for a woman aged x_i.

Test of H_0 that $\beta_1 = 0$

We reach a conclusion about the adequacy of the logistic model by testing the null hypothesis that the slope of the regression line is zero. The test statistic is $z = \hat{\beta}_1 / s_{\hat{\beta}_1}$ where z is the standard normal statistic, $\hat{\beta}_1$ is the sample slope (−.0379), and $s_{\hat{\beta}_1}$ is its standard error (.0146) as shown in Figure 11.4.2. From these numbers we compute $z = -.0379/.0146 = -2.5959$, which has an associated two-sided p value of .0094. We conclude, therefore, that the logistic model is adequate. The square of z is chi-square with 1 degree of freedom, a statistic that is shown in Figure 11.4.2.

Using the Logistic Regression to Estimate p

We may use Equation 11.4.5 and the results of our analysis to estimate p, the probability that a woman of a given age (within the range of ages represented by the data) will attend a cardiac rehabilitation program. Suppose, for example, that we wish to estimate the probability that a woman who is 50 years of age will participate in a rehabilitation program. Substituting 50 and the results shown in Figure 11.4.2 into Equation 11.4.5 gives

$$\hat{p} = \frac{\exp[1.8744 - (.0379)(50)]}{1 + \exp[1.8744 - (.0379)(50)]} = .49485$$

SAS® calculates the estimated probabilities for the given values of X. We can see the estimated probabilities of attending cardiac rehabilitation programs for the age range of the subjects enrolled in the study in Figure 11.4.3. Since the slope was negative, we see a decreasing probability of attending a cardiac rehabilitation program for older women.

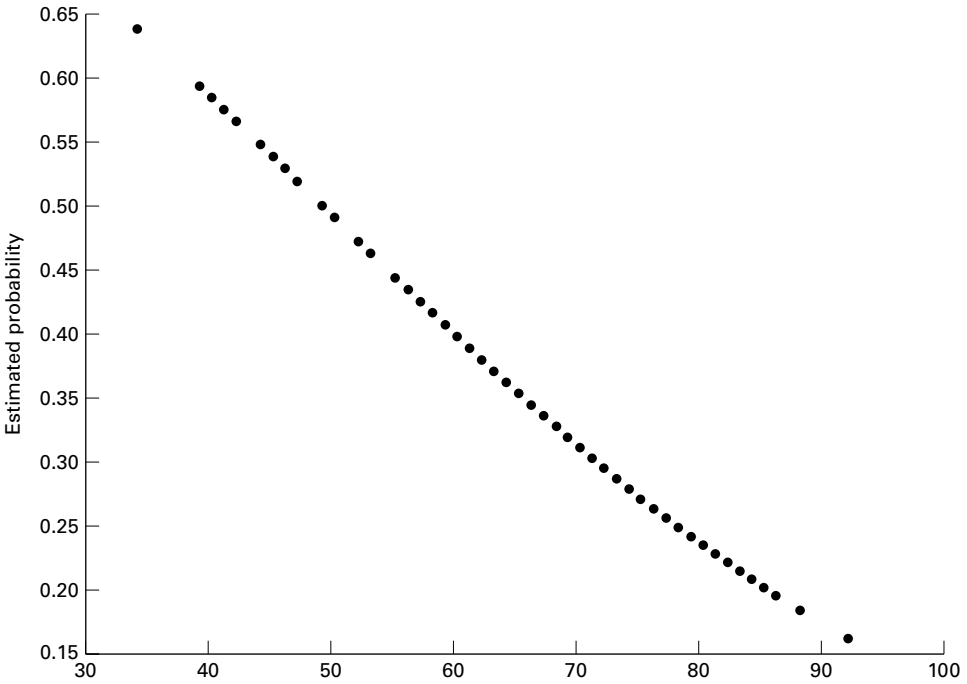


FIGURE 11.4.3 Estimated probabilities of attendance for ages within the study for Example 11.4.2. ■

Multiple Logistic Regression Practitioners often are interested in the relationships of several independent variables to a response variable. These independent variables may be either continuous or discrete or a combination of the two.

Multiple logistic models are constructed by expanding Equations (11.4.1) to (11.4.4). If we begin with Equation 11.4.4, multiple logistic regression can be represented as

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} \tag{11.4.6}$$

Using the logit transformation, we now have

$$p = \frac{\exp\left(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj}\right)} \tag{11.4.7}$$

EXAMPLE 11.4.3

Consider the data presented in Review Exercise 24. In this study by Fils-Aime et al. (A-21), data were gathered and classified with regard to alcohol use. Subjects were classified as having either early (< 25 years) or late (> 25 years) onset of excessive alcohol use.

Parameter	B	S.E.	Wald	Df	Sig.	Exp(B)
5-HIAA	−.013	.006	5.878	1	.015	.987
TRYPT	.000	.000	.000	1	.983	1.000
Constant	2.076	1.049	3.918	1	.048	7.970

FIGURE 11.4.4 SPSS output for the data in Example 11.4.3.

Levels of cerebrospinal fluid (CSF) tryptophan (TRYPT) and 5-hydroxyindoleacetic acid (5-HIAA) concentrations were also obtained.

Solution: The independent variables are the concentrations of TRYPT and 5-HIAA, and the dependent variable is the dichotomous response for onset of excessive alcohol use. We use SPSS software to analyze the data. The output is presented in Figure 11.4.4.

The equation can be written as

$$\hat{y}_i = 2.076 - .013x_{1j} + 0x_{2j}$$

Note that the coefficient for TRYPT is 0, and therefore it is not playing a role in the model.

Test of H_0 that $\beta_1 = 0$

Tests for significance of the regression coefficients can be obtained directly from Figure 11.4.4. Note that both the constant (intercept) and the 5-HIAA variables are significant in the model (both have p values, noted as "Sig." in the table, $<.05$); however, TRYPT is not significant and therefore need not be in the model, suggesting that it is not useful for identifying those study participants with early or late alcoholism onset.

As above, probabilities can be easily obtained by using Equation 11.4.7 and substituting the values obtained from the analysis. ■

Assessing Goodness of Fit A natural question that arises when doing logistic regression is: "How good is my model?" In classical linear regression we discussed measures such as R^2 for determining how much variation is explained by the model, with values of R^2 approaching 1 as a good indicator of model adequacy based on the predictors chosen to model the outcome. Given the nature of the response variable in logistic regression, a coefficient of determination does not provide the same information as it does in linear regression. This is because in logistic regression values of the parameters are not derived to minimize sums of squares, but rather are iterative estimates; hence, there is no equivalent measure of R^2 in logistic regression. Below, we provide an explanation of some commonly used approaches to evaluate logistic regression models, and follow these explanations with two illustrative examples.

Many authors have attempted to develop what are known as "pseudo-R^2" values that range from 0 to 1, with higher values indicating better fit. In general, these measures are

based on comparisons of a derived model with a model that contains only an intercept. In other words, they are comparative measures designed to indicate "how much better" a model with predictor variables is when compared to a model with no predictors. Two commonly used pseudo-R^2 statistics were developed by Cox and Snell (4) and Nagelkerke (5). These are often provided in standard outputs of statistical software. The value of these measures is the fact that they may be useful for comparing models with different predictor variables, but provide little relative use for examining a single model. Both of these approaches are based on the idea of using a measure of fit known as the log-likelihood statistic. The log-likelihood for the intercept-only model is used to represent the total sum of squares, while the log-likelihood for the model with predictor variables is used to represent the error sum of squares. Interested readers may find an explanation of the log-likelihood statistic in Hosmer and Lemeshow (2).

Another intuitive approach is to consider a classification table. Using this method, one develops a contingency table that provides frequency counts of the number of data points that were observed to be either 0 or 1 in the raw data, along with whether the raw data were classified as 0 or 1 based on the predictive equation. One can then estimate how many of the data points were correctly classified. As a general rule-of-thumb, correctly classifying 70 percent or greater is considered evidence of a satisfactory model from a statistical viewpoint. However, the model may not provide great enough predictive ability to be useful in a practice sense. A problem does arise, however, in that reclassifying the same data used to build a model with the model itself may bias the results. There are two practical ways to deal with this issue. First, one may use part of the data set to construct the model and the other part of the data set to develop a classification table. This strategy, of course, requires a sample large enough to accommodate adequately the needs of both procedures. A second approach is to construct a model using the data in hand and then collect additional data to test the adequacy of the model using a classification table. This strategy, too, has its shortcomings, as the collection of additional data can be both time-consuming and expensive.

A third approach that also has intuitive visual appeal is to develop a plot that shows the frequency of observations against their predicted probability. In this type of plot, one would hope to see a complete separation of 0 and 1 values. When there is misclassification of the outcome variable, this type of plot provides a means of determining where the misclassification occurred, and how frequently observations were misclassified.

Finally, in a commonly used approach known as the Hosmer and Lemeshow test, one develops a table of observed and expected frequencies and uses a chi-square test to determine if there is a significant deviation between the observed and expected frequencies. For the interested reader, we suggest the text by Hosmer and Lemeshow (2).

EXAMPLE 11.4.4

Consider the logistic regression model that was constructed from the cardiac rehabilitation program data in Example 11.4.2.

Figure 11.4.5 shows standard SPSS output for this logistic regression model. In this figure, we see that both the Cox and Snell and the Nagelkerke pseudo-R^2 values are provided. Since they are both > 0, the model with the predictor provides more information than the intercept-only model. One can readily see that only 63% of the data were correctly

Model Summary

Cox & Snell R Square	Nagelkerke R Square
.037	.051

Classification Table[a]

Observed		Predicted		
		att		Percentage
		0	1	Correct
att	0	111	10	91.7
	1	58	5	7.9
Overall Percentage				63.0

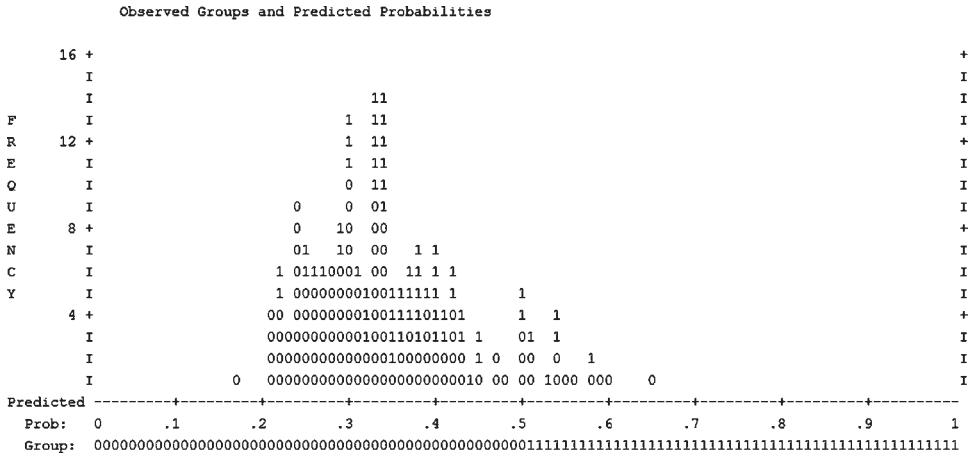


FIGURE 11.4.5 Partial SPSS output for the logistic regression analysis of the data in Example 11.4.2.

reclassified, with those participating in the rehabilitation program much more poorly classified than those who did not attend the program. The frequency distribution shows the large number of ATT = 1 subjects who were misclassified as ATT = 0 based on the model. ■

EXAMPLE 11.4.5

Consider the logistic regression model that was constructed from the cardiac rehabilitation program data in Example 11.4.3.

Figure 11.4.6 shows standard SPSS output for this logistic regression model. In this figure, we see that both the Cox and Snell and the Nagelkerke pseudo-R^2 values are provided, and since they are both > 0, the model with the predictors provides more information than the intercept-only model. One can readily see that only 69% of the data were correctly reclassified, with the model reclassifying those with onset of excessive alcohol use at a much

Model Summary

Cox & Snell R Square	Nagelkerke R Square
0.49	.069

Classification Table[a]

Observed		Predicted		
		Onset		Percentage Correct
		0	1	
onset	0	2	37	5.1
	1	3	87	96.7
Overall Percentage				69.0

a. The cut value is. 500

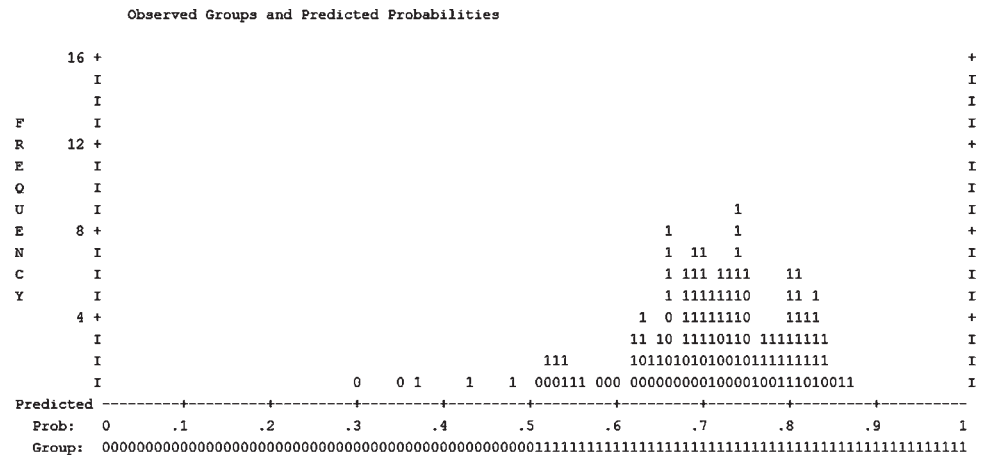


FIGURE 11.4.6 Partial SPSS output for the logistic regression analysis of the data in Example 11.4.3.

higher rate than those without such onset. The frequency distribution shows the large number of those without onset of excessive alcohol use predicted by the model to develop early onset of alcoholism. ■

Polytomous Logistic Regression Thus far we have limited our discussion to situations in which there is a dichotomous response variable (e.g., successful or unsuccessful). Often, we have a situation in which multiple categories make up the response. We may, for example, have subjects that are classified as positive, negative, and undetermined for a given disease (a standard polytomous response). There may also be times when we have a response variable that is ordered. We may, for example, classify our subjects by BMI as underweight, ideal weight, overweight, or obese (an ordinal polytomous response). The modeling process is slightly more complex and requires the use of a computer program. For those interested in exploring these valuable methods further, we recommend the book by Hosmer and Lemeshow (2).

Further Reading We have discussed only the basic concepts and applications of logistic regression. The technique has much wider application. Stepwise regression analysis may be used with logistic regression. There are also techniques available for constructing confidence intervals for odds ratios. The reader who wishes to learn more about logistic regression may consult the books by Hosmer and Lemeshow (2) and Kleinbaum (3).

EXERCISES

11.4.1 In a study of violent victimization of women and men, Porcerelli et al. (A-11) collected information from 679 women and 345 men ages 18 to 64 years at several family-practice centers in the metropolitan Detroit area. Patients filled out a health history questionnaire that included a question about victimization. The following table shows the sample subjects cross-classified by gender and whether the subject self-identified as being "hit, kicked, punched, or otherwise hurt by someone within the past year." Subjects answering yes to that question are classified "violently victimized." Use logistic regression analysis to find the regression coefficients and the estimate of the odds ratio. Write an interpretation of your results.

Victimization	Women	Men	Total
No victimization	611	308	919
Violently victimized	68	37	105
Total	679	345	1024

Source: John H. Porcerelli, Rosemary Cogan, Patricia P. West, Edward A. Rose, Dawn Lambrecht, Karen E. Wilson, Richard K. Severson, and Dunia Karana, "Violent Victimization of Women and Men: Physical and Psychiatric Symptoms," *Journal of the American Board of Family Practice, 16* (2003), 32–39.

11.4.2 Refer to the research of Gallagher et al. (A-10) discussed in Example 11.4.2. Another covariate of interest was a score using the Hospital Anxiety and Depression Index. A higher value for this score indicates a higher level of anxiety and depression. Use the following data to predict whether a woman in the study participated in a cardiac rehabilitation program.

Hospital Anxiety and Depression Index Scores for Nonparticipating Women				Hospital Anxiety and Depression Index Scores for Participating Women	
17	14	19	16	23	25
7	21	6	9	3	6
19	13	8	22	24	29
16	15	13	17	13	22
23	21	4	14	26	11
27	12	15	14	19	12
23	9	23	5	25	20
18	29	19	5	15	18
21	4	14	14	22	24
27	18	19	20	13	18

(*Continued*)

Hospital Anxiety and Depression Index Scores for Nonparticipating Women					Hospital Anxiety and Depression Index Scores for Participating Women
14	22	17	21	21	8
25	5	13	17	15	10
19	27	14	17	12	17
23	16	14	10	25	14
6	11	17	13	29	21
8	19	26	10	17	25
15	23	15	20	21	25
30	22	19	3	8	16
18	25	16	18	19	23
10	11	10	9	16	19
29	20	15	10	24	24
8	11	22	5	17	11
12	28	8	15	26	17
27	12	15	13	12	19
12	19	20	16	19	20
9	18	12		13	17
16	13	2		23	31
6	12	6		11	0
22	7	14		17	18
10	12	19		29	18
9	14	14		6	15
11	13	19		20	

Source: Data provided courtesy of Robyn Gallagher, R.N., Ph.D.

11.5 SUMMARY

This chapter is included for the benefit of those who wish to extend their understanding of regression analysis and their ability to apply techniques to models that are more complex than those covered in Chapters 9 and 10. In this chapter we present some additional topics from regression analysis. We discuss the analysis that is appropriate when one or more of the independent variables is dichotomous. In this discussion the concept of dummy variable coding is presented. A second topic that we discuss is how to select the most useful independent variables when we have a long list of potential candidates. The technique we illustrate for the purpose is stepwise regression analysis. Finally, we present the basic concepts and procedures that are involved in logistic regression analysis. We cover two situations: the case in which the independent variable is dichotomous, and the case in which the independent variable is continuous.

Since the calculations involved in obtaining useful results from data that are appropriate for analysis by means of the techniques presented in this chapter are complicated and time-consuming when attempted by hand, it is recommended that a computer be used to work the exercises.

SUMMARY OF FORMULAS FOR CHAPTER 11

Formula Number	Name	Formula
11.4.1–11.4.3	Representations of the simple linear regression model	$y = \beta_0 + \beta_1 x + \epsilon$ $\mu_{y\|x} = \beta_0 + \beta_1 x$ $E_{(y\|x)} = \beta_0 + \beta_1 x$
11.4.4	Simple logistic regression model	$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x$
11.4.5	Alternative representation of the simple logistic regression model	$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$
11.4.6	Alternative representation of the multiple logistic regression model	$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj}$
11.4.7	Alternative representation of the multiple logistic regression model	$p = \frac{\exp(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj})}{1 + \exp(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj})}$
Symbol Key	• β_0 = regression intercept • β_i = regression coefficient • ϵ = regression model error term • $E_{(y\|x)}$ = expected value of y at x • $\ln\left[\frac{p}{1-p}\right]$ = log it transformation • $\mu_{y\|x}$ = mean of y at x • x_i = value of independent variable at i	

REVIEW QUESTIONS AND EXERCISES

1. What is a qualitative variable?

2. What is a dummy variable?

3. Explain and illustrate the technique of dummy variable coding.

4. Why is a knowledge of variable selection techniques important to the health sciences researcher?

5. What is stepwise regression?

6. Explain the basic concept involved in stepwise regression.

7. When is logistic regression used?

8. Write out and explain the components of the logistic regression model.

9. Define the word *odds*.

10. What is an odds ratio?

11. Give an example in your field in which logistic regression analysis would be appropriate when the independent variable is dichotomous.

12. Give an example in your field in which logistic regression analysis would be appropriate when the independent variable is continuous.

13. Find a published article in the health sciences field in which each of the following techniques is employed:

(a) Dummy variable coding

(b) Stepwise regression

(c) Logistic regression

Write a report on the article in which you identify the variables involved, the reason for the choice of the technique, and the conclusions that the authors reach on the basis of their analysis.

14. In Example 10.3.1, we saw that the purpose of a study by Jansen and Keller (A-12) was to predict the capacity to direct attention (CDA) in elderly subjects. The study collected information on 71 community-dwelling older women with normal mental status. Higher CDA scores indicate better attentional functioning. In addition to the variables age and education level, the researchers performed stepwise regression with two additional variables: IADL, a measure of activities of daily living (higher values indicate greater number of daily activities), and ADS, a measure of attentional demands (higher values indicate more attentional demands). Perform stepwise regression with the data in the following table and report your final model, p values, and conclusions.

CDA	Age	Edyrs	IADL	ADS	CDA	Age	Edyrs	IADL	ADS
4.57	72	20	28	27	3.17	79	12	28	18
−3.04	68	12	27	96	−1.19	87	12	21	61
1.39	65	13	24	97	0.99	71	14	28	55
−3.55	85	14	27	48	−2.94	81	16	27	124
−2.56	84	13	28	50	−2.21	66	16	28	42
−4.66	90	15	27	47	−0.75	81	16	28	64
−2.70	79	12	28	71	5.07	80	13	28	26
0.30	74	10	24	48	−5.86	82	12	28	84
−4.46	69	12	28	67	5.00	65	13	28	43
−6.29	87	15	21	81	0.63	73	16	26	70
−4.43	84	12	27	44	2.62	85	16	28	20
0.18	79	12	28	39	1.77	83	17	23	80
−1.37	71	12	28	124	−3.79	83	8	27	21
3.26	76	14	29	43	1.44	76	20	28	26
−1.12	73	14	29	30	−5.77	77	12	28	53
−0.77	86	12	26	44	−5.77	83	12	22	69
3.73	69	17	28	47	−4.62	79	14	27	82
−5.92	66	11	28	49	−2.03	69	12	28	77
5.74	65	16	28	48	−2.22	66	14	28	38
2.83	71	14	28	46	0.80	75	12	28	28
−2.40	80	18	28	25	−0.75	77	16	27	85
−0.29	81	11	28	27	−4.60	78	12	22	82
4.44	66	14	29	54	2.68	83	20	28	34
3.35	76	17	29	26	−3.69	85	10	20	72
−3.13	70	12	25	100	4.85	76	18	28	24
−2.14	76	12	27	38	−0.08	75	14	29	49
9.61	67	12	26	84	0.63	70	16	28	29
7.57	72	20	29	44	5.92	79	16	27	83

(*Continued*)

CDA	Age	Edyrs	IADL	ADS	CDA	Age	Edyrs	IADL	ADS
2.21	68	18	28	52	3.63	75	18	28	32
−2.30	102	12	26	18	−7.07	94	8	24	80
1.73	67	12	27	80	6.39	76	18	28	41
6.03	66	14	28	54	−0.08	84	18	27	75
−0.02	75	18	26	67	1.07	79	17	27	21
−7.65	91	13	21	101	5.31	78	16	28	18
4.17	74	15	28	90	0.30	79	12	28	38

Source: Data provided courtesy of Debra Jansen, Ph.D., R.N.

15. In the following table are the cardiac output (L/min) and oxygen consumption (V_{O_2}) values for a sample of adults (A) and children (C), who participated in a study designed to investigate the relationship among these variables. Measurements were taken both at rest and during exercise. Treat cardiac output as the dependent variable and use dummy variable coding and analyze the data by regression techniques. Explain the results. Plot the original data and the fitted regression equations.

Cardiac Output (L/min)	V_{O_2} (L/min)	Age Group	Cardiac Output (L/min)	V_{O_2} (L/min)	Age Group
4.0	.21	A	4.0	.25	C
7.5	.91	C	6.1	.22	A
3.0	.22	C	6.2	.61	C
8.9	.60	A	4.9	.45	C
5.1	.59	C	14.0	1.55	A
5.8	.50	A	12.9	1.11	A
9.1	.99	A	11.3	1.45	A
3.5	.23	C	5.7	.50	C
7.2	.51	A	15.0	1.61	A
5.1	.48	C	7.1	.83	C
6.0	.74	C	8.0	.61	A
5.7	.70	C	8.1	.82	A
14.2	1.60	A	9.0	1.15	C
4.1	.30	C	6.1	.39	A

16. A simple random sample of normal subjects between the ages of 6 and 18 yielded the data on total body potassium (mEq) and total body water (liters) shown in the following table. Let total potassium be the dependent variable and use dummy variable coding to quantify the qualitative variable. Analyze the data using regression techniques. Explain the results. Plot the original data and the fitted regression equations.

Total Body Potassium	Total Body Water	Sex	Total Body Potassium	Total Body Water	Sex
795	13	M	950	12	F
1590	16	F	2400	26	M
1250	15	M	1600	24	F
1680	21	M	2400	30	M

(*Continued*)

Total Body Potassium	Total Body Water	Sex	Total Body Potassium	Total Body Water	Sex
800	10	F	1695	26	F
2100	26	M	1510	21	F
1700	15	F	2000	27	F
1260	16	M	3200	33	M
1370	18	F	1050	14	F
1000	11	F	2600	31	M
1100	14	M	3000	37	M
1500	20	F	1900	25	F
1450	19	M	2200	30	F
1100	14	M			

17. The data shown in the following table were collected as part of a study in which the subjects were preterm infants with low birth weights born in three different hospitals. Use dummy variable coding and multiple regression techniques to analyze these data. May we conclude that the three sample hospital populations differ with respect to mean birth weight when gestational age is taken into account? May we conclude that there is interaction between hospital of birth and gestational age? Plot the original data and the fitted regression equations.

Birth Weight (kg)	Gestation Age (weeks)	Hospital of Birth	Birth Weight (kg)	Gestation Age (weeks)	Hospital of Birth
1.4	30	A	1.0	29	C
.9	27	B	1.4	33	C
1.2	33	A	.9	28	A
1.1	29	C	1.0	28	C
1.3	35	A	1.9	36	B
.8	27	B	1.3	29	B
1.0	32	A	1.7	35	C
.7	26	A	1.0	30	A
1.2	30	C	.9	28	A
.8	28	A	1.0	31	A
1.5	32	B	1.6	31	B
1.3	31	A	1.6	33	B
1.4	32	C	1.7	34	B
1.5	33	B	1.6	35	C
1.0	27	A	1.2	28	A
1.8	35	B	1.5	30	B
1.4	36	C	1.8	34	B
1.2	34	A	1.5	34	C
1.1	28	B	1.2	30	A
1.2	30	B	1.2	32	C

18. Refer to Chapter 9, Review Exercise 18. In the study cited in that exercise, Maria Mathias (A-13) investigated the relationship between ages (AGE) of boys and improvement in measures of hyperactivity, attitude, and social behavior. In the study, subjects were randomly assigned to two different treatments. The control group (TREAT = 0) received standard therapy for hyperactivity, and the treatment group (TREAT = 1) received standard therapy plus pet therapy. The results are

shown in the following table. Create a scatter plot with age as the independent variable and ATT (change in attitude with positive numbers indicating positive change in attitude) as the dependent variable. Use different symbols for the two different treatment groups. Use multiple regression techniques to determine whether age, treatment, or the interaction are useful in predicting ATT. Report your results.

Subject	TREAT	AGE	ATT	Subject	TREAT	AGE	ATT
1	1	9	−1.2	17	0	10	0.4
2	1	9	0.0	18	0	7	0.0
3	1	13	−0.4	19	0	12	1.1
4	1	6	−0.4	20	0	9	0.2
5	1	9	1.0	21	0	7	0.4
6	1	8	0.8	22	0	6	0.0
7	1	8	−0.6	23	1	11	0.6
8	1	9	−1.2	24	1	11	0.4
9	0	7	0.0	25	1	11	1.0
10	0	12	0.4	26	1	11	0.8
11	0	9	−0.8	27	1	11	1.2
12	0	10	1.0	28	1	11	0.2
13	0	12	1.4	29	1	11	0.8
14	0	9	1.0	30	1	8	0.0
15	0	12	0.8	31	1	9	0.4
16	0	9	1.0				

Source: Data provided courtesy of Maria Mathias, M.D. and the Wright State University Statistical Consulting Center.

For each study described in Exercises 19 through 21, answer as many of the following questions as possible:

(a) Which is the dependent variable?

(b) What are the independent variables?

(c) What are the appropriate null and alternative hypotheses?

(d) Which null hypotheses do you think were rejected? Why?

(e) Which is the more relevant objective, prediction or estimation, or are the two equally relevant? Explain your answer.

(f) What is the sampled population?

(g) What is the target population?

(h) Which variables are related to which other variables? Are the relationships direct or inverse?

(i) Write out the regression equation using appropriate numbers for parameter estimates.

(j) Give numerical values for any other statistics that you can.

(k) Identify each variable as to whether it is quantitative or qualitative.

(l) Explain the meaning of any statistics for which numerical values are given.

19. Golfinopoulos and Arhonditsis (A-14) used a multiple regression model in a study of trihalomethanes (THMs) in drinking water in Athens, Greece. THMs are of concern since they have been related to cancer and reproductive outcomes. The researchers found the following regression model useful in

predicting THM:

$$THM = -.26chla + 1.57\,pH + 28.74Br - 66.72Br^2 \\ -43.63S + 1.13Sp + 2.62T \times S - .72T \times CL$$

The variables were as follows: $chla$ = chlorophyll concentration, pH = acid/base scale, Br = bromide concentration, S = dummy variable for summer, Sp = dummy variable for spring, T = Temperature, and CL = chlorine concentration. The researchers reported $R = .52$, $p < .001$.

20. In a study by Takata et al. (A-15), investigators evaluated the relationship between chewing ability and teeth number and measures of physical fitness in a sample of subjects ages 80 or higher in Japan. One of the outcome variables that measured physical fitness was leg extensor strength. To measure the ability to chew foods, subjects were asked about their ability to chew 15 foods (peanuts, vinegared octopus, and French bread, among others). Consideration of such variables as height, body weight, gender, systolic blood pressure, serum albumin, fasting glucose concentration, back pain, smoking, alcohol consumption, marital status, regular medical treatment, and regular exercise revealed that the number of chewable foods was significant in predicting leg extensor strength ($\hat{\beta}_1 = .075, p = .0366$). However, in the presence of the other variables, number of teeth was not a significant predictor ($\hat{\beta}_1 = .003, p = .9373$).

21. Varela et al. (A-16) examined 515 patients who underwent lung resection for bronchogenic carcinoma. The outcome variable was the occurrence of cardiorespiratory morbidity after surgery. Any of the following postoperative events indicated morbidity: pulmonary atelectasis or pneumonia, respiratory or ventilatory insufficiency at discharge, need for mechanical ventilation at any time after extubation in the operating room, pulmonary thromboembolism, arrhythmia, myocardial ischemia or infarct, and clinical cardiac insufficiency. Performing a stepwise logistic regression, the researchers found that age ($p < .001$) and postoperative forced expiratory volume ($p = .003$) were statistically significant in predicting the occurrence of cardiorespiratory morbidity.

For each of the data sets given in Exercises 22 through 29, do as many of the following as you think appropriate:

(a) Apply one or more of the techniques discussed in this chapter.

(b) Apply one or more of the techniques discussed in previous chapters.

(c) Construct graphs.

(d) Formulate relevant hypotheses, perform the appropriate tests, and find p values.

(e) State the statistical decisions and clinical conclusions that the results of your hypothesis tests justify.

(f) Describe the population(s) to which you think your inferences are applicable.

22. A study by Davies et al. (A-17) was motivated by the fact that, in previous studies of contractile responses to β-adrenoceptor agonists in single myocytes from failing and nonfailing human hearts, they had observed an age-related decline in maximum response to isoproterenol, at frequencies where the maximum response to high Ca^{2+} in the same cell was unchanged. For the present study, the investigators computed the isoproterenol/Ca^{2+} ratio (ISO/CA) from measurements taken on myocytes from patients ranging in age from 7 to 70 years. Subjects were classified as older (> 50 years) and younger. The following are the (ISO/CA) values, age, and myocyte source of subjects in the study. Myocyte sources were reported as donor and biopsy.

Age	ISO/CA	Myocyte Source
7	1.37	Donor
21	1.39	Donor
28	1.17	Donor
35	0.71	Donor
38	1.14	Donor
50	0.95	Donor
51	0.86	Biopsy
52	0.72	Biopsy
55	0.53	Biopsy
56	0.81	Biopsy
61	0.86	Biopsy
70	0.77	Biopsy

Source: Data provided courtesy of Dr. Sian E. Harding.

23. Hayton et al. (A-18) investigated the pharmacokinetics and bioavailability of cefetamet and cefetamet pivoxil in infants between the ages of 3.5 and 17.3 months who had received the antibiotic during and after urological surgery. Among the pharmacokinetic data collected were the following measurements of the steady-state apparent volume of distribution (V). Also shown are previously collected data on children ages 3 to 12 years (A-19) and adults (A-20). Weights (W) of subjects are also shown.

Infants		Children		Adults	
W (kg)	V (liters)	W (kg)	V (liters)	W (kg)	V (liters)
6.2	2.936	13	4.72	61	19.7
7.5	3.616	14	5.23	80	23.7
7.0	1.735	14	5.85	96	20.0
7.1	2.557	15	4.17	75	19.5
7.8	2.883	16	5.01	60	19.6
8.2	2.318	17	5.81	68	21.5
8.3	3.689	17	7.03	72.2	21.9
8.5	4.133	17.5	6.62	87	30.9
8.6	2.989	17	4.98	66.5	20.4
8.8	3.500	17.5	6.45		
10.0	4.235	20	7.73		
10.0	4.804	23	7.67		
10.2	2.833	25	9.82		
10.3	4.068	37	14.40		
10.6	3.640	28	10.90		
10.7	4.067	47	15.40		
10.8	8.366	29	9.86		
11.0	4.614	37	14.40		
12.5	3.168				
13.1	4.158				

Source: Data provided courtesy of Dr. Klaus Stoeckel.

24. According to Fils-Aime et al. (A-21), epidemiologic surveys have found that alcoholism is the most common mental or substance abuse disorder among men in the United States. Fils-Aime and associates investigated the interrelationships of age at onset of excessive alcohol consumption, family history of alcoholism, psychiatric comorbidity, and cerebrospinal fluid (CSF) monoamine metabolite concentrations in abstinent, treatment-seeking alcoholics. Subjects were mostly white males classified as experiencing early (25 years or younger) or late (older than 25 years) onset of excessive alcohol consumption. Among the data collected were the following measurements on CSF tryptophan (TRYPT) and 5-hydroxyindoleacetic acid (5-HIAA) concentrations (pmol/ml).

5-HIAA	TRYPT	Onset 1 = Early 0 = Late	5-HIAA	TRYPT	Onset 1 = Early 0 = Late
57	3315	1	102	3181	1
116	2599	0	51	2513	1
81	3334	1	92	2764	1
78	2505	0	104	3098	1
206	3269	0	50	2900	1
64	3543	1	93	4125	1
123	3374	0	146	6081	1
147	2345	1	96	2972	1
102	2855	1	112	3962	0
93	2972	1	23	4894	1
128	3904	0	109	3543	1
69	2564	1	80	2622	1
20	8832	1	111	3012	1
66	4894	0	85	2685	1
90	6017	1	131	3059	0
103	3143	0	58	3946	1
68	3729	0	110	3356	0
81	3150	1	80	3671	1
143	3955	1	42	4155	1
121	4288	1	80	1923	1
149	3404	0	91	3589	1
82	2547	1	102	3839	0
100	3633	1	93	2627	0
117	3309	1	98	3181	0
41	3315	1	78	4428	0
223	3418	0	152	3303	0
96	2295	1	108	5386	1
87	3232	0	102	3282	1
96	3496	1	122	2754	1
34	2656	1	81	4321	1
98	4318	1	81	3386	1
86	3510	0	99	3344	1
118	3613	1	73	3789	1
84	3117	1	163	2131	1
99	3496	1	109	3030	0
114	4612	1	90	4731	1

(*Continued*)

5-HIAA	TRYPT	Onset 1 = Early 0 = Late	5-HIAA	TRYPT	Onset 1 = Early 0 = Late
140	3051	1	110	4581	1
74	3067	1	48	3292	0
45	2782	1	77	4494	0
51	5034	1	67	3453	1
99	2564	1	92	3373	1
54	4335	1	86	3787	0
93	2596	1	101	3842	1
50	2960	1	88	2882	1
118	3916	0	38	2949	1
96	2797	0	75	2248	0
49	3699	1	35	3203	0
133	2394	0	53	3248	1
105	2495	0	77	3455	0
61	2496	1	179	4521	1
197	2123	1	151	3240	1
87	3320	0	57	3905	1
50	3117	1	45	3642	1
109	3308	0	76	5233	0
59	3280	1	46	4150	1
107	3151	1	98	2579	1
85	3955	0	84	3249	1
156	3126	0	119	3381	0
110	2913	0	41	4020	1
81	3786	1	40	4569	1
53	3616	1	149	3781	1
64	3277	1	116	2346	1
57	2656	1	76	3901	1
29	4953	0	96	3822	1
34	4340	1			

Source: Data provided courtesy of Dr. Markku Linnoila.

25. The objective of a study by Abrahamsson et al. (A-22) was to investigate the anti-thrombotic effects of an inhibitor of the plasminogen activator inhibitor-1 (PAI-1) in rats given endotoxin. Experimental subjects were male Sprague–Dawley rats weighing between 300 and 400 grams. Among the data collected were the following measurements on PAI-1 activity and the lung ^{125}I-concentration in anesthetized rats given three drugs:

Drugs	Plasma PAI-1 Activity (U/ml)	^{125}I-Fibrin in the Lungs (% of Ref. Sample)
Endotoxin	127	158
	175	154
	161	118
	137	77
	219	172

(*Continued*)

Drugs	Plasma PAI-1 Activity (U/ml)	^{125}I-Fibrin in the Lungs (% of Ref. Sample)
	260	277
	203	216
	195	169
	414	272
	244	192
Endotoxin + PRAP = 1 low dose	107	49
	103	28
	248	187
	164	109
	176	96
	230	126
	184	148
	276	17
	201	97
	158	86
Endotoxin + PRAP = 1 high dose	132	86
	130	24
	75	17
	140	41
	166	114
	194	110
	121	26
	111	53
	208	71
	211	90

Source: Data provided courtesy of Dr. Tommy Abrahamsson.

26. Pearse and Sylvester (A-23) conducted a study to determine the separate contributions of ischemia and extracorporeal perfusion to vascular injury occurring in isolated sheep lungs and to determine the oxygen dependence of this injury. Lungs were subjected to ischemia alone, extracorporeal perfusion alone, and both ischemia and extracorporeal perfusion. Among the data collected were the following observations on change in pulmonary arterial pressure (mm Hg) and pulmonary vascular permeability assessed by estimation of the reflection coefficient for albumin in perfused lungs with and without preceding ischemia:

Ischemic–Perfused Lungs		Perfused Lungs	
Change in Pulmonary Pressure	Reflection Coefficient	Change in Pulmonary Pressure	Reflection Coefficient
8.0	0.220	34.0	0.693
3.0	0.560	31.0	0.470
10.0	0.550	4.0	0.651
23.0	0.806	48.0	0.999

Ischemic–Perfused Lungs		Perfused Lungs	
Change in Pulmonary Pressure	Reflection Coefficient	Change in Pulmonary Pressure	Reflection Coefficient
15.0	0.472	32.0	0.719
43.0	0.759	27.0	0.902
18.0	0.489	25.0	0.736
27.0	0.546	25.0	0.718
13.0	0.548		
0.0	0.467		

Source: Data provided courtesy of Dr. David B. Pearse.

27. The purpose of a study by Balzamo et al. (A-24) was to investigate, in anesthetized rabbits, the effects of mechanical ventilation on the concentration of substance P (SP) measured by radioimmunoassay in nerves and muscles associated with ventilation and participating in the sensory innervation of the respiratory apparatus and heart. SP is a neurotransmitter located in primary sensory neurons in the central and autonomic nervous systems. Among the data collected were the following measures of SP concentration in cervical vagus nerves (X) and corresponding nodose ganglia (NG), right and left sides:

SPXright	SPNGright	SPXleft	SPNGleft
0.6500	9.6300	3.3000	1.9300
2.5600	3.7800	0.6200	2.8700
1.1300	7.3900	0.9600	1.3100
1.5500	3.2800	2.7000	5.6400
35.9000	22.0000	4.5000	9.1000
19.0000	22.8000	8.6000	8.0000
13.6000	2.3000	7.0000	8.3000
8.0000	15.8000	4.1000	4.7000
7.4000	1.6000	5.5000	2.5000
3.3000	11.6000	9.7000	8.0000
19.8000	18.0000	13.8000	8.0000
8.5000	6.2000	11.0000	17.2000
5.4000	7.8000	11.9000	5.3000
11.9000	16.9000	8.2000	10.6000
47.7000	35.9000	3.9000	3.3000
14.2000	10.2000	3.2000	1.9000
2.9000	1.6000	2.7000	3.5000
6.6000	3.7000	2.8000	2.5000
3.7000	1.3000		

Source: Data provided courtesy of Dr. Yves Jammes.

28. Scheeringa and Zeanah (A-25) examined the presence of posttraumatic stress disorder (PTSD), the severity of posttraumatic symptomatology, and the pattern of expression of symptom clusters in relation to six independent variables that may be salient to the development of a posttraumatic disorder in children under 48 months of age. The following data were collected during the course of the study.

Predictor Variables						Response Variables			
Gender	Age	Acute/Rept.	Injury	Wit./ Exper.	Threat to Caregiver	Reexp	Numb	Arous	FrAgg
0	1	0	1	1	1	3	0	0	1
0	1	0	0	0	1	2	2	1	1
1	1	0	0	0	1	3	1	1	1
0	1	0	0	0	1	3	1	0	4
1	0	1	1	1	0	1	3	1	1
1	1	0	1	1	0	3	1	0	1
0	1	0	1	1	0	4	2	0	1
0	1	0	0	1	0	5	2	0	4
1	1	0	0	0	1	2	1	3	2
1	1	1	1	1	0	4	1	0	0
0	0	1	1	1	0	1	3	0	1
1	0	1	0	1	0	1	3	0	2
1	0	1	1	1	0	0	3	0	0
1	1	0	1	1	0	4	1	2	1
1	0	0	1	1	1	3	2	1	3
1	0	0	1	1	1	3	1	2	1
0	1	0	1	1	1	3	1	2	2
0	1	0	0	0	1	5	2	1	1
0	1	0	0	0	1	1	2	2	2
0	1	0	1	1	0	4	4	0	3
1	0	1	1	1	0	2	1	2	3
1	0	0	1	1	1	1	1	2	1
1	1	0	0	0	1	4	1	1	1
0	1	0	0	0	1	3	2	1	0
0	1	0	0	0	1	3	1	2	4
0	1	0	0	0	1	3	1	2	4
0	1	0	0	1	0	2	2	0	0
1	1	0	0	0	1	2	0	3	0
1	1	0	0	0	1	2	0	1	2
0	1	0	1	0	1	2	3	1	3
1	1	1	0	1	0	1	2	1	1
1	1	0	1	1	1	3	2	0	4
1	1	0	0	0	0	2	4	2	0
0	1	0	0	0	1	1	1	0	2
0	0	1	0	0	1	2	3	2	3
0	0	1	0	0	1	3	1	4	3
0	0	1	0	0	1	3	1	2	3
0	0	0	0	1	0	1	1	0	0
1	0	0	0	0	1	4	3	2	3
1	0	0	1	1	0	4	2	3	2
0	0	1	1	1	0	1	2	2	1

Predictor Variables						Response Variables			
Gender	Age	Acute/Rept.	Injury	Wit./ Exper.	Threat to Caregiver	Reexp	Numb	Arous	FrAgg

Key:

Gender 0 = male; 1 = female

Age 0 = younger than 18 months at time of trauma; 1 = older than 18 months

Acute/Rept. 0 = trauma was acute, single blow; 1 = trauma was repealed or chronic

Injury 0 = subject was not injured in the trauma; 1 = subject was physically injured in the trauma

Wit./Exper. 0 = subject witnessed but did not directly experience trauma; 1 = subject directly experienced the trauma

Threat to Caregiver 0 = caregiver was not threatened in the trauma; 1 = caregiver was threatened in the trauma

Reexp = Reexperiencing cluster symptom count
Numb = Numbing of responsiveness/avoidance cluster symptom count
Arous = Hyperarousal cluster symptom count
FrAgg = New fears/aggression cluster symptom count

Source: Data provided courtesy of Dr. Michael S. Scheeringa.

29. One of the objectives of a study by Mulloy and McNicholas (A-26) was to compare ventilation and gas exchange during sleep and exercise in chronic obstructive pulmonary disease (COPD). The investigators wished also to determine whether exercise studies could aid in the prediction of nocturnal desaturation in COPD. Subjects (13 male, 6 female) were ambulatory patients attending an outpatient respiratory clinic. The mean age of the patients, all of whom had severe, stable COPD, was 64.8 years with a standard deviation of 5.2. Among the data collected were measurements on the following variables:

Age (years)	BMI	PaO_2 (mm Hg)	$PaCO_2$ (mm Hg)	FEV_1 (% Predicted)	Lowest Ex. Sao_2[a]	Mean Sleep Sao_2[a]	Lowest Sleep Sao_2[a]	Fall Sleep Sao_2[a]
67	23.46	52.5	54	22	74	70.6	56	29.6
62	25.31	57.75	49.575	19	82	85.49	76	11.66
68	23.11	72	43.8	41	95	88.72	82	11.1
61	25.15	72	47.4	38	88	91.11	76	18.45
70	24.54	78	40.05	40	88	92.86	92	0.8
71	25.47	63.75	45.375	31	85	88.95	80	13
60	19.49	80.25	42.15	28	91	94.78	90	4
57	21.37	84.75	40.2	20	91	93.72	89	5.8
69	25.78	68.25	43.8	32	85	90.91	79	13
57	22.13	83.25	43.725	20	88	94.39	86	9.5
74	26.74	57.75	51	33	75	89.89	80	14.11
63	19.07	78	44.175	36	81	93.95	82	13
64	19.61	90.75	40.35	27	90	95.07	92	4
73	30.30	69.75	38.85	53	87	90	76	18

(*Continued*)

Age (years)	BMI	PaO_2 (mm Hg)	$PaCO_2$ (mm Hg)	FEV_1 (% Predicted)	Lowest Ex. Sao_2[a]	Mean Sleep Sao_2[a]	Lowest Sleep Sao_2[a]	Fall Sleep Sao_2[a]
63	26.12	51.75	46.8	39	67	69.31	46	34.9
62	21.71	72	41.1	27	88	87.95	72	22
67	24.75	84.75	40.575	45	87	92.95	90	2.17
57	25.98	84.75	40.05	35	94	93.4	86	8.45
66	32.00	51.75	53.175	30	83	80.17	71	16

[a] Treated as dependent variable in the authors' analyses. BMI = body mass index; Pao_2 = arterial oxygen tension: $Paco_2$ = arterial carbon dioxide pressure; FEV_1 = forced expiratory volume in 1 second; Sao_2 = arterial oxygen saturation.
Source: Data provided courtesy of Dr. Eithne Mulloy.

Exercises for Use with the Large Data Sets Available on the Following Website: www.wiley.com/college/daniel

1. The goal of a study by Gyurcsik et al. (A-27) was to examine the usefulness of aquatic exercise-related goals, task self-efficacy, and scheduling self-efficacy for predicting aquatic exercise attendance by individuals with arthritis. The researchers collected data on 142 subjects participating in Arthritis Foundation Aquatics Programs. The outcome variable was the percentage of sessions attended over an 8-week period (ATTEND). The following predictor variables are all centered values. Thus, for each participant, the mean for all participants is subtracted from the individual score. The variables are:

GOALDIFF—higher values indicate setting goals of higher participation.

GOALSPEC—higher values indicate higher specificity of goals related to aquatic exercise.

INTER—interaction of GOALDIFF and GOALSPEC.

TSE—higher values indicate participants' confidence in their abilities to attend aquatic classes.

SSE—higher values indicate participants' confidence in their abilities to perform eight tasks related to scheduling exercise into their daily routine for 8 weeks.

MONTHS—months of participation in aquatic exercise prior to start of study.

With the data set AQUATICS, perform a multiple regression to predict ATTEND with each of the above variables. What is the multiple correlation coefficient? What variables are significant in predicting ATTEND? What are your conclusions?

2. Rodehorst (A-28) conducted a prospective study of 212 rural elementary school teachers. The main outcome variable was the teachers' intent to manage children demonstrating symptoms of asthma in their classrooms. This variable was measured with a single-item question that used a seven-point Likert scale (INTENT, with possible responses of 1 = extremely probable to 7 = extremely improbable). Rodehorst used the following variables as independent variables to predict INTENT:

SS = Social Support. Scores range from 7 to 49, with higher scores indicating higher perceived social support for managing children with asthma in a school setting.

ATT = Attitude. Scores range from 15 to 90, with higher scores indicating more favorable attitudes toward asthma.

KNOW = Knowledge. Scores range from 0 to 24, with higher scores indicating higher general knowledge about asthma.

CHILD =Number of children with asthma the teacher has had in his or her class during his or her entire teaching career.

SE = Self-efficacy. Scores range from 12 to 60, with higher scores indicating higher self-efficacy for managing children with asthma in the school setting.

YRS =Years of teaching experience.

With the data TEACHERS, use stepwise regression analysis to select the most useful variables to include in a model for predicting INTENT.

3. Refer to the weight loss data on 588 cancer patients and 600 healthy controls (WGTLOSS). Weight loss among cancer patients is a well-known phenomenon. Of interest to clinicians is the role played in the process by metabolic abnormalities. One investigation into the relationships among these variables yielded data on whole-body protein turnover (Y) and percentage of ideal body weight for height (X). Subjects were lung cancer patients and healthy controls of the same age. Select a simple random sample of size 15 from each group and do the following:

(a) Draw a scatter diagram of the sample data using different symbols for each of the two groups.

(b) Use dummy variable coding to analyze these data.

(c) Plot the two regression lines on the scatter diagram. May one conclude that the two sampled populations differ with respect to mean protein turnover when percentage of ideal weight is taken into account?

May one conclude that there is interaction between health status and percentage of ideal body weight? Prepare a verbal interpretation of the results of your analysis and compare your results with those of your classmates.

REFERENCES

Methodology References

1. Bruce L. Bowerman and Richard T. O'connell, *Linear Statistical Models: An Applied Approach*, 2nd ed. PWS-Kent Publishing, Boston, 1990.
2. David W. Hosmer and Stanley Lemeshow, *Applied Logistic Regression*, 2nd Ed. Wiley, New York, 2000.
3. David G. Kleinbaum, *Logistic Regression: A Self-Learning Text*, New York, Springer, 1994.
4. D. R. Cox and E. J. Snell, *Analysis of Binary Data*, 2nd ed. Chapman and Hall/CRC, New York, 1989.
5. N. J. D. Nagelkerke, "A Note on a General Definition of the Coefficient of Determination", *Biometrika*, *78* (1991), 691–692.

Applications References

A-1. North Carolina State Center for Health Statistics and Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill, Birth Data set for 2001 found at www.irss.unc.edu/ncvital/bfd1down.html. All sampling and coding performed by John Holcomb and do not represent the findings of the Center or Institute.

A-2. B. Bolwell, R. Sobecks, B. Pohlman, S. Andresen, K. Theil, S. Serafino, L. Rybicki, and M. Kalaycio, "Etoposide (VP-16) Plus G-CSF Mobilizes Different Dendritic Cell Subsets than Does G-CSF Alone," *Bone Marrow Transplantation*, *31* (2003), 95–98.

A-3. Manoj Pandey, Lal B. Sharma, and Vijay K. Shukla, "Cytochrome P-450 Expression and Lipid Peroxidation in Gallbladder Cancer," *Journal of Surgical Oncology*, *82* (2003), 180–183.

A-4. Mori J. Krantz, Ilana B. Kutinsky, Alastair D. Robertson, and Philip S. Mehler, "Dose-Related Effects of Methadone on QT Prolongation in a Series of Patients with Torsade de Pointes," *Pharmacotherapy*, *23* (2003), 802–805.

A-5. Robert A. Reiss, Curtis E. Haas, Deborah L. Griffis, Bernadette Porter, and Mary Ann Tara, "Point-of-Care versus Laboratory Monitoring of Patients Receiving Different Anticoagulant Therapies," *Pharmacotherapy 22*, (2002), 677–685.

A-6. Gwi-Ryung Son, May L. Wykle, and Jaclene A. Zauszniewski, "Korean Adult Child Caregivers of Older Adults with Dementia," *Journal of Gerontological Nursing*, *29* (2003), 19–28.

A-7. M. Naeije, "Local Kinematic and Anthropometric Factors Related to the Maximum Mouth Opening in Healthy Individuals," *Journal of Oral Rehabilitation*, *29* (2002), 534–539.

A-8. Daniel F. Connor, Ronald J. Steingard, Jennifer J. Anderson, and Richard H. Melloni, "Gender Differences in Reactive and Proactive Aggression," *Child Psychiatry and Human Development*, *33* (2003), 279–294.

A-9. Daniel H. LaMont, Matthew J. Budoff, David M. Shavelle, Robert Shavelle, Bruce H. Brundage, and James M. Hagar, "Coronary Calcium Scanning Adds Incremental Value to Patients with Positive Stress Tests," *American Heart Journal*, *143* (2002), 861–867.

A-10. Robyn Gallagher, Sharon McKinley, and Kathleen Dracup, "Predictors of Women's Attendance at Cardiac Rehabilitation Programs," *Progress in Cardiovascular Nursing*, *18* (2003), 121–126.

A-11. John H. Porcerelli, Rosemary Cogan, Patricia P. West, Edward A. Rose, Dawn Lambrecht, Karen E. Wilson, Richard K. Severson, and Dunia Karana, "Violent Victimization of Women and Men: Physical and Psychiatric Symptoms," *Journal of the American Board of Family Practice*, *16* (2003), 32–39.

A-12. Debra A. Jansen, and Mary L. Keller, "Cognitive Function in Community-Dwelling Elderly Women," *Journal of Gerontological Nursing*, *29* (2003), 34–43.

A-13. Maria Mathias and the Wright State University Statistical Consulting Center.

A-14. Spyros K. Golfinopoulos and George B. Arhonditsis, "Multiple Regression Models: A Methodology for Evaluating Trihalomethane Concentrations in Drinking Water from Raw Water Characteristics," *Chemosphere*, *47* (2002), 1007–1018.

A-15. Y. Takata, T. Ansai, S. Awano, T. Hamasaki, Y. Yoshitake, Y. Kimura, K. Sonoki, M. Wakisaka, M. Fukuhara, and T. Takehara, "Relationship of Physical Fitness to Chewing in an 80-Year-Old Population," *Oral Diseases*, *10* (2004), 44–49.

A-16. G. Varela, N. Novoa, M. F. Jiménez, and G. Santos, "Applicability of Logistic Regression (LR) Risk Modeling to Decision Making in Lung Cancer Resection," *Interactive Cardiovascular and Thoracic Surgery*, *2* (2003), 12–15.

A-17. C. H. Davies, N. Ferrara, and S. E. Harding, "β-Adrenoceptor Function Changes with Age of Subject in Myocytes from Non-Failing Human Ventricle," *Cardiovascular Research*, *31* (1996), 152–156.

A-18. William L. Hayton, Johannes Kneer, Ronald Groot, and Klaus Stoeckel, "Influence of Maturation and Growth on Cefetamet Pivoxil Pharmacokinetics: Rational Dosing for Infants," *Antimicrobial Agents and Chemotherapy*, *40* (1996), 567–574.

A-19. W. L. Hayton, R. A. Walstad, E. Thurmann-Nielsen, T. Kufaas, J. Kneer, R. J. Ambros, H. E. Rugstad, E. Monn, E. Bodd, and K. Stoeckel, "Pharmacokinetics of Intravenous Cefetamet and Oral Cefetamet Pivoxil in Children," *Antimicrobial Agents and Chemotherapy*, *35* (1991), 720–725. Erratum, *36* (1992), 2575.

A-20. M. P. Ducharme, D. J. Edwards, P. J. McNamara, and K. Stoeckel, "Bioavailability of Syrup and Tablet Formulations of Cefetamet Pivoxil," *Antimicrobial Agents and Chemotherapy*, *37* (1993), 2706–2709.

A-21. Marie-Lourdes Fils-Aime, Michael J. Eckardt, David T. George, Gerald L. Brown, Ivan Mefford, and Markku Linnoila, "Early-Onset Alcoholics Have Lower Cerebrospinal Fluid 5-Hydroxyindoleacetic Acid Levels than Late-Onset Alcoholics," *Archives of General Psychiatry*, *53* (1996), 211–216.

A-22. T. Abrahamsson, V. Nerme, M. Strömqvist, B. Åkerblom, A. Legnehed, K. Pettersson, and A. Westin Eriksson, "Anti-thrombotic Effect of PAI-1 Inhibitor in Rats Given Endotoxin," *Thrombosis and Haemostasis*, *75* (1996), 118–126.

A-23. David B. Pearse and J. T. Sylvester, "Vascular Injury in Isolated Sheep Lungs: Role of Ischemia, Extracorporeal Perfusion, and Oxygen," *American Journal of Respiratory and Critical Care Medicine*, *153* (1996), 196–202.

A-24. Emmanuel Balzamo, Pierre Joanny, Jean Guillaume Steinberg, Charles Oliver, and Yves Jammes, "Mechanical Ventilation Increases Substance P Concentration in the Vagus, Sympathetic, and Phrenic Nerves," *American Journal of Respiratory and Critical Care Medicine*, *153* (1996), 153–157.

A-25. Michael S. Scheeringa and Charles H. Zeanah, "Symptom Expression and Trauma Variables in Children Under 48 Months of Age," *Infant Mental Health Journal*, *16* (1995), 259–270.

A-26. Eithne Mulloy and Walter T. McNicholas, "Ventilation and Gas Exchange During Sleep and Exercise in Severe COPD," *Chest*, *109* (1996), 387–394.

A-27. Nancy C. Gyurcsik, Paul A. Estabrooks, and Melissa J. Frahm-Templar, "Exercise-Related Goals and Self-Efficacy as Correlates of Aquatic Exercise in Individuals with Arthritis," *Arthritis Care and Research*, *49* (2003), 306–313.

A-28. T. Kim Rodehurst, "Rural Elementary School Teachers' Intent to Manage Children with Asthma Symptoms," *Pediatric Nursing*, *29* (2003), 184–194.

CHAPTER **12**

THE CHI-SQUARE DISTRIBUTION AND THE ANALYSIS OF FREQUENCIES

CHAPTER OVERVIEW

This chapter explores techniques that are commonly used in the analysis of count or frequency data. Uses of the chi-square distribution, which was mentioned briefly in Chapter 6, are discussed and illustrated in greater detail. Additionally, statistical techniques often used in epidemiological studies are introduced and demonstrated by means of examples.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand the mathematical properties of the chi-square distribution.
2. be able to use the chi-square distribution for goodness-of-fit tests.
3. be able to construct and use contingency tables to test independence and homogeneity.
4. be able to apply Fisher's exact test for 2×2 tables.
5. understand how to calculate and interpret the epidemiological concepts of relative risk, odds ratios, and the Mantel-Haenszel statistic.

12.1 INTRODUCTION

In the chapters on estimation and hypothesis testing, brief mention is made of the chi-square distribution in the construction of confidence intervals for, and the testing of, hypotheses concerning a population variance. This distribution, which is one of the most widely used distributions in statistical applications, has many other uses. Some of the more common ones are presented in this chapter along with a more complete description of the distribution itself, which follows in the next section.

The chi-square distribution is the most frequently employed statistical technique for the analysis of count or frequency data. For example, we may know for a sample of hospitalized patients how many are male and how many are female. For the same sample we may also know how many have private insurance coverage, how many have Medicare insurance, and how many are on Medicaid assistance. We may wish to know, for the population from which the sample was drawn, if the type of insurance coverage differs according to gender. For another sample of patients, we may have frequencies for each diagnostic category represented and for each geographic area represented. We might want to know if, in the population from which the same was drawn, there is a relationship between area of residence and diagnosis. We will learn how to use chi-square analysis to answer these types of questions.

There are other statistical techniques that may be used to analyze frequency data in an effort to answer other types of questions. In this chapter we will also learn about these techniques.

12.2 THE MATHEMATICAL PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

The chi-square distribution may be derived from normal distributions. Suppose that from a normally distributed random variable Y with mean μ and variance σ^2 we randomly and independently select samples of size $n = 1$. Each value selected may be transformed to the standard normal variable z by the familiar formula

$$z_i = \frac{y_i - \mu}{\sigma} \tag{12.2.1}$$

Each value of z may be squared to obtain z^2. When we investigate the sampling distribution of z^2, we find that it follows a chi-square distribution with 1 degree of freedom. That is,

$$\chi^2_{(1)} = \left(\frac{y - \mu}{\sigma}\right)^2 = z^2$$

Now suppose that we randomly and independently select samples of size $n = 2$ from the normally distributed population of Y values. Within each sample we may transform each

value of y to the standard normal variable z and square as before. If the resulting values of z^2 for each sample are added, we may designate this sum by

$$\chi^2_{(2)} = \left(\frac{y_1 - \mu}{\sigma}\right)^2 + \left(\frac{y_2 - \mu}{\sigma}\right)^2 = z_1^2 + z_2^2$$

since it follows the chi-square distribution with 2 degrees of freedom, the number of independent squared terms that are added together.

The procedure may be repeated for any sample size n. The sum of the resulting z^2 values in each case will be distributed as chi-square with n degrees of freedom. In general, then,

$$\chi^2_{(n)} = z_1^2 + z_2^2 + \cdots + z_n^2 \tag{12.2.2}$$

follows the chi-square distribution with n degrees of freedom. The mathematical form of the chi-square distribution is as follows:

$$f(u) = \frac{1}{\left(\frac{k}{2} - 1\right)!} \frac{1}{2^{k/2}} u^{(k/2)-1} e^{-(u/2)}, \quad u > 0 \tag{12.2.3}$$

where e is the irrational number 2.71828 . . . and k is the number of degrees of freedom. The variate u is usually designated by the Greek letter chi (χ) and, hence, the distribution is called the chi-square distribution. As we pointed out in Chapter 6, the chi-square distribution has been tabulated in Appendix Table F. Further use of the table is demonstrated as the need arises in succeeding sections.

The mean and variance of the chi-square distribution are k and $2k$, respectively. The modal value of the distribution is $k - 2$ for values of k greater than or equal to 2 and is zero for $k = 1$.

The shapes of the chi-square distributions for several values of k are shown in Figure 6.9.1. We observe in this figure that the shapes for $k = 1$ and $k = 2$ are quite different from the general shape of the distribution for $k > 2$. We also see from this figure that chi-square assumes values between 0 and infinity. It cannot take on negative values, since it is the sum of values that have been squared. A final characteristic of the chi-square distribution worth noting is that the sum of two or more independent chi-square variables also follows a chi-square distribution.

Types of Chi-Square Tests As already noted, we make use of the chi-square distribution in this chapter in testing hypotheses where the data available for analysis are in the form of frequencies. These hypothesis testing procedures are discussed under the topics of *tests of goodness-of-fit, tests of independence,* and *tests of homogeneity.* We will discover that, in a sense, all of the chi-square tests that we employ may be thought of as goodness-of-fit tests, in that they test the goodness-of-fit of observed frequencies to frequencies that one would expect if the data were generated under some particular theory or hypothesis. We, however, reserve the phrase "goodness-of-fit" for use in a more

restricted sense. We use it to refer to a comparison of a sample distribution to some theoretical distribution that it is assumed describes the population from which the sample came. The justification of our use of the distribution in these situations is due to Karl Pearson (1), who showed that the chi-square distribution may be used as a test of the agreement between observation and hypothesis whenever the data are in the form of frequencies. An extensive treatment of the chi-square distribution is to be found in the book by Lancaster (2). Nikulin and Greenwood (3) offer practical advice for conducting chi-square tests.

Observed Versus Expected Frequencies The chi-square statistic is most appropriate for use with categorical variables, such as marital status, whose values are the categories married, single, widowed, and divorced. The quantitative data used in the computation of the test statistic are the frequencies associated with each category of the one or more variables under study. There are two sets of frequencies with which we are concerned, *observed frequencies* and *expected frequencies*. The observed frequencies are the number of subjects or objects in our sample that fall into the various categories of the variable of interest. For example, if we have a sample of 100 hospital patients, we may observe that 50 are married, 30 are single, 15 are widowed, and 5 are divorced. Expected frequencies are the number of subjects or objects in our sample that we would expect to observe if some null hypothesis about the variable is true. For example, our null hypothesis might be that the four categories of marital status are equally represented in the population from which we drew our sample. In that case we would expect our sample to contain 25 married, 25 single, 25 widowed, and 25 divorced patients.

The Chi-Square Test Statistic The test statistic for the chi-square tests we discuss in this chapter is

$$X^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] \tag{12.2.4}$$

When the null hypothesis is true, X^2 is distributed approximately as χ^2 with $k - r$ degrees of freedom. In determining the degrees of freedom, k is equal to the number of groups for which observed and expected frequencies are available, and r is the number of restrictions or constraints imposed on the given comparison. A restriction is imposed when we force the sum of the expected frequencies to equal the sum of the observed frequencies, and an additional restriction is imposed for each parameter that is estimated from the sample.

In Equation 12.2.4, O_i is the observed frequency for the ith category of the variable of interest, and E_i is the expected frequency (given that H_0 is true) for the ith category.

The quantity X^2 is a measure of the extent to which, in a given situation, pairs of observed and expected frequencies agree. As we will see, the nature of X^2 is such that when there is close agreement between observed and expected frequencies it is small, and when the agreement is poor it is large. Consequently, only a sufficiently large value of X^2 will cause rejection of the null hypothesis.

If there is perfect agreement between the observed frequencies and the frequencies that one would expect, given that H_0 is true, the term $O_i - E_i$ in Equation 12.2.4 will be

equal to zero for each pair of observed and expected frequencies. Such a result would yield a value of X^2 equal to zero, and we would be unable to reject H_0.

When there is disagreement between observed frequencies and the frequencies one would expect given that H_0 is true, at least one of the $O_i - E_i$ terms in Equation 12.2.4 will be a nonzero number. In general, the poorer the agreement between the O_i and the E_i, the greater or the more frequent will be these nonzero values. As noted previously, if the agreement between the O_i and the E_i is sufficiently poor (resulting in a sufficiently large X^2 value,) we will be able to reject H_0.

When there is disagreement between a pair of observed and expected frequencies, the difference may be either positive or negative, depending on which of the two frequencies is the larger. Since the measure of agreement, X^2, is a sum of component quantities whose magnitudes depend on the difference $O_i - E_i$, positive and negative differences must be given equal weight. This is achieved by squaring each $O_i - E_i$ difference. Dividing the squared differences by the appropriate expected frequency converts the quantity to a term that is measured in original units. Adding these individual $(O_i - E_i)^2/E_i$ terms yields X^2, a summary statistic that reflects the extent of the overall agreement between observed and expected frequencies.

The Decision Rule The quantity $\sum[(O_i - E_i)^2/E_i]$ will be small if the observed and expected frequencies are close together and will be large if the differences are large.

The computed value of X^2 is compared with the tabulated value of χ^2 with $k - r$ degrees of freedom. The decision rule, then, is: Reject H_0 if X^2 is greater than or equal to the tabulated χ^2 for the chosen value of α.

Small Expected Frequencies Frequently in applications of the chi-square test the expected frequency for one or more categories will be small, perhaps much less than 1. In the literature the point is frequently made that the approximation of X^2 to χ^2 is not strictly valid when some of the expected frequencies are small. There is disagreement among writers, however, over what size expected frequencies are allowable before making some adjustment or abandoning χ^2 in favor of some alternative test. Some writers, especially the earlier ones, suggest lower limits of 10, whereas others suggest that all expected frequencies should be no less than 5. Cochran (4,5), suggests that for goodness-of-fit tests of unimodal distributions (such as the normal), the minimum expected frequency can be as low as 1. If, in practice, one encounters one or more expected frequencies less than 1, adjacent categories may be combined to achieve the suggested minimum. Combining reduces the number of categories and, therefore, the number of degrees of freedom. Cochran's suggestions appear to have been followed extensively by practitioners in recent years.

12.3 TESTS OF GOODNESS-OF-FIT

As we have pointed out, a goodness-of-fit test is appropriate when one wishes to decide if an observed distribution of frequencies is incompatible with some preconceived or hypothesized distribution.

We may, for example, wish to determine whether or not a sample of observed values of some random variable is compatible with the hypothesis that it was drawn from a population of values that is normally distributed. The procedure for reaching a decision consists of placing the values into mutually exclusive categories or class intervals and noting the frequency of occurrence of values in each category. We then make use of our knowledge of normal distributions to determine the frequencies for each category that one could expect if the sample had come from a normal distribution. If the discrepancy is of such magnitude that it could have come about due to chance, we conclude that the sample may have come from a normal distribution. In a similar manner, tests of goodness-of-fit may be carried out in cases where the hypothesized distribution is the binomial, the Poisson, or any other distribution. Let us illustrate in more detail with some examples of tests of hypotheses of goodness-of-fit.

EXAMPLE 12.3.1 *The Normal Distribution*

Cranor and Christensen (A-1) conducted a study to assess short-term clinical, economic, and humanistic outcomes of pharmaceutical care services for patients with diabetes in community pharmacies. For 47 of the subjects in the study, cholesterol levels are summarized in Table 12.3.1.

We wish to know whether these data provide sufficient evidence to indicate that the sample did not come from a normally distributed population. Let $\alpha = .05$

Solution:

1. **Data.** See Table 12.3.1.
2. **Assumptions.** We assume that the sample available for analysis is a simple random sample.

TABLE 12.3.1 Cholesterol Levels as Described in Example 12.3.1

Cholesterol Level (mg/dl)	Number of Subjects
100.0–124.9	1
125.0–149.9	3
150.0–174.9	8
175.0–199.9	18
200.0–224.9	6
225.0–249.9	4
250.0–274.9	4
275.0–299.9	3

Source: Data provided courtesy of Carole W. Cranor, and Dale B. Christensen, "The Asheville Project: Short-Term Outcomes of a Community Pharmacy Diabetes Care Program," *Journal of the American Pharmaceutical Association,* 43 (2003), 149–159.

3. **Hypotheses.**
   H_0: In the population from which the sample was drawn, cholesterol levels are normally distributed.
   H_A: The sampled population is not normally distributed.
4. **Test statistic.** The test statistic is

$$X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

5. **Distribution of test statistic.** If H_0 is true, the test statistic is distributed approximately as chi-square with $k - r$ degrees of freedom. The values of k and r will be determined later.
6. **Decision rule.** We will reject H_0 if the computed value of X^2 is equal to or greater than the critical value of chi-square.
7. **Calculation of test statistic.** Since the mean and variance of the hypothesized distribution are not specified, the sample data must be used to estimate them. These parameters, or their estimates, will be needed to compute the frequency that would be expected in each class interval when the null hypothesis is true. The mean and standard deviation computed from the grouped data of Table 12.3.1 are

$$\bar{x} = 198.67$$
$$s = 41.31$$

As the next step in the analysis, we must obtain for each class interval the frequency of occurrence of values that we would expect when the null hypothesis is true, that is, if the sample were, in fact, drawn from a normally distributed population of values. To do this, we first determine the expected relative frequency of occurrence of values for each class interval and then multiply these expected relative frequencies by the total number of values to obtain the expected number of values for each interval.

The Expected Relative Frequencies

It will be recalled from our study of the normal distribution that the relative frequency of occurrence of values equal to or less than some specified value, say, x_0, of the normally distributed random variable X is equivalent to the area under the curve and to the left of x_0 as represented by the shaded area in Figure 12.3.1. We obtain the numerical value of this area by converting x_0 to a standard normal deviation by the formula $z_0 = (x_0 - \mu)/\sigma$ and finding the appropriate value in Appendix Table D. We use this procedure to obtain the expected relative frequencies corresponding to each of the class intervals in Table 12.3.1. We estimate μ and σ with $\bar{x}$ and s as computed from the grouped sample data. The first step consists of obtaining z values corresponding to the lower limit of each class interval. The area between two successive z values will give the expected relative frequency of occurrence of values for the corresponding class interval.

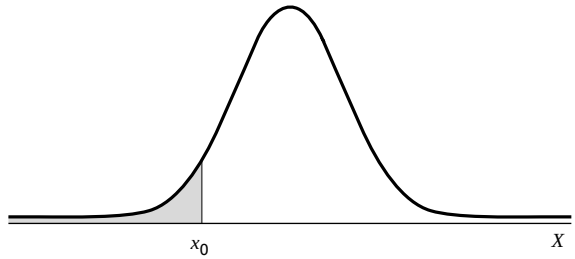


FIGURE 12.3.1 A normal distribution showing the relative frequency of occurrence of values less than or equal to x_0. The shaded area represents the relative frequency of occurrence of values equal to or less than x_0.

For example, to obtain the expected relative frequency of occurrence of values in the interval 100.0 to 124.9 we proceed as follows:

$$\text{The } z \text{ value corresponding to } X = 100.0 \text{ is } z = \frac{100.0 - 198.67}{41.31} = -2.39$$

$$\text{The } z \text{ value corresponding to } X = 125.0 \text{ is } z = \frac{125.0 - 198.67}{41.31} = -1.78$$

In Appendix Table D we find that the area to the left of -2.39 is .0084, and the area to the left of -1.78 is .0375. The area between -1.78 and -2.39 is equal to $.0375 - .0084 = .0291$, which is equal to the expected relative frequency of occurrence of cholesterol levels within the interval 100.0 to 124.9. This tells us that if the null hypothesis is true, that is, if the cholesterol levels are normally distributed, we should expect 2.91 percent of the values in our sample to be between 100.0 and 124.9. When we multiply our total sample size, 47, by .0291 we find the expected frequency for the interval to be 1.4. Similar calculations will give the expected frequencies for the other intervals as shown in Table 12.3.2.

TABLE 12.3.2 Class Intervals and Expected Frequencies for Example 12.3.1

Class Interval	$z(x_i - \bar{x})/s$ At Lower Limit of Interval	Expected Relative Frequency	Expected Frequency	
< 100		.0084	.4	1.8
100.0–124.9	−2.39	.0291	1.4	
125.0–149.9	−1.78	.0815	3.8	
150.0–174.9	−1.18	.1653	7.8	
175.0–199.9	−.57	.2277	10.7	
200.0–224.9	.03	.2269	10.7	
225.0–249.9	.64	.1536	7.2	
250.0–274.9	1.24	.0753	3.5	
275.0–299.9	1.85	.0251	1.2	1.5
300.0 and greater	2.45	.0071	.3	

Comparing Observed and Expected Frequencies

We are now interested in examining the magnitudes of the discrepancies between the observed frequencies and the expected frequencies, since we note that the two sets of frequencies do not agree. We know that even if our sample were drawn from a normal distribution of values, sampling variability alone would make it highly unlikely that the observed and expected frequencies would agree perfectly. We wonder, then, if the discrepancies between the observed and expected frequencies are small enough that we feel it reasonable that they could have occurred by chance alone, when the null hypothesis is true. If they are of this magnitude, we will be unwilling to reject the null hypothesis that the sample came from a normally distributed population.

If the discrepancies are so large that it does not seem reasonable that they could have occurred by chance alone when the null hypothesis is true, we will want to reject the null hypothesis. The criterion against which we judge whether the discrepancies are "large" or "small" is provided by the chi-square distribution.

The observed and expected frequencies along with each value of $(O_i - E_i)^2/E_i$ are shown in Table 12.3.3. The first entry in the last column, for example, is computed from $(1 - 1.8)^2/1.8 = .356$. The other values of $(O_i - E_i)^2/E_i$ are computed in a similar manner.

From Table 12.3.3 we see that $X^2 = \sum[(O_i - E_i)^2/E_i] = 10.566$. The appropriate degrees of freedom are 8 (the number of groups or class intervals) -3 (for the three restrictions: making $\sum E_i = \sum O_i$, and estimating μ and σ from the sample data) $= 5$.

8. **Statistical decision.** When we compare $X^2 = 10.566$ with values of χ^2 in Appendix Table F, we see that it is less than $\chi^2_{.95} = 11.070$, so that, at the .05 level of significance, we cannot reject the null hypothesis that the sample came from a normally distributed population.

TABLE 12.3.3 Observed and Expected Frequencies and $(O_i - E_i)^2/E_i$ for Example 12.3.1

Class Interval	Observed Frequency (O_i)	Expected Frequency (E_i)	$(O_i - E_i)^2/E_i$
< 100	0	.4 } 1.8	.356
100.0–124.9	1	1.4 }	
125.0–149.9	3	3.8	.168
150.0–174.9	8	7.8	.005
175.0–199.9	18	10.7	4.980
200.0–224.9	6	10.7	2.064
225.0–249.9	4	7.2	1.422
250.0–274.9	4	3.5	.071
275.0–299.9	3	1.2 } 1.5	1.500
300.0 and greater	0	.3 }	
Total	47	47	10.566

9. **Conclusion.** We conclude that in the sampled population, cholesterol levels may follow a normal distribution.
10. ***p* value.** Since $11.070 > 10.566 > 9.236$, $.05 < p < .10$. In other words, the probability of obtaining a value of X^2 as large as 10.566, when the null hypothesis is true, is between .05 and .10. Thus we conclude that such an event is not sufficiently rare to reject the null hypothesis that the data come from a normal distribution. ■

Sometimes the parameters are specified in the null hypothesis. It should be noted that had the mean and variance of the population been specified as part of the null hypothesis in Example 12.3.1, we would not have had to estimate them from the sample and our degrees of freedom would have been $8 - 1 = 7$.

Alternatives Although one frequently encounters in the literature the use of chi-square to test for normality, it is not the most appropriate test to use when the hypothesized distribution is continuous. The Kolmogorov–Smirnov test, described in Chapter 13, was especially designed for goodness-of-fit tests involving continuous distributions.

EXAMPLE 12.3.2 *The Binomial Distribution*

In a study designed to determine patient acceptance of a new pain reliever, 100 physicians each selected a sample of 25 patients to participate in the study. Each patient, after trying the new pain reliever for a specified period of time, was asked whether it was preferable to the pain reliever used regularly in the past.

The results of the study are shown in Table 12.3.4.

TABLE 12.3.4 Results of Study Described in Example 12.3.2

Number of Patients Out of 25 Preferring New Pain Reliever	Number of Doctors Reporting this Number	Total Number of Patients Preferring New Pain Reliever by Doctor
0	5	0
1	6	6
2	8	16
3	10	30
4	10	40
5	15	75
6	17	102
7	10	70
8	10	80
9	9	81
10 or more	0	0
Total	100	500

We are interested in determining whether or not these data are compatible with the hypothesis that they were drawn from a population that follows a binomial distribution. Again, we employ a chi-square goodness-of-fit test.

Solution: Since the binomial parameter, p, is not specified, it must be estimated from the sample data. A total of 500 patients out of the 2500 patients participating in the study said they preferred the new pain reliever, so that our point estimate of p is $\hat{p} = 500/2500 = .20$. The expected relative frequencies can be obtained by evaluating the binomial function

$$f(x) = {}_{25}C_x(.2)^x(.8)^{25-x}$$

for $x = 0, 1, \ldots, 25$. For example, to find the probability that out of a sample of 25 patients none would prefer the new pain reliever, when in the total population the true proportion preferring the new pain reliever is .2, we would evaluate

$$f(0) = {}_{25}C_0(.2)^0(.8)^{25-0}$$

This can be done most easily by consulting Appendix Table B, where we see that $P(X = 0) = .0038$. The relative frequency of occurrence of samples of size 25 in which no patients prefer the new pain reliever is .0038. To obtain the corresponding expected frequency, we multiply .0038 by 100 to get .38. Similar calculations yield the remaining expected frequencies, which, along with the observed frequencies, are shown in Table 12.3.5. We see in this table

TABLE 12.3.5 Calculations for Example 12.3.2

Number of Patients Out of 25 Preferring New Pain Reliever	Number of Doctors Reporting This Number (Observed Frequency, O_i)	Expected Relative Frequency	Expected Frequency E_i
0	5 } 11	.0038	.38 } 2.74
1	6 }	.0236	2.36 }
2	8	.0708	7.08
3	10	.1358	13.58
4	10	.1867	18.67
5	15	.1960	19.60
6	17	.1633	16.33
7	10	.1109	11.09
8	10	.0623	6.23
9	9	.0295	2.95
10 or more	0	.0173	1.73
Total	100	1.0000	100.00

that the first expected frequency is less than 1, so that we follow Cochran's suggestion and combine this group with the second group. When we do this, all the expected frequencies are greater than 1.

From the data, we compute

$$X^2 = \frac{(11 - 2.74)^2}{2.74} + \frac{(8 - 7.08)^2}{7.08} + \cdots + \frac{(0 - 1.73)^2}{1.73} = 47.624$$

The appropriate degrees of freedom are 10 (the number of groups left after combining the first two) less 2, or 8. One degree of freedom is lost because we force the total of the expected frequencies to equal the total observed frequencies, and one degree of freedom is sacrificed because we estimated p from the sample data.

We compare our computed X^2 with the tabulated χ^2 with 8 degrees of freedom and find that it is significant at the .005 level of significance; that is, $p < .005$. We reject the null hypothesis that the data came from a binomial distribution. ■

EXAMPLE 12.3.3 *The Poisson Distribution*

A hospital administrator wishes to test the null hypothesis that emergency admissions follow a Poisson distribution with $\lambda = 3$. Suppose that over a period of 90 days the numbers of emergency admissions were as shown in Table 12.3.6.

TABLE 12.3.6 Number of Emergency Admissions to a Hospital During a 90-Day Period

Day	Emergency Admissions	Day	Emergency Admissions	Day	Emergency Admissions	Day	Emergency Admissions
1	2	24	5	47	4	70	3
2	3	25	3	48	2	71	5
3	4	26	2	49	2	72	4
4	5	27	4	50	3	73	1
5	3	28	4	51	4	74	1
6	2	29	3	52	2	75	6
7	3	30	5	53	3	76	3
8	0	31	1	54	1	77	3
9	1	32	3	55	2	78	5
10	0	33	2	56	3	79	2
11	1	34	4	57	2	80	1
12	0	35	2	58	5	81	7
13	6	36	5	59	2	82	7
14	4	37	0	60	7	83	1
15	4	38	6	61	8	84	5
16	4	39	4	62	3	85	1

(*Continued*)

Day	Emergency Admissions	Day	Emergency Admissions	Day	Emergency Admissions	Day	Emergency Admissions
17	3	40	4	63	1	86	4
18	4	41	5	64	3	87	4
19	3	42	1	65	1	88	9
20	3	43	3	66	0	89	2
21	3	44	1	67	3	90	3
22	4	45	2	68	2		
23	3	46	3	69	1		

The data of Table 12.3.6 are summarized in Table 12.3.7.

Solution: To obtain the expected frequencies we first obtain the expected relative frequencies by evaluating the Poisson function given by Equation 4.4.1 for each entry in the left-hand column of Table 12.3.7. For example, the first expected relative frequency is obtained by evaluating

$$f(0) = \frac{e^{-3}3^0}{0!}$$

We may use Appendix Table C to find this and all the other expected relative frequencies that we need. Each of the expected relative frequencies

TABLE 12.3.7 Summary of Data Presented in Table 12.3.6

Number of Emergency Admissions in a Day	Number of Days This Number of Emergency Admissions Occurred
0	5
1	14
2	15
3	23
4	16
5	9
6	3
7	3
8	1
9	1
10 or more	0
Total	90

TABLE 12.3.8 Observed and Expected Frequencies and Components of X^2 for Example 12.3.3

Number of Emergency Admissions	Number of Days this Number Occurred, O_i	Expected Relative Frequency	Expected Frequency	$\frac{(O_i - E_i)^2}{E_i}$
0	5	.050	4.50	.056
1	14	.149	13.41	.026
2	15	.224	20.16	1.321
3	23	.224	20.16	.400
4	16	.168	15.12	.051
5	9	.101	9.09	.001
6	3	.050	4.50	.500
7	3	.022	1.98	.525
8	1	.008	.72	
9	1 } 2	.003	.27 } 1.08	.784
10 or more	0	.001	.09	
Total	90	1.000	90.00	3.664

is multiplied by 90 to obtain the corresponding expected frequencies. These values along with the observed and expected frequencies and the components of X^2, $(O_i - E_i)^2/E_i$, are displayed in Table 12.3.8, in which we see that

$$X^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{(5 - 4.50)^2}{4.50} + \cdots + \frac{(2 - 1.08)^2}{1.08} = 3.664$$

We also note that the last three expected frequencies are less than 1, so that they must be combined to avoid having any expected frequencies less than 1. This means that we have only nine effective categories for computing degrees of freedom. Since the parameter, λ, was specified in the null hypothesis, we do not lose a degree of freedom for reasons of estimation, so that the appropriate degrees of freedom are $9 - 1 = 8$. By consulting Appendix Table F, we find that the critical value of χ^2 for 8 degrees of freedom and $\alpha = .05$ is 15.507, so that we cannot reject the null hypothesis at the .05 level, or for that matter any reasonable level, of significance ($p > .10$). We conclude, therefore, that emergency admissions at this hospital may follow a Poisson distribution with $\lambda = 3$. At least the observed data do not cast any doubt on that hypothesis.

If the parameter λ has to be estimated from sample data, the estimate is obtained by multiplying each value x by its frequency, summing these products, and dividing the total by the sum of the frequencies. ■

EXAMPLE 12.3.4 *The Uniform Distribution*

The flu season in southern Nevada for 2005–2006 ran from December to April, the coldest months of the year. The Southern Nevada Health District reported the numbers of vaccine-preventable influenza cases shown in Table 12.3.9. We are interested in knowing whether the numbers of flu cases in the district are equally distributed among the five flu season months. That is, we wish to know if flu cases follow a uniform distribution.

Solution:

1. **Data.** See Table 12.3.9.
2. **Assumptions.** We assume that the reported cases of flu constitute a simple random sample of cases of flu that occurred in the district.
3. **Hypotheses.**
   H_0: Flu cases in southern Nevada are uniformly distributed over the five flu season months.
   H_A: Flu cases in southern Nevada are not uniformly distributed over the five flu season months.
   Let $\alpha = .01$.
4. **Test statistic.** The test statistic is

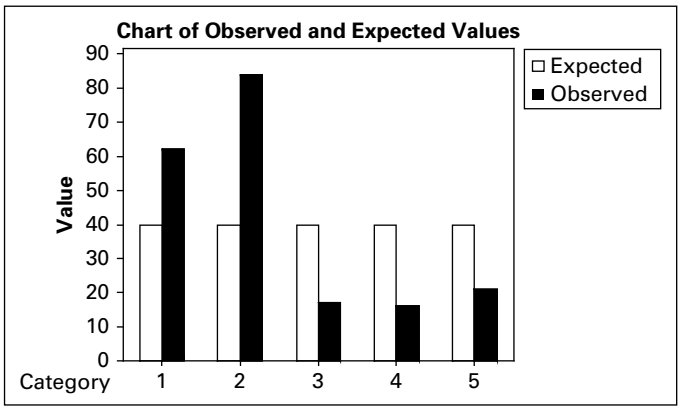
$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

5. **Distribution of test statistic.** If H_0 is true, X^2 is distributed approximately as χ^2 with $(5 - 1) = 4$ degrees of freedom.
6. **Decision rule.** Reject H_0 if the computed value of X^2 is equal to or greater than 13.277.

TABLE 12.3.9 Reported Vaccine-Preventable Influenza Cases from Southern Nevada, December 2005–April 2006

Month	Number of Reported Cases of Influenza
December 2005	62
January 2006	84
February 2006	17
March 2006	16
April 2006	21
Total	200

Source: http://www.southernnevadahealthdistrict.org/epidemiology/disease_statistics.htm.



Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: C1

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
1	62	0.2	40	12.100
2	84	0.2	40	48.400
3	17	0.2	40	13.225
4	16	0.2	40	14.400
5	21	0.2	40	9.025

N	DF	Chi-Sq	P-Value
200	4	97.15	0.000

FIGURE 12.3.2 MINITAB output for Example 12.3.4.

7. **Calculation of test statistic.** If the null hypothesis is true, we would expect to observe $200/5 = 40$ cases per month. Figure 12.3.2 shows the computer printout obtained from MINITAB. The bar graph shows the observed and expected frequencies per month. The chi-square table provides the observed frequencies, the expected frequencies based on a uniform distribution, and the individual chi-square contribution for each test value.
8. **Statistical decision.** Since 97.15, the computed value of X^2, is greater than 13.277, we reject, based on these data, the null hypothesis of a

uniform distribution of flu cases during the flu season in southern Nevada.

9. **Conclusion.** We conclude that the occurrence of flu cases does not follow a uniform distribution.
10. ***p* value.** From the MINITAB output we see that $p = .000$ (i.e., $< .001$). ■

EXAMPLE 12.3.5

A certain human trait is thought to be inherited according to the ratio 1:2:1 for homozygous dominant, heterozygous, and homozygous recessive. An examination of a simple random sample of 200 individuals yielded the following distribution of the trait: dominant, 43; heterozygous, 125; and recessive, 32. We wish to know if these data provide sufficient evidence to cast doubt on the belief about the distribution of the trait.

Solution:

1. **Data.** See statement of the example.
2. **Assumptions.** We assume that the data meet the requirements for the application of the chi-square goodness-of-fit test.
3. **Hypotheses.**
   H_0: The trait is distributed according to the ratio 1:2:1 for homozygous dominant, heterozygous, and homozygous recessive.
   H_A: The trait is not distributed according to the ratio 1:2:1.
4. **Test statistic.** The test statistic is

$$X^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

5. **Distribution of test statistic.** If H_0 is true, X^2 is distributed as chi-square with 2 degrees of freedom.
6. **Decision rule.** Suppose we let the probability of committing a type I error be .05. Reject H_0 if the computed value of X^2 is equal to or greater than 5.991.
7. **Calculation of test statistic.** If H_0 is true, the expected frequencies for the three manifestations of the trait are 50, 100, and 50 for dominant, heterozygous, and recessive, respectively. Consequently,

$$X^2 = (43 - 50)^2/50 + (125 - 100)2/100 + (32 - 50)^2/50 = 13.71$$

8. **Statistical decision.** Since $13.71 > 5.991$, we reject H_0.
9. **Conclusion.** We conclude that the trait is not distributed according to the ratio 1:2:1.
10. ***p* value.** Since $13.71 > 10.597$, the p value for the test is $p < .005$. ■

EXERCISES

12.3.1 The following table shows the distribution of uric acid determinations taken on 250 patients. Test the goodness-of-fit of these data to a normal distribution with $\mu = 5.74$ and $\sigma = 2.01$. Let $\alpha = .01$.

Uric Acid Determination	Observed Frequency	Uric Acid Determination	Observed Frequency
< 1	1	6 to 6.99	45
1 to 1.99	5	7 to 7.99	30
2 to 2.99	15	8 to 8.99	22
3 to 3.99	24	9 to 9.99	10
4 to 4.99	43	10 or higher	5
5 to 5.99	50		
Total			250

12.3.2 The following data were collected on 300 eight-year-old girls. Test, at the .05 level of significance, the null hypothesis that the data are drawn from a normally distributed population. The sample mean and standard deviation computed from grouped data are 127.02 and 5.08.

Height in Centimeters	Observed Frequency	Height in Centimeters	Observed Frequency
114 to 115.9	5	128 to 129.9	43
116 to 117.9	10	130 to 131.9	42
118 to 119.9	14	132 to 133.9	30
120 to 121.9	21	134 to 135.9	11
122 to 123.9	30	136 to 137.9	5
124 to 125.9	40	138 to 139.9	4
126 to 127.9	45		
Total			300

12.3.3 The face sheet of patients' records maintained in a local health department contains 10 entries. A sample of 100 records revealed the following distribution of erroneous entries:

Number of Erroneous Entries Out of 10	Number of Records
0	8
1	25
2	32
3	24
4	10
5 or more	1
Total	100

Test the goodness-of-fit of these data to the binomial distribution with $p = .20$. Find the p value for this test.

12.3.4 In a study conducted by Byers et al. (A-2), researchers tested a Poisson model for the distribution of activities of daily living (ADL) scores after a 7-month prehabilitation program designed to prevent functional decline among physically frail, community-living older persons. ADL measured the ability of individuals to perform essential tasks, including walking inside the house, bathing, upper and lower body dressing, transferring from a chair, toileting, feeding, and grooming. The scoring method used in this study assigned a value of 0 for no (personal) help and no difficulty, 1 for difficulty but no help, and 2 for help regardless of difficulty. Scores were summed to produce an overall score ranging from 0 to 16 (for eight tasks). There were 181 subjects who completed the study. Suppose we use the authors' scoring method to assess the status of another group of 181 subjects relative to their activities of daily living. Let us assume that the following results were obtained.

X	Observed Frequency X	Expected Frequency	X	Observed Frequency X	Expected Frequency
0	74	11.01	7	4	2.95
1	27	30.82	8	3	1.03
2	14	43.15	9	2	0.32
3	14	40.27	10	3	0.09
4	11	28.19	11	4	0.02
5	7	15.79	12 or more	13	0.01
6	5	7.37			

Source: Hypothetical data based on procedure reported by Amy L. Byers, Heather Allore, Thomas M. Gill, and Peter N. Peduzzi, "Application of Negative Binomial Modeling for Discrete Outcomes: A Case Study in Aging Research," *Journal of Clinical Epidemiology, 56* (2003), 559–564.

Test the null hypothesis that these data were drawn from a Poisson distribution with $\lambda = 2.8$. Let $\alpha = .01$.

12.3.5 The following are the numbers of a particular organism found in 100 samples of water from a pond:

Number of Organisms per Sample	Frequency	Number of Organisms per Sample	Frequency
0	15	4	5
1	30	5	4
2	25	6	1
3	20	7	0
Total			100

Test the null hypothesis that these data were drawn from a Poisson distribution. Determine the p value for this test.

12.3.6 A research team conducted a survey in which the subjects were adult smokers. Each subject in a sample of 200 was asked to indicate the extent to which he or she agreed with the statement: "I would like to quit smoking." The results were as follows:

Response:	Strongly agree	Agree	Disagree	Strongly Disagree
Number Responding:	102	30	60	8

Can one conclude on the basis of these data that, in the sampled population, opinions are not equally distributed over the four levels of agreement? Let the probability of committing a type I error be .05 and find the p value.

12.4 TESTS OF INDEPENDENCE

Another, and perhaps the most frequent, use of the chi-square distribution is to test the null hypothesis that two criteria of classification, when applied to the same set of entities, are independent. We say that two criteria of classification are independent if the distribution of one criterion is the same no matter what the distribution of the other criterion. For example, if socioeconomic status and area of residence of the inhabitants of a certain city are independent, we would expect to find the same proportion of families in the low, medium, and high socioeconomic groups in all areas of the city.

The Contingency Table The classification, according to two criteria, of a set of entities, say, people, can be shown by a table in which the r rows represent the various levels of one criterion of classification and the c columns represent the various levels of the second criterion. Such a table is generally called a *contingency table*, with dimension $r \times c$. The classification according to two criteria of a finite population of entities is shown in Table 12.4.1.

We will be interested in testing the null hypothesis that in the population the two criteria of classification are independent. If the hypothesis is rejected, we will conclude that

TABLE 12.4.1 Two-Way Classification of a Finite Population of Entities

Second Criterion of Classification Level	**First Criterion of Classification Level**					
	1	**2**	**3**	. . .	***c***	**Total**
1	N_{11}	N_{12}	N_{13}	. . .	N_{1c}	$N_{1.}$
2	N_{21}	N_{22}	N_{23}	. . .	N_{2c}	$N_{2.}$
3	N_{31}	N_{32}	N_{33}	. . .	N_{3c}	$N_{3.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	N_{r1}	N_{r2}	N_{r3}	. . .	N_{rc}	$N_{r.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	. . .	$N_{.c}$	N

TABLE 12.4.2 Two-Way Classification of a Sample of Entities

Second Criterion of Classification Level	First Criterion of Classification Level					
	1	2	3	. . .	c	Total
1	n_{11}	n_{12}	n_{13}	. . .	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	. . .	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	. . .	n_{3c}	$n_{3.}$
⋮	⋮	⋮	⋮		⋮	⋮
r	n_{r1}	n_{r2}	n_{r3}	. . .	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	. . .	$n_{.c}$	n

the two criteria of classification are not independent. A sample of size n will be drawn from the population of entities, and the frequency of occurrence of entities in the sample corresponding to the cells formed by the intersections of the rows and columns of Table 12.4.1 along with the marginal totals will be displayed in a table such as Table 12.4.2.

Calculating the Expected Frequencies The expected frequency, under the null hypothesis that the two criteria of classification are independent, is calculated for each cell.

We learned in Chapter 3 (see Equation 3.4.4) that if two events are independent, the probability of their joint occurrence is equal to the product of their individual probabilities. Under the assumption of independence, for example, we compute the probability that one of the n subjects represented in Table 12.4.2 will be counted in Row 1 and Column 1 of the table (that is, in Cell 11) by multiplying the probability that the subject will be counted in Row 1 by the probability that the subject will be counted in Column 1. In the notation of the table, the desired calculation is

$$\left(\frac{n_{1.}}{n}\right)\left(\frac{n_{.1}}{n}\right)$$

To obtain the expected frequency for Cell 11, we multiply this probability by the total number of subjects, n. That is, the expected frequency for Cell 11 is given by

$$\left(\frac{n_{1.}}{n}\right)\left(\frac{n_{.1}}{n}\right)(n)$$

Since the n in one of the denominators cancels into numerator n, this expression reduces to

$$\frac{(n_{1.})(n_{.1})}{n}$$

In general, then, we see that to obtain the expected frequency for a given cell, we multiply the total of the row in which the cell is located by the total of the column in which the cell is located and divide the product by the grand total.

Observed Versus Expected Frequencies The expected frequencies and observed frequencies are compared. If the discrepancy is sufficiently small, the null hypothesis is tenable. If the discrepancy is sufficiently large, the null hypothesis is rejected, and we conclude that the two criteria of classification are not independent. The decision as to whether the discrepancy between observed and expected frequencies is sufficiently large to cause rejection of H_0 will be made on the basis of the size of the quantity computed when we use Equation 12.2.4, where O_i and E_i refer, respectively, to the observed and expected frequencies in the cells of Table 12.4.2. It would be more logical to designate the observed and expected frequencies in these cells by O_{ij} and E_{ij}, but to keep the notation simple and to avoid the introduction of another formula, we have elected to use the simpler notation. It will be helpful to think of the cells as being numbered from 1 to k, where 1 refers to Cell 11 and k refers to Cell rc. It can be shown that X^2 as defined in this manner is distributed approximately as χ^2 with $(r-1)(c-1)$ degrees of freedom when the null hypothesis is true. If the computed value of X^2 is equal to or larger than the tabulated value of χ^2 for some α, the null hypothesis is rejected at the α level of significance. The hypothesis testing procedure is illustrated with the following example.

EXAMPLE 12.4.1

In 1992, the U.S. Public Health Service and the Centers for Disease Control and Prevention recommended that all women of childbearing age consume 400 μg of folic acid daily to reduce the risk of having a pregnancy that is affected by a neural tube defect such as spina bifida or anencephaly. In a study by Stepanuk et al. (A-3), 693 pregnant women called a teratology information service about their use of folic acid supplementation. The researchers wished to determine if preconceptional use of folic acid and race are independent. The data appear in Table 12.4.3.

Solution:

1. **Data.** See Table 12.4.3.
2. **Assumptions.** We assume that the sample available for analysis is equivalent to a simple random sample drawn from the population of interest.

TABLE 12.4.3 Race of Pregnant Caller and Use of Folic Acid

	Preconceptional Use of Folic Acid		
	Yes	No	Total
White	260	299	559
Black	15	41	56
Other	7	14	21
Total	282	354	636

Source: Kathleen M. Stepanuk, Jorge E. Tolosa, Dawneete Lewis, Victoria Meyers, Cynthia Royds, Juan Carlos Saogal, and Ron Librizzi, "Folic Acid Supplementation Use Among Women Who Contact a Teratology Information Service," *American Journal of Obstetrics and Gynecology, 187* (2002), 964–967.

3. **Hypotheses.**
   H_0: Race and preconceptional use of folic acid are independent.
   H_A: The two variables are not independent.
   Let $\alpha = .05$.

4. **Test statistic.** The test statistic is

$$X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

5. **Distribution of test statistic.** When H_0 is true, X^2 is distributed approximately as χ^2 with $(r-1)(c-1) = (3-1)(2-1) = (2)(1) = 2$ degrees of freedom.

6. **Decision rule.** Reject H_0 if the computed value of X^2 is equal to or greater than 5.991.

7. **Calculation of test statistic.** The expected frequency for the first cell is $(559 \times 282)/636 = 247.86$. The other expected frequencies are calculated in a similar manner. Observed and expected frequencies are displayed in Table 12.4.4. From the observed and expected frequencies we may compute

$$\begin{aligned} X^2 &= \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] \\ &= \frac{(260 - 247.86)^2}{247.86} + \frac{(299 - 311.14)^2}{311.14} + \ldots + \frac{(14 - 11.69)^2}{11.69} \\ &= .59461 + .47368 + \ldots + .45647 = 9.08960 \end{aligned}$$

8. **Statistical decision.** We reject H_0 since $9.08960 > 5.991$.

9. **Conclusion.** We conclude that H_0 is false, and that there is a relationship between race and preconceptional use of folic acid.

10. ***p* value.** Since $7.378 < 9.08960 < 9.210$, $.01 < p < .025$.

TABLE 12.4.4 Observed and Expected Frequencies for Example 12.4.1

	Preconceptional Use of Folic Acid		
	Yes	**No**	**Total**
White	260 (247.86)	299 (311.14)	559
Black	15 (24.83)	41 (31.17)	56
Other	7 (9.31)	14 (11.69)	21
Total	282	354	636

■

Computer Analysis The computer may be used to advantage in calculating X^2 for tests of independence and tests of homogeneity. Figure 12.4.1 shows the procedure and printout for Example 12.4.1 when the MINITAB program for computing X^2 from contingency tables is used. The data were entered into MINITAB Columns 1 and 2, corresponding to the columns of Table 12.4.3.

We may use SAS® to obtain an analysis and printout of contingency table data by using the PROC FREQ statement. Figure 12.4.2 shows a partial SAS® printout reflecting the analysis of the data of Example 12.4.1.

Data:

```
C1: 260 15  7
C2: 299 41 14
```

Dialog Box: **Session command:**

Stat ➤ Tables ➤ Chi-square Test `MTB > CHISQUARE C1-C3`

Type *C1-C2* in Columns containing the table.
Click **OK.**

Output:

Chi-Square Test: C1, C2

```
Expected counts are printed below observed counts

            C1        C2    Total
    1      260       299      559
         247.86    311.14

    2       15        41       56
          24.83     31.17

    3        7        14       21
           9.31     11.69

Total      282       354      636

Chi-Sq  =  0.595 + 0.474 +
           3.892 + 3.100 +
           0.574 + 0.457 = 9.091
DF = 2, P-Value = 0.011
```

FIGURE 12.4.1 MINITAB procedure and output for chi-square analysis of data in Table 12.4.3.

The SAS System

The FREQ Procedure

Table of race by folic

race Frequency Percent Row Pct Col Pct	folic No	Yes	Total
Black	41 6.45 73.21 11.58	15 2.36 26.79 5.32	56 8.81
Other	14 2.20 66.67 3.95	7 1.10 33.33 2.48	21 3.30
White	299 47.01 53.49 84.46	260 40.88 46.51 92.20	559 87.89
Total	354 55.66	282 44.34	636 100.00

Statistics for Table of race by folic

Statistic	DF	Value	Prob
Chi-Square	2	9.0913	0.0106
Likelihood Ratio Chi-Square	2	9.4808	0.0087
Mantel–Haenszel Chi-Square	1	8.9923	0.0027
Phi Coefficient		0.1196	
Contingency Coefficient		0.1187	
Cramer's V		0.1196	

Sample Size = 636

FIGURE 12.4.2 Partial SAS® printout for the chi-square analysis of the data from Example 12.4.1.

Note that the SAS® printout shows, in each cell, the percentage that cell frequency is of its row total, its column total, and the grand total. Also shown, for each row and column total, is the percentage that the total is of the grand total. In addition to the X^2 statistic, SAS® gives the value of several other statistics that may be computed from contingency table data. One of these, the Mantel–Haenszel chi-square statistic, will be discussed in a later section of this chapter.

Small Expected Frequencies The problem of small expected frequencies discussed in the previous section may be encountered when analyzing the data of contingency tables. Although there is a lack of consensus on how to handle this problem, many authors currently follow the rule given by Cochran (5). He suggests that for contingency tables with more than 1 degree of freedom a minimum expectation of 1 is allowable if no more than 20 percent of the cells have expected frequencies of less than 5. To meet this rule, adjacent rows and/or adjacent columns may be combined when to do so is logical in light of other considerations. If X^2 is based on less than 30 degrees of freedom, expected frequencies as small as 2 can be tolerated. We did not experience the problem of small expected frequencies in Example 12.4.1, since they were all greater than 5.

The 2×2 Contingency Table Sometimes each of two criteria of classification may be broken down into only two categories, or levels. When data are cross-classified in this manner, the result is a contingency table consisting of two rows and two columns. Such a table is commonly referred to as a 2×2 table. The value of X^2 may be computed by first calculating the expected cell frequencies in the manner discussed above. In the case of a 2×2 contingency table, however, X^2 may be calculated by the following shortcut formula:

$$X^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)} \tag{12.4.1}$$

where a, b, c, and d are the observed cell frequencies as shown in Table 12.4.5. When we apply the $(r - 1)(c - 1)$ rule for finding degrees of freedom to a 2×2 table, the result is 1 degree of freedom. Let us illustrate this with an example.

TABLE 12.4.5 A 2×2 Contingency Table

Second Criterion of Classification	First Criterion of Classification		
	1	2	Total
1	a	b	$a + b$
2	c	d	$c + d$
Total	$a + c$	$b + d$	n

EXAMPLE 12.4.2

According to Silver and Aiello (A-4), falls are of major concern among polio survivors. Researchers wanted to determine the impact of a fall on lifestyle changes. Table 12.4.6 shows the results of a study of 233 polio survivors on whether fear of falling resulted in lifestyle changes.

Solution:

1. **Data.** From the information given we may construct the 2×2 contingency table displayed as Table 12.5.6.
2. **Assumptions.** We assume that the sample is equivalent to a simple random sample.
3. **Hypotheses.**

   H_0: Fall status and lifestyle change because of fear of falling are independent.

   H_1: The two variables are not independent.

   Let $\alpha = .05$.
4. **Test statistic.** The test statistic is

$$X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

5. **Distribution of test statistic.** When H_0 is true, X^2 is distributed approximately as χ^2 with $(r-1)(c-1) = (2-1)(2-1) = (1)(1) = 1$ degree of freedom.
6. **Decision rule.** Reject H_0 if the computed value of X^2 is equal to or greater than 3.841.
7. **Calculation of test statistic.** By Equation 12.4.1 we compute

$$X^2 = \frac{233[(131)(36) - (52)(14)]^2}{(145)(88)(183)(50)} = 31.7391$$

8. **Statistical decision.** We reject H_0 since $31.7391 > 3.841$.

TABLE 12.4.6 Contingency Table for the Data of Example 12.4.2

	Made Lifestyle Changes Because of Fear of Falling		
	Yes	**No**	**Total**
Fallers	131	52	183
Nonfallers	14	36	50
Total	145	88	233

Source: J. K. Silver and D. D. Aiello, "Polio Survivors: Falls and Subsequent Injuries," *American Journal of Physical Medicine and Rehabilitation, 81* (2002), 567–570.

9. **Conclusion.** We conclude that H_0 is false, and that there is a relationship between experiencing a fall and changing one's lifestyle because of fear of falling.
10. ***p* value.** Since $31.7391 > 7.879$, $p < .005$. ■

Small Expected Frequencies The problems of how to handle small expected frequencies and small total sample sizes may arise in the analysis of 2×2 contingency tables. Cochran (5) suggests that the χ^2 test should not be used if $n < 20$ or if $20 < n < 40$ and any expected frequency is less than 5. When $n = 40$, an expected cell frequency as small as 1 can be tolerated.

Yates's Correction The observed frequencies in a contingency table are discrete and thereby give rise to a discrete statistic, X^2, which is approximated by the χ^2 distribution, which is continuous. Yates (6) in 1934 proposed a procedure for correcting for this in the case of 2×2 tables. The correction, as shown in Equation 12.4.2, consists of subtracting half the total number of observations from the absolute value of the quantity $ad - bc$ before squaring. That is,

$$X^2_{\text{corrected}} = \frac{n(|ad - bc| - .5n)^2}{(a + c)(b + d)(a + b)(c + d)} \tag{12.4.2}$$

It is generally agreed that no correction is necessary for larger contingency tables. Although Yates's correction for 2×2 tables has been used extensively in the past, more recent investigators have questioned its use. As a result, some practitioners recommend against its use.

We may, as a matter of interest, apply the correction to our current example. Using Equation 12.4.2 and the data from Table 12.4.6, we may compute

$$X^2 = \frac{233[|(131)(36) - (52)(14)| - .5(233)]^2}{(145)(88)(183)(50)} = 29.9118$$

As might be expected, with a sample this large, the difference in the two results is not dramatic.

Tests of Independence: Characteristics The characteristics of a chi-square test of independence that distinguish it from other chi-square tests are as follows:

1. A single sample is selected from a population of interest, and the subjects or objects are cross-classified on the basis of the two variables of interest.
2. The rationale for calculating expected cell frequencies is based on the probability law, which states that if two events (here the two criteria of classification) are independent, the probability of their joint occurrence is equal to the product of their individual probabilities.
3. The hypotheses and conclusions are stated in terms of the independence (or lack of independence) of two variables.

EXERCISES

In the exercises that follow perform the test at the indicated level of significance and determine the p value.

12.4.1 In the study by Silver and Aiello (A-4) cited in Example 12.4.2, a secondary objective was to determine if the frequency of falls was independent of wheelchair use. The following table gives the data for falls and wheelchair use among the subjects of the study.

	Wheelchair Use	
	Yes	**No**
Fallers	62	121
Nonfallers	18	32

Source: J. K. Silver and D. D. Aiello, "Polio Survivors: Falls and Subsequent Injuries," *American Journal of Physical Medicine and Rehabilitation*, *81* (2002), 567–570.

Do these data provide sufficient evidence to warrant the conclusion that wheelchair use and falling are related? Let $\alpha = .05$.

12.4.2 Sternal surgical site infection (SSI) after coronary artery bypass graft surgery is a complication that increases patient morbidity and costs for patients, payers, and the health care system. Segal and Anderson (A-5) performed a study that examined two types of preoperative skin preparation before performing open heart surgery. These two preparations used aqueous iodine and insoluble iodine with the following results.

	Comparison of Aqueous and Insoluble Preps	
Prep Group	**Infected**	**Not Infected**
Aqueous iodine	14	94
Insoluble iodine	4	97

Source: Cynthia G. Segal and Jacqueline J. Anderson, "Preoperative Skin Preparation of Cardiac Patients," *AORN Journal, 76* (2002), 821–827.

Do these data provide sufficient evidence at the $\alpha = .05$ level to justify the conclusion that the type of skin preparation and infection are related?

12.4.3 The side effects of nonsteroidal antiinflammatory drugs (NSAIDs) include problems involving peptic ulceration, renal function, and liver disease. In 1996, the American College of Rheumatology issued and disseminated guidelines recommending baseline tests (CBC, hepatic panel, and renal tests) when prescribing NSAIDs. A study was conducted by Rothenberg and Holcomb (A-6) to determine if physicians taking part in a national database of computerized medical records performed the recommended baseline tests when prescribing NSAIDs. The researchers classified physicians in the study into four categories—those practicing in internal medicine, family practice, academic family practice, and multispeciality groups. The data appear in the following table.

Practice Type	Performed Baseline Tests	
	Yes	No
Internal medicine	294	921
Family practice	98	2862
Academic family practice	50	3064
Multispecialty groups	203	2652

Source: Ralph Tothenberg and John P. Holcomb, "Guidelines for Monitoring of NSAIDs: Who Listened?," *Journal of Clinical Rheumatology, 6* (2000), 258–265.

Do the data above provide sufficient evidence for us to conclude that type of practice and performance of baseline tests are related? Use $\alpha = .01$.

12.4.4 Boles and Johnson (A-7) examined the beliefs held by adolescents regarding smoking and weight. Respondents characterized their weight into three categories: underweight, overweight, or appropriate. Smoking status was categorized according to the answer to the question, "Do you currently smoke, meaning one or more cigarettes per day?" The following table shows the results of a telephone study of adolescents in the age group 12–17.

	Smoking	
	Yes	No
Underweight	17	97
Overweight	25	142
Appropriate	96	816

Source: Sharon M. Boles and Patrick B. Johnson, "Gender, Weight Concerns, and Adolescent Smoking," *Journal of Addictive Diseases, 20* (2001), 5–14.

Do the data provide sufficient evidence to suggest that weight perception and smoking status are related in adolescents? $\alpha = .05$.

12.4.5 A sample of 500 college students participated in a study designed to evaluate the level of college students' knowledge of a certain group of common diseases. The following table shows the students classified by major field of study and level of knowledge of the group of diseases:

Major	Knowledge of Diseases		
	Good	Poor	Total
Premedical	31	91	122
Other	19	359	378
Total	50	450	500

Do these data suggest that there is a relationship between knowledge of the group of diseases and major field of study of the college students from which the present sample was drawn? Let $\alpha = .05$.

12.4.6 The following table shows the results of a survey in which the subjects were a sample of 300 adults residing in a certain metropolitan area. Each subject was asked to indicate which of three policies they favored with respect to smoking in public places.

Highest Education Level	Policy Favored: No Restrictions on Smoking	Policy Favored: Smoking Allowed in Designated Areas Only	Policy Favored: No Smoking at All	No Opinion	Total
College graduate	5	44	23	3	75
High-school graduate	15	100	30	5	150
Grade-school graduate	15	40	10	10	75
Total	35	184	63	18	300

Can one conclude from these data that, in the sampled population, there is a relationship between level of education and attitude toward smoking in public places? Let $\alpha = .05$.

12.5 TESTS OF HOMOGENEITY

A characteristic of the examples and exercises presented in the last section is that, in each case, the total sample was assumed to have been drawn before the entities were classified according to the two criteria of classification. That is, the observed number of entities falling into each cell was determined after the sample was drawn. As a result, the row and column totals are chance quantities not under the control of the investigator. We think of the sample drawn under these conditions as a single sample drawn from a single population. On occasion, however, either row or column totals may be under the control of the investigator; that is, the investigator may specify that independent samples be drawn from each of several populations. In this case, one set of marginal totals is said to be *fixed*, while the other set, corresponding to the criterion of classification applied to the samples, is *random*. The former procedure, as we have seen, leads to a chi-square test of independence. The latter situation leads to a chi-square *test of homogeneity*. The two situations not only involve different sampling procedures; they lead to different questions and null hypotheses. The test of independence is concerned with the question: Are the two criteria of classification independent? The homogeneity test is concerned with the question: Are the samples drawn from populations that are homogeneous with respect to some criterion of classification? In the latter case the null hypothesis states that the samples are drawn from the same population. Despite these differences in concept and sampling procedure, the two tests are mathematically identical, as we see when we consider the following example.

Calculating Expected Frequencies Either the row categories or the column categories may represent the different populations from which the samples are drawn. If, for example, three populations are sampled, they may be designated as populations 1, 2, and 3, in which case these labels may serve as either row or column headings. If the variable of interest has three categories, say, A, B, and C, these labels may serve as headings for rows or columns, whichever is not used for the populations. If we use notation similar to that adopted for Table 12.4.2, the contingency table for this situation, with columns used to represent the populations, is shown as Table 12.5.1. Before computing our test statistic we need expected frequencies for each of the cells in Table 12.5.1. If the populations are indeed

TABLE 12.5.1 A Contingency Table for Data for a Chi-Square Test of Homogeneity

	Population			
Variable Category	**1**	**2**	**3**	**Total**
A	n_{A1}	n_{A2}	n_{A3}	$n_{A.}$
B	n_{B1}	n_{B2}	n_{B3}	$n_{B.}$
C	n_{C1}	n_{C2}	n_{C3}	$n_{C.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

homogeneous, or, equivalently, if the samples are all drawn from the same population, with respect to the categories A, B, and C, our best estimate of the proportion in the combined population who belong to category A is $n_{A.}/n$. By the same token, if the three populations are homogeneous, we interpret this probability as applying to each of the populations individually. For example, under the null hypothesis, $n_{A.}$ is our best estimate of the probability that a subject picked at random from the combined population will belong to category A. We would expect, then, to find $n_{.1}(n_{A.}/n)$ of those in the sample from population 1 to belong to category A, $n_{.2}(n_{A.}/n)$ of those in the sample from population 2 to belong to category A, and $n_{.3}(n_{A.}/n)$ of those in the sample from population 3 to belong to category A. These calculations yield the expected frequencies for the first row of Table 12.5.1. Similar reasoning and calculations yield the expected frequencies for the other two rows.

We see again that the shortcut procedure of multiplying appropriate marginal totals and dividing by the grand total yields the expected frequencies for the cells.

From the data in Table 12.5.1 we compute the following test statistic:

$$X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

EXAMPLE 12.5.1

Narcolepsy is a disease involving disturbances of the sleep–wake cycle. Members of the German Migraine and Headache Society (A-8) studied the relationship between migraine headaches in 96 subjects diagnosed with narcolepsy and 96 healthy controls. The results are shown in Table 12.5.2. We wish to know if we may conclude, on the basis of these data,

TABLE 12.5.2 Frequency of Migraine Headaches by Narcolepsy Status

	Reported Migraine Headaches		
	Yes	**No**	**Total**
Narcoleptic subjects	21	75	96
Healthy controls	19	77	96
Total	40	152	192

Source: The DMG Study Group, "Migraine and Idiopathic Narcolepsy—A Case-Control Study," *Cephalagia, 23* (2003), 786–789.

that the narcolepsy population and healthy populations represented by the samples are not homogeneous with respect to migraine frequency.

Solution:

1. **Data.** See Table 12.5.2.
2. **Assumptions.** We assume that we have a simple random sample from each of the two populations of interest.
3. **Hypotheses.**
   H_0: The two populations are homogeneous with respect to migraine frequency.
   H_A: The two populations are not homogeneous with respect to migraine frequency.
   Let $\alpha = .05$.
4. **Test statistic.** The test statistic is
$$X^2 = \sum \left[(O_i - E_i)^2 / E_i \right]$$
5. **Distribution of test statistic.** If H_0 is true, X^2 is distributed approximately as χ^2 with $(2-1)(2-1) = (1)(1) = 1$ degree of freedom.
6. **Decision rule.** Reject H_0 if the computed value of X^2 is equal to or greater than 3.841.
7. **Calculation of test statistic.** The MINITAB output is shown in Figure 12.5.1.

Chi-Square Test

```
Expected counts are printed below observed counts

Rows: Narcolepsy   Columns: Migraine

            No        Yes         All

 No         77         19          96
         76.00      20.00       96.00

Yes         75         21          96
         76.00      20.00       96.00

All        152         40         192
        152.00      40.00      192.00

Chi-Square = 0.126, DF = 1, P-Value = 0.722
```

FIGURE 12.5.1 MINITAB output for Example 12.5.1.

8. **Statistical decision.** Since .126 is less than the critical value of 3.841, we are unable to reject the null hypothesis.
9. **Conclusion.** We conclude that the two populations may be homogeneous with respect to migraine frequency.
10. ***p* value.** From the MINITAB output we see that $p = .722$. ■

Small Expected Frequencies The rules for small expected frequencies given in the previous section are applicable when carrying out a test of homogeneity.

In summary, the chi-square test of homogeneity has the following characteristics:

1. Two or more populations are identified in advance, and an independent sample is drawn from each.
2. Sample subjects or objects are placed in appropriate categories of the variable of interest.
3. The calculation of expected cell frequencies is based on the rationale that if the populations are homogeneous as stated in the null hypothesis, the best estimate of the probability that a subject or object will fall into a particular category of the variable of interest can be obtained by pooling the sample data.
4. The hypotheses and conclusions are stated in terms of homogeneity (with respect to the variable of interest) of populations.

Test of Homogeneity and $H_0: p_1 = p_2$ The chi-square test of homogeneity for the two-sample case provides an alternative method for testing the null hypothesis that two population proportions are equal. In Section 7.6, it will be recalled, we learned to test $H_0: p_1 = p_2$ against $H_A: p_1 \neq p_2$ by means of the statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (\hat{p}_1 - \hat{p}_2)_0}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

where $\bar{p}$ is obtained by pooling the data of the two independent samples available for analysis.

Suppose, for example, that in a test of $H_0: p_1 = p_2$ against $H_A: p_1 \neq p_2$, the sample data were as follows: $n_1 = 100$, $\hat{p}_1 = .60$, $n_2 = 120$, $\hat{p}_2 = .40$. When we pool the sample data we have

$$\bar{p} = \frac{.60(100) + .40(120)}{100 + 120} = \frac{108}{220} = .4909$$

and

$$z = \frac{.60 - .40}{\sqrt{\frac{(.4909)(.5091)}{100} + \frac{(.4909)(.5091)}{120}}} = 2.95469$$

which is significant at the .05 level since it is greater than the critical value of 1.96.

If we wish to test the same hypothesis using the chi-square approach, our contingency table will be

	Characteristic Present		
Sample	**Yes**	**No**	**Total**
1	60	40	100
2	48	72	120
Total	108	112	220

By Equation 12.4.1 we compute

$$X^2 = \frac{220[(60)(72) - (40)(48)]^2}{(108)(112)(100)(120)} = 8.7302$$

which is significant at the .05 level because it is greater than the critical value of 3.841. We see, therefore, that we reach the same conclusion by both methods. This is not surprising because, as explained in Section 12.2, $\chi^2_{(1)} = z^2$. We note that $8.7302 = (2.95469)^2$ and that $3.841 = (1.96)^2$.

EXERCISES

In the exercises that follow perform the test at the indicated level of significance and determine the p value.

12.5.1 Refer to the study by Carter et al. [A-9], who investigated the effect of age at onset of bipolar disorder on the course of the illness. One of the variables studied was subjects' family history. Table 3.4.1 shows the frequency of a family history of mood disorders in the two groups of interest: early age at onset (18 years or younger) and later age at onset (later than 18 years).

Family History of Mood Disorders	**Early $\leq$ 18(E)**	**Later $>$ 18(L)**	**Total**
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

Can we conclude on the basis of these data that subjects 18 or younger differ from subjects older than 18 with respect to family histories of mood disorders? Let $\alpha = .05$.

12.5.2 Coughlin et al. (A-10) examined breast and cervical screening practices of Hispanic and non-Hispanic women in counties that approximate the U.S. southern border region. The study used data from the Behavioral Risk Factor Surveillance System surveys of adults ages 18 years or older conducted in 1999 and 2000. The following table shows the number of observations of Hispanic and non-Hispanic women who had received a mammogram in the past 2 years cross-classified by marital status.

Marital Status	Hispanic	Non-Hispanic	Total
Currently married	319	738	1057
Divorced or separated	130	329	459
Widowed	88	402	490
Never married or living as an unmarried couple	41	95	136
Total	578	1564	2142

Source: Steven S. Coughlin, Robert J. Uhler, Thomas Richards, and Katherine M. Wilson, "Breast and Cervical Cancer Screening Practices Among Hispanic and Non-Hispanic Women Residing Near the United States–Mexico Border, 1999–2000," *Family and Community Health, 26*, (2003), 130–139.

We wish to know if we may conclude on the basis of these data that marital status and ethnicity (Hispanic and non-Hispanic) in border counties of the southern United States are not homogeneous. Let $\alpha = .05$.

12.5.3 Swor et al. (A-11) examined the effectiveness of cardiopulmonary resuscitation (CPR) training in people over 55 years of age. They compared the skill retention rates of subjects in this age group who completed a course in traditional CPR instruction with those who received chest-compression–only cardiopulmonary resuscitation (CC-CPR). Independent groups were tested 3 months after training. Among the 27 subjects receiving traditional CPR, 12 were rated as competent. In the CC-CPR group, 15 out of 29 were rated competent. Do these data provide sufficient evidence for us to conclude that the two populations are not homogeneous with respect to competency rating 3 months after training? Let $\alpha = .05$.

12.5.4 In an air pollution study, a random sample of 200 households was selected from each of two communities. A respondent in each household was asked whether or not anyone in the household was bothered by air pollution. The responses were as follows:

	Any Member of Household Bothered by Air Pollution?		
Community	Yes	No	Total
I	43	157	200
II	81	119	200
Total	124	276	400

Can the researchers conclude that the two communities differ with respect to the variable of interest? Let $\alpha = .05$.

12.5.5 In a simple random sample of 250 industrial workers with cancer, researchers found that 102 had worked at jobs classified as "high exposure" with respect to suspected cancer-causing agents. Of the remainder, 84 had worked at "moderate exposure" jobs, and 64 had experienced no known exposure because of their jobs. In an independent simple random sample of 250 industrial workers from the same area who had no history of cancer, 31 worked in "high exposure" jobs, 60 worked in "moderate exposure" jobs, and 159 worked in jobs involving no known exposure to suspected cancer-causing agents. Does it appear from these data that persons working in jobs that expose them to suspected cancer-causing agents have an increased risk of contracting cancer? Let $\alpha = .05$.

12.6 THE FISHER EXACT TEST

Sometimes we have data that can be summarized in a 2×2 contingency table, but these data are derived from very small samples. The chi-square test is not an appropriate method of analysis if minimum expected frequency requirements are not met. If, for example, n is less than 20 or if n is between 20 and 40 and one of the expected frequencies is less than 5, the chi-square test should be avoided.

A test that may be used when the size requirements of the chi-square test are not met was proposed in the mid-1930s almost simultaneously by Fisher (7,8), Irwin (9), and Yates (10). The test has come to be known as the *Fisher exact test*. It is called exact because, if desired, it permits us to calculate the exact probability of obtaining the observed results or results that are more extreme.

Data Arrangement When we use the Fisher exact test, we arrange the data in the form of a 2×2 contingency table like Table 12.6.1. We arrange the frequencies in such a way that $A > B$ and choose the characteristic of interest so that $a/A > b/B$.

Some theorists believe that Fisher's exact test is appropriate only when both marginal totals of Table 12.6.1 are fixed by the experiment. This specific model does not appear to arise very frequently in practice. Many experimenters, therefore, use the test when both marginal totals are not fixed.

Assumptions The following are the assumptions for the Fisher exact test.

1. The data consist of A sample observations from population 1 and B sample observations from population 2.
2. The samples are random and independent.
3. Each observation can be categorized as one of two mutually exclusive types.

TABLE 12.6.1 A 2×2 Contingency Table for the Fisher Exact Test

Sample	With Characteristic	Without Characteristic	Total
1	a	$A - a$	A
2	b	$B - b$	B
Total	$a + b$	$A + B - a - b$	$A + B$

Hypotheses The following are the null hypotheses that may be tested and their alternatives.

1. (Two-sided)

   H_0: The proportion with the characteristic of interest is the same in both populations; that is, $p_1 = p_2$.

   H_A: The proportion with the characteristic of interest is not the same in both populations; $p_1 \neq p_2$.

2. (One-sided)

   H_0: The proportion with the characteristic of interest in population 1 is less than or the same as the proportion in population 2; $p_1 \leq p_2$.

   H_A: The proportion with the characteristic of interest is greater in population 1 than in population 2; $p_1 > p_2$.

Test Statistic The test statistic is b, the number in sample 2 with the characteristic of interest.

Decision Rule Finney (11) has prepared critical values of b for $A \leq 15$. Latscha (12) has extended Finney's tables to accommodate values of A up to 20. Appendix Table J gives these critical values of b for A between 3 and 20, inclusive. Significance levels of .05, .025, .01, and .005 are included. The specific decision rules are as follows:

1. **Two-sided test.** Enter Table J with A, B, and a. If the observed value of b is equal to or less than the integer in a given column, reject H_0 at a level of significance equal to twice the significance level shown at the top of that column. For example, suppose $A = 8$, $B = 7$, $a = 7$, and the observed value of b is 1. We can reject the null hypothesis at the $2(.05) = .10$, the $2(.025) = .05$, and the $2(.01) = .02$ levels of significance, but not at the $2(.005) = .01$ level.
2. **One-sided test.** Enter Table J with A, B, and a. If the observed value of b is less than or equal to the integer in a given column, reject H_0 at the level of significance shown at the top of that column. For example, suppose that $A = 16$, $B = 8$, $a = 4$, and the observed value of b is 3. We can reject the null hypothesis at the .05 and .025 levels of significance, but not at the .01 or .005 levels.

Large-Sample Approximation For sufficiently large samples we can test the null hypothesis of the equality of two population proportions by using the normal approximation. Compute

$$z = \frac{(a/A) - (b/B)}{\sqrt{\hat{p}(1-\hat{p})(1/A + 1/B)}} \tag{12.6.1}$$

where

$$\hat{p} = (a + b)/(A + B) \tag{12.6.2}$$

and compare it for significance with appropriate critical values of the standard normal distribution. The use of the normal approximation is generally considered satisfactory if a,

b, $A - a$, and $B - b$ are all greater than or equal to 5. Alternatively, when sample sizes are sufficiently large, we may test the null hypothesis by means of the chi-square test.

Further Reading The Fisher exact test has been the subject of some controversy among statisticians. Some feel that the assumption of fixed marginal totals is unrealistic in most practical applications. The controversy then centers around whether the test is appropriate when both marginal totals are not fixed. For further discussion of this and other points, see the articles by Barnard (13–15), Fisher (16), and Pearson (17).

Sweetland (18) compared the results of using the chi-square test with those obtained using the Fisher exact test for samples of size $A + B = 3$ to $A + B = 69$. He found close agreement when A and B were close in size and the test was one-sided.

Carr (19) presents an extension of the Fisher exact test to more than two samples of equal size and gives an example to demonstrate the calculations. Neave (20) presents the Fisher exact test in a new format; the test is treated as one of independence rather than of homogeneity. He has prepared extensive tables for use with his approach.

The sensitivity of Fisher's exact test to minor perturbations in 2×2 contingency tables is discussed by Dupont (21).

EXAMPLE 12.6.1

The purpose of a study by Justesen et al. (A-12) was to evaluate the long-term efficacy of taking indinavir/ritonavir twice a day in combination with two nucleoside reverse transcriptase inhibitors among HIV-positive subjects who were divided into two groups. Group 1 consisted of patients who had no history of taking protease inhibitors (PI Naïve). Group 2 consisted of patients who had a previous history taking a protease inhibitor (PI Experienced). Table 12.6.2 shows whether these subjects remained on the regimen for the 120 weeks of follow-up. We wish to know if we may conclude that patients classified as group 1 have a lower probability than subjects in group 2 of remaining on the regimen for 120 weeks.

TABLE 12.6.2 Regimen Status at 120 Weeks for PI Naïve and PI Experienced Subjects Taking Indinavir/Ritonavir as Described in Example 12.6.1

		Remained in the Regimen for 120 Weeks	
	Total	Yes	No
1 (PI Naïve)	9	2	7
2 (PA Experienced)	12	8	4
Total	21	10	11

Source: U.S. Justesen, A. M. Lervfing, A. Thomsen, J. A. Lindberg, C. Pedersen, and P. Tauris, "Low-Dose Indinavir in Combination with Low-Dose Ritonavir: Steady-State Pharmacokinetics and Long-Term Clinical Outcome Follow-Up," *HIV Medicine, 4* (2003), 250–254.

TABLE 12.6.3 Data of Table 12.6.2 Rearranged to Conform to the Layout of Table 12.6.1

	Remained in Regimen for 120 Weeks		
	Yes	No	Total
2 (PI Experienced)	$8 = a$	$4 = A - a$	$12 = A$
1 (PI Naïve)	$2 = b$	$7 = B - b$	$9 = B$
Total	$10 = a + b$	$11 = A + B - a - b$	$21 = A + B$

Solution:

1. **Data.** The data as reported are shown in Table 12.6.2. Table 12.6.3 shows the data rearranged to conform to the layout of Table 12.6.1. Remaining on the regimen is the characteristic of interest.
2. **Assumptions.** We presume that the assumptions for application of the Fisher exact test are met.
3. **Hypotheses.**
   H_0: The proportion of subjects remaining 120 weeks on the regimen in a population of patients classified as group 2 is the same as or less than the proportion of subjects remaining on the regimen 120 weeks in a population classified as group 1.
   H_A: Group 2 patients have a higher rate than group 1 patients of remaining on the regimen for 120 weeks.
4. **Test statistic.** The test statistic is the observed value of b as shown in Table 12.6.3.
5. **Distribution of test statistic.** We determine the significance of b by consulting Appendix Table J.
6. **Decision rule.** Suppose we let $\alpha = .05$. The decision rule, then, is to reject H_0 if the observed value of b is equal to or less than 1, the value of b in Table J for $A = 12$, $B = 9$, $a = 8$, and $\alpha = .05$.
7. **Calculation of test statistic.** The observed value of b, as shown in Table 12.6.3, is 2.
8. **Statistical decision.** Since $2 > 1$, we fail to reject H_0.
9. **Conclusion.** Since we fail to reject H_0, we conclude that the null hypothesis may be true. That is, it may be true that the rate of remaining on the regimen for 120 weeks is the same or less for the PI experienced group compared to the PI naïve group.
10. ***p* value.** We see in Table J that when $A = 12, B = 9, a = 8$, the value of $b = 2$ has an exact probability of occurring by chance alone, when H_0 is true, greater than .05. ■

PI * Remained Cross-Tabulation

Count

		Remained		
		Yes	No	Total
PI	Experienced	8	4	12
	Naive	2	7	9
Total		10	11	21

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.073[b]	1	.044		
Continuity Correction[a]	2.486	1	.115		
Likelihood Ratio	4.253	1	.039		
Fisher's Exact Test				.080	.056
Linear-by-Linear Association	3.879	1	.049		
N of Valid Cases	21				

a. Computed only for a 2 × 2 table

b. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 4.29.

FIGURE 12.6.1 SPSS output for Example 12.6.1.

Various statistical software programs perform the calculations for the Fisher exact test. Figure 12.6.1 shows the results of Example 12.6.1 as computed by SPSS. The exact p value is provided for both a one-sided and a two-sided test. Based on these results, we fail to reject H_0 (p value $>.05$), just as we did using the statistical tables in the Appendix. Note that in addition to the Fisher exact test several alternative tests are provided. The reader should be aware that these alternative tests are not appropriate if the assumptions underlying them have been violated.

EXERCISES

12.6.1 The goal of a study by Tahmassebi and Curzon (A-13) was to determine if drooling in children with cerebral palsy is due to hypersalivation. One of the procedures toward that end was to examine the salivary buffering capacity of cerebral palsied children and controls. The following table gives the results.

	Buffering Capacity	
Group	**Medium**	**High**
Cerebral palsy	2	8
Control	3	7

Source: J. F. Tahmassebi and M. E. J. Curzon, "The Cause of Drooling in Children with Cerebral Palsy—Hypersalivation or Swallowing Defect?" *International Journal of Paediatric Dentistry, 13* (2003), 106–111.

Test for a significant difference between cerebral palsied children and controls with respect to high or low buffering capacity. Let $\alpha = .05$ and find the p value.

12.6.2 In a study by Xiao and Shi (A-14), researchers studied the effect of cranberry juice in the treatment and prevention of *Helicobacter pylori* infection in mice. The eradication of *Helicobacter pylori* results in the healing of peptic ulcers. Researchers compared treatment with cranberry juice to "triple therapy (amoxicillin, bismuth subcitrate, and metronidazole) in mice infected with *Helicobacter pylori*. After 4 weeks, they examined the mice to determine the frequency of eradication of the bacterium in the two treatment groups. The following table shows the results.

	No. of Mice with *Helicobacter pylori* Eradicated	
	Yes	**No**
Triple therapy	8	2
Cranberry juice	2	8

Source: Shu Dong Xiao and Tong Shi, "Is Cranberry Juice Effective in the Treatment and Prevention of *Helicobacter Pylori* Infection of Mice," *Chinese Journal of Digestive Diseases, 4* (2003), 136–139.

May we conclude, on the basis of these data, that triple therapy is more effective than cranberry juice at eradication of the bacterium? Let $\alpha = .05$ and find the p value.

12.6.3 In a study by Shaked et al. (A-15), researchers studied 26 children with blunt pancreatic injuries. These injuries occurred from a direct blow to the abdomen, bicycle handlebars, fall from height, or car accident. Nineteen of the patients were classified as having minor injuries, and seven were classified as having major injuries. Pseudocyst formation was suspected when signs of clinical deterioration developed, such as increased abdominal pain, epigastric fullness, fever, and increased pancreatic enzyme levels. In the major injury group, six of the seven children developed pseudocysts while in the minor injury group, three of the 19 children developed pseudocysts. Is this sufficient evidence to allow us to conclude that the proportion of children developing pseudocysts is higher in the major injury group than in the minor injury group? Let $\alpha = .01$.

12.7 RELATIVE RISK, ODDS RATIO, AND THE MANTEL–HAENSZEL STATISTIC

In Chapter 8 we learned to use analysis of variance techniques to analyze data that arise from designed experiments, investigations in which at least one variable is manipulated in some way. Designed experiments, of course, are not the only sources of data that are

of interest to clinicians and other health sciences professionals. Another important class of scientific investigation that is widely used is the *observational study.*

DEFINITION

An *observational study* is a scientific investigation in which neither the subjects under study nor any of the variables of interest are manipulated in any way.

An observational study, in other words, may be defined simply as an investigation that is not an experiment. The simplest form of observational study is one in which there are only two variables of interest. One of the variables is called the *risk factor*, or independent variable, and the other variable is referred to as the *outcome,* or dependent variable.

DEFINITION

The term *risk factor* is used to designate a variable that is thought to be related to some outcome variable. The risk factor may be a suspected cause of some specific state of the outcome variable.

In a particular investigation, for example, the outcome variable might be subjects' status relative to cancer and the risk factor might be their status with respect to cigarette smoking. The model is further simplified if the variables are categorical with only two categories per variable. For the outcome variable the categories might be cancer present and cancer absent. With respect to the risk factor subjects might be categorized as smokers and nonsmokers.

When the variables in observational studies are categorical, the data pertaining to them may be displayed in a contingency table, and hence the inclusion of the topic in the present chapter. We shall limit our discussion to the situation in which the outcome variable and the risk factor are both dichotomous variables.

Types of Observational Studies There are two basic types of observational studies, *prospective studies* and *retrospective studies*.

DEFINITION

A *prospective study* is an observational study in which two random samples of subjects are selected. One sample consists of subjects who possess the risk factor, and the other sample consists of subjects who do not possess the risk factor. The subjects are followed into the future (that is, they are followed prospectively), and a record is kept on the number of subjects in each sample who, at some point in time, are classifiable into each of the categories of the outcome variable.

The data resulting from a prospective study involving two dichotomous variables can be displayed in a 2×2 contingency table that usually provides information regarding the number of subjects with and without the risk factor and the number who did and did not

TABLE 12.7.1 Classification of a Sample of Subjects with Respect to Disease Status and Risk Factor

	Disease Status		
Risk Factor	**Present**	**Absent**	**Total at Risk**
Present	a	b	$a+b$
Absent	c	d	$c+d$
Total	$a+c$	$b+d$	n

succumb to the disease of interest as well as the frequencies for each combination of categories of the two variables.

DEFINITION

A *retrospective study* is the reverse of a prospective study. The samples are selected from those falling into the categories of the outcome variable. The investigator then looks back (that is, takes a retrospective look) at the subjects and determines which ones have (or had) and which ones do not have (or did not have) the risk factor.

From the data of a retrospective study we may construct a contingency table with frequencies similar to those that are possible for the data of a prospective study.

In general, the prospective study is more expensive to conduct than the retrospective study. The prospective study, however, more closely resembles an experiment.

Relative Risk The data resulting from a prospective study in which the dependent variable and the risk factor are both dichotomous may be displayed in a 2×2 contingency table such as Table 12.7.1. The risk of the development of the disease among the subjects with the risk factor is $a/(a+b)$. The risk of the development of the disease among the subjects without the risk factor is $c/(c+d)$. We define relative risk as follows.

DEFINITION

***Relative risk* is the ratio of the risk of developing a disease among subjects with the risk factor to the risk of developing the disease among subjects without the risk factor.**

We represent the relative risk from a prospective study symbolically as

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)} \tag{12.7.1}$$

where a, b, c, and d are as defined in Table 12.7.1, and $\widehat{RR}$ indicates that the relative risk is computed from a sample to be used as an estimate of the relative risk, RR, for the population from which the sample was drawn.

We may construct a confidence interval for *RR*

$$100(1-\alpha)\%\text{CI} = \widehat{RR}^{1\pm\left(z_\alpha/\sqrt{X^2}\right)} \tag{12.7.2}$$

where z_α is the two-sided z value corresponding to the chosen confidence coefficient and X^2 is computed by Equation 12.4.1.

Interpretation of *RR* The value of *RR* may range anywhere between zero and infinity. A value of 1 indicates that there is no association between the status of the risk factor and the status of the dependent variable. In most cases the two possible states of the dependent variable are disease present and disease absent. We interpret an *RR* of 1 to mean that the risk of acquiring the disease is the same for those subjects with the risk factor and those without the risk factor. A value of *RR* greater than 1 indicates that the risk of acquiring the disease is greater among subjects with the risk factor than among subjects without the risk factor. An *RR* value that is less than 1 indicates less risk of acquiring the disease among subjects with the risk factor than among subjects without the risk factor. For example, a risk factor of 2 is taken to mean that those subjects with the risk factor are twice as likely to acquire the disease as compared to subjects without the risk factor.

We illustrate the calculation of relative risk by means of the following example.

EXAMPLE 12.7.1

In a prospective study of pregnant women, Magann et al. (A-16) collected extensive information on exercise level of low-risk pregnant working women. A group of 217 women did no voluntary or mandatory exercise during the pregnancy, while a group of 238 women exercised extensively. One outcome variable of interest was experiencing preterm labor. The results are summarized in Table 12.7.2.

We wish to estimate the relative risk of preterm labor when pregnant women exercise extensively.

Solution: By Equation 12.7.1 we compute

$$\widehat{RR} = \frac{22/238}{18/217} = \frac{.0924}{.0829} = 1.1$$

TABLE 12.7.2 Subjects with and without the Risk Factor Who Became Cases of Preterm Labor

Risk Factor	Cases of Preterm Labor	Noncases of Preterm Labor	Total
Extreme exercising	22	216	238
Not exercising	18	199	217
Total	40	415	455

Source: Everett F. Magann, Sharon F. Evans, Beth Weitz, and John Newnham, "Antepartum, Intrapartum, and Neonatal Significance of Exercise on Healthy Low-Risk Pregnant Working Women," *Obstetrics and Gynecology, 99* (2002), 466–472.

Odds Ratio and Relative Risk Section

Parameter	Common Odds Ratio	Original Odds Ratio	Iterated Odds Ratio	Log Odds Ratio	Relative Risk
Upper 95% C.L.		2.1350	2.2683	0.7585	2.1192
Estimate	1.1260	1.1207	1.1207	0.1140	1.1144
Lower 95% C.L.		0.5883	0.5606	−0.5305	0.5896

FIGURE 12.7.1 NCSS output for the data in Example 12.7.1.

These data indicate that the risk of experiencing preterm labor when a woman exercises heavily is 1.1 times as great as it is among women who do not exercise at all.

We compute the 95 percent confidence interval for *RR* as follows. By Equation 12.4.1, we compute from the data in Table 12.7.2:

$$X^2 = \frac{455[(22)(199) - (216)(18)]^2}{(40)(415)(238)(217)} = .1274$$

By Equation 12.7.2, the lower and upper confidence limits are, respectively, $1.1^{1-1.96/\sqrt{.1274}} = .65$ and $1.1^{1+1.96/\sqrt{.1274}} = 1.86$. Since the interval includes 1, we conclude, at the .05 level of significance, that the population risk may be 1. In other words, we conclude that, in the population, there may not be an increased risk of experiencing preterm labor when a pregnant woman exercises extensively.

The data were processed by NCSS. The results are shown in Figure 12.7.1. The relative risk calculation is shown in the column at the far right of the output, along with the 95% confidence limits. Because of rounding errors, these values differ slightly from those given in the example. ■

Odds Ratio When the data to be analyzed come from a retrospective study, relative risk is not a meaningful measure for comparing two groups. As we have seen, a retrospective study is based on a sample of subjects with the disease (cases) and a separate sample of subjects without the disease (controls or noncases). We then retrospectively determine the distribution of the risk factor among the cases and controls. Given the results of a retrospective study involving two samples of subjects, cases, and controls, we may display the data in a 2×2 table such as Table 12.7.3, in which subjects are dichotomized with respect to the presence and absence of the risk factor. Note that the column headings in Table 12.7.3 differ from those in Table 12.7.1 to emphasize the fact that the data are from a retrospective study and that the subjects were selected because they were either cases or controls. When the data from a retrospective study are displayed as in Table 12.7.3, the ratio $a/(a + b)$, for example, is not an estimate of the risk of disease for subjects with the risk factor. The appropriate measure for comparing cases and controls in a retrospective study is the *odds ratio*. As noted in Chapter 11, in order to understand the concept of

TABLE 12.7.3 Subjects of a Retrospective Study Classified According to Status Relativeto a Risk Factor and Whether They Are Cases or Controls

	Sample		
Risk Factor	**Cases**	**Controls**	**Total**
Present	a	b	$a+b$
Absent	c	d	$c+d$
Total	$a+c$	$b+d$	n

the odds ratio, we must understand the term *odds,* which is frequently used by those who place bets on the outcomes of sporting events or participate in other types of gambling activities.

DEFINITION

The odds for success are the ratio of the probability of success to the probability of failure.

We use this definition of odds to define two odds that we can calculate from data displayed as in Table 12.7.3:

1. The odds of being a case (having the disease) to being a control (not having the disease) among subjects with the risk factor is $[a/(a+b)]/[b/(a+b)] = a/b$.
2. The odds of being a case (having the disease) to being a control (not having the disease) among subjects without the risk factor is $[c/(c+d)]/[d/(c+d)] = c/d$.

We now define the odds ratio that we may compute from the data of a retrospective study. We use the symbol $\widehat{OR}$ to indicate that the measure is computed from sample data and used as an estimate of the population odds ratio, OR.

DEFINITION

The estimate of the population odds ratio is

$$\widehat{OR} = \frac{a/b}{c/d} = \frac{ad}{bc} \tag{12.7.3}$$

where *a, b, c,* and *d* are as defined in Table 12.7.3.

We may construct a confidence interval for OR by the following method:

$$100(1-\alpha)\%\,\text{CI} = \widehat{OR}^{1\pm\left(z_\alpha/\sqrt{X^2}\right)} \tag{12.7.4}$$

where z_α is the two-sided z value corresponding to the chosen confidence coefficient and X^2 is computed by Equation 12.4.1.

Interpretation of the Odds Ratio In the case of a rare disease, the population odds ratio provides a good approximation to the population relative risk. Consequently, the sample odds ratio, being an estimate of the population odds ratio, provides an indirect estimate of the population relative risk in the case of a rare disease.

The odds ratio can assume values between zero and ∞. A value of 1 indicates no association between the risk factor and disease status. A value less than 1 indicates reduced odds of the disease among subjects with the risk factor. A value greater than 1 indicates increased odds of having the disease among subjects in whom the risk factor is present.

EXAMPLE 12.7.2

Toschke et al. (A-17) collected data on obesity status of children ages 5–6 years and the smoking status of the mother during the pregnancy. Table 12.7.4 shows 3970 subjects classified as cases or noncases of obesity and also classified according to smoking status of the mother during pregnancy (the risk factor). We wish to compare the odds of obesity at ages 5–6 among those whose mother smoked throughout the pregnancy with the odds of obesity at age 5–6 among those whose mother did not smoke during pregnancy.

Solution: The odds ratio is the appropriate measure for answering the question posed. By Equation 12.7.3 we compute

$$\widehat{OR} = \frac{(64)(3496)}{(342)(68)} = 9.62$$

We see that obese children (cases) are 9.62 times as likely as nonobese children (noncases) to have had a mother who smoked throughout the pregnancy.

We compute the 95 percent confidence interval for OR as follows. By Equation 12.4.1 we compute from the data in Table 12.7.4

$$X^2 = \frac{3970[(64)(3496) - (342)(68)]^2}{(132)(3838)(406)(3564)} = 217.6831$$

TABLE 12.7.4 Subjects Classified According to Obesity Status and Mother's Smoking Status during Pregnancy

	Obesity Status		
Smoking Status During Pregnancy	**Cases**	**Noncases**	**Total**
Smoked throughout	64	342	406
Never smoked	68	3496	3564
Total	132	3838	3970

Source: A. M. Toschke, S. M. Montgomery, U. Pfeiffer, and R. von Kries, "Early Intrauterine Exposure to Tobacco-Inhaled Products and Obesity," *American Journal of Epidemiology, 158* (2003), 1068–1074.

Smoking_status * Obsesity_status Cross-Tabulation

Count

		Obesity status		
		Cases	Noncases	Total
Smoking_status	Smoked throughout	64	342	406
	Never smoked	68	3496	3564
Total		132	3838	3970

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Smoking_status (Smoked throughout /Never smoked)	9.621	6.719	13.775
For cohort Obesity_ status = Cases	8.262	5.966	11.441
For cohort Obesity_ status = Noncases	.859	.823	.896
N of Valid Cases	3970		

FIGURE 12.7.2 SPSS output for Example 12.7.2.

The lower and upper confidence limits for the population *OR*, respectively, are $9.62^{1-1.96/\sqrt{217.6831}} = 7.12$ and $9.62^{1+1.96/\sqrt{217.6831}} = 13.00$. We conclude with 95 percent confidence that the population *OR* is somewhere between 7.12 and 13.00. Because the interval does not include 1, we conclude that, in the population, obese children (cases) are more likely than nonobese children (noncases) to have had a mother who smoked throughout the pregnancy.

The data from Example 12.7.2 were processed using SPSS. The results are shown in Figure 12.7.2. The odds ratio calculation, along with the 95% confidence limits, are shown in the top line of the Risk Estimate box. These values differ slightly from those in the example because of rounding error. ■

The Mantel–Haenszel Statistic Frequently when we are studying the relationship between the status of some disease and the status of some risk factor, we are

aware of another variable that may be associated with the disease, with the risk factor, or with both in such a way that the true relationship between the disease status and the risk factor is masked. Such a variable is called a *confounding variable*. For example, experience might indicate the possibility that the relationship between some disease and a suspected risk factor differs among different ethnic groups. We would then treat ethnic membership as a confounding variable. When they can be identified, it is desirable to control for confounding variables so that an unambiguous measure of the relationship between disease status and risk factor may be calculated. A technique for accomplishing this objective is the Mantel–Haenszel (22) procedure, so called in recognition of the two men who developed it. The procedure allows us to test the null hypothesis that there is no association between status with respect to disease and risk factor status. Initially used only with data from retrospective studies, the Mantel–Haenszel procedure is also appropriate for use with data from prospective studies, as discussed by Mantel (23).

In the application of the Mantel–Haenszel procedure, case and control subjects are assigned to strata corresponding to different values of the confounding variable. The data are then analyzed within individual strata as well as across all strata. The discussion that follows assumes that the data under analysis are from a retrospective or a prospective study with case and noncase subjects classified according to whether they have or do not have the suspected risk factor. The confounding variable is categorical, with the different categories defining the strata. If the confounding variable is continuous it must be categorized. For example, if the suspected confounding variable is age, we might group subjects into mutually exclusive age categories. The data before stratification may be displayed as shown in Table 12.7.3.

Application of the Mantel–Haenszel procedure consists of the following steps.

1. Form k strata corresponding to the k categories of the confounding variable. Table 12.7.5 shows the data display for the ith stratum.
2. For each stratum compute the expected frequency e_i of the upper left-hand cell of Table 12.7.5 as follows:

$$e_i = \frac{(a_i + b_i)(a_i + c_i)}{n_i} \tag{12.7.5}$$

TABLE 12.7.5 Subjects in the *i*th Stratum of a Confounding Variable Classified According to Status Relative to a Risk Factor and Whether They Are Cases or Controls

	Sample		
Risk Factor	**Cases**	**Controls**	**Total**
Present	a_i	b_i	$a_i + b_i$
Absent	c_i	d_i	$c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	n_i

3. For each stratum compute

$$v_i = \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)} \tag{12.7.6}$$

4. Compute the Mantel–Haenszel test statistic, χ^2_{MH} as follows:

$$\chi^2_{\text{MH}} = \frac{\left(\sum_{i=1}^{k} a_i - \sum_{i=1}^{k} e_i\right)^2}{\sum_{i=1}^{k} v_i} \tag{12.7.7}$$

5. Reject the null hypothesis of no association between disease status and suspected risk factor status in the population if the computed value of χ^2_{MH} is equal to or greater than the critical value of the test statistic, which is the tabulated chi-square value for 1 degree of freedom and the chosen level of significance.

Mantel–Haenszel Estimator of the Common Odds Ratio When we have k strata of data, each of which may be displayed in a table like Table 12.7.5, we may compute the Mantel–Haenszel estimator of the common odds ratio, $\widehat{OR}_{\text{MH}}$ as follows:

$$\widehat{OR}_{\text{MH}} = \frac{\sum_{i=1}^{k} (a_i d_i / n_i)}{\sum_{i=1}^{k} (b_i c_i / n_i)} \tag{12.7.8}$$

When we use the Mantel–Haenszel estimator given by Equation 12.7.4, we assume that, in the population, the odds ratio is the same for each stratum.

We illustrate the use of the Mantel–Haenszel statistics with the following examples.

EXAMPLE 12.7.3

In a study by LaMont et al. (A-18), researchers collected data on obstructive coronary artery disease (OCAD), hypertension, and age among subjects identified by a treadmill stress test as being at risk. In Table 12.7.6, counts on subjects in two age strata are presented with hypertension as the risk factor and the presence of OCAD as the case/noncase variable.

Solution:

1. **Data.** See Table 12.7.6.
2. **Assumptions.** We assume that the assumptions discussed earlier for the valid use of the Mantel–Haenszel statistic are met.

TABLE 12.7.6 Patients Stratified by Age and Classified by Status Relative to Hypertension (the Risk Factor) and OCAD (Case/Noncase Variable)

Stratum 1 (55 and under)			
Risk Factor (Hypertension)	**Cases (OCAD)**	**Noncases**	**Total**
Present	21	11	32
Absent	16	6	22
Total	37	17	54
Stratum 2 (over 55)			
Risk Factor (Hypertension)	**Cases (OCAD)**	**Noncases**	**Total**
Present	50	14	64
Absent	18	6	24
Total	68	20	88

Source: Data provided courtesy of Matthew J. Budoff, MD.

3. **Hypotheses.**

   H_0: There is no association between the presence of hypertension and occurrence of OCAD in subjects 55 and under and subjects over 55.

   H_A: There is a relationship between the two variables.

4. **Test statistic.**

$$\chi^2_{MH} = \frac{\left(\sum_{i=1}^{k} a_i - \sum_{i=1}^{k} e_i\right)^2}{\sum_{i=1}^{k} v_i}$$

   as given in Equation 12.7.7.

5. **Distribution of test statistic.** Chi-square with 1 degree of freedom.

6. **Decision rule.** Suppose we let $\alpha = .05$. Reject H_0 if the computed value of the test statistic is greater than or equal to 3.841.

7. **Calculation of test statistic.** By Equation 12.7.5 we compute the following expected frequencies:

$$e_1 = (21 + 11)(21 + 16)/54 = (32)(37)/54 = 21.93$$
$$e_2 = (50 + 14)(50 + 18)/88 = (64)(68)/88 = 49.45$$

By Equation 12.7.6 we compute

$$v_1 = (32)(22)(37)(17)/(2916)(54-1) = 2.87$$
$$v_2 = (64)(24)(68)(20)/(7744)(88-1) = 3.10$$

Finally, by Equation 12.7.7 we compute

$$\chi^2_{\text{MH}} = \frac{[(21+50)-(21.93+49.45)]^2}{2.87+3.10} = .0242$$

8. **Statistical decision.** Since $.0242 < 3.841$, we fail to reject H_0.
9. **Conclusion.** We conclude that there may not be an association between hypertension and the occurrence of OCAD.
10. ***p* value.** Since $.0242 < 2.706$, the p value for this test is $p > .10$.

We now illustrate the calculation of the Mantel–Haenszel estimator of the common odds ratio. ■

EXAMPLE 12.7.4

Let us refer to the data in Table 12.7.6 and compute the common odds ratio.

Solution: From the stratified data in Table 12.7.6 we compute the numerator of the ratio as follows:

$$(a_1d_1/n_1) + (a_2d_2/n_2) = [(21)(6)/54] + [(50)(6)/88]$$
$$= 5.7424$$

The denominator of the ratio is

$$(b_1c_1/n_1) + (b_2c_2/n_2) = [(11)(16)/54] + [(14)(18)/88]$$
$$= 6.1229$$

Now, by Equation 12.7.7, we compute the common odds ratio:

$$\widehat{OR}_{\text{MH}} = \frac{5.7424}{6.1229} = .94$$

From these results we estimate that, regardless of age, patients who have hypertension are less likely to have OCAD than patients who do not have hypertension. ■

Hand calculation of the Mantel–Haenszel test statistics can prove to be a cumbersome task. Fortunately, the researcher can find relief in one of several statistical software packages that are available. To illustrate, results from the use of SPSS to process the data of Example 12.7.3 are shown in Figure 12.7.3. These results differ from those given in the example because of rounding error.

Smoking_status * Obsesity_status * Stratum Cross-Tabulation

Count

Stratum			Obesity status: Cases	Obesity status: Noncases	Total
55 and under	Smoking_status	Smoked throughout	21	11	32
		Never smoked	16	6	22
	Total		37	17	54
Over 55	Smoking_status	Smoked throughout	50	14	64
		Never smoked	18	6	24
	Total		68	20	88

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	.025	1	.875
Mantel-Haenszel	.002	1	.961

Mantel–Haenszel Common Odds Ratio Estimate

Estimate			.938
ln(Estimate)			−.064
Std. Error of ln(Estimate)			.412
Asymp. Sig. (2-sided)			.876
Asymp. 95% confidence Interval	Common Odds Ratio	Lower Bound	.418
		Upper Bound	2.102
	ln(Common Odds Ratio)	Lower Bound	−.871
		Upper Bound	.743

FIGURE 12.7.3 SPSS output for Example 12.7.3.

EXERCISES

12.7.1 Davy et al. (A-19) reported the results of a study involving survival from cervical cancer. The researchers found that among subjects younger than age 50, 16 of 371 subjects had not survived for 1 year after diagnosis. In subjects age 50 or older, 219 of 376 had not survived for 1 year after diagnosis. Compute the relative risk of death among subjects age 50 or older. Does it appear from these data that older subjects diagnosed as having cervical cancer are prone to higher mortality rates?

12.7.2 The objective of a prospective study by Stenestrand et al. (A-20) was to compare the mortality rate following an acute myocardial infarction (AMI) among subjects receiving early revascularization to the mortality rate among subjects receiving conservative treatments. Among 2554 patients receiving revascularization within 14 days of AMI, 84 died in the year following the AMI. In the conservative treatment group (risk factor present), 1751 of 19,358 patients died within a year of AMI. Compute the relative risk of mortality in the conservative treatment group as compared to the revascularization group in patients experiencing AMI.

12.7.3 Refer to Example 12.7.2. Toschke et al. (A-17), who collected data on obesity status of children ages 5–6 years and the smoking status of the mother during the pregnancy, also reported on another outcome variable: whether the child was born premature (37 weeks or fewer of gestation). The following table summarizes the results of this aspect of the study. The same risk factor (smoking during pregnancy) is considered, but a case is now defined as a mother who gave birth prematurely.

	Premature Birth Status		
Smoking Status During Pregnancy	**Cases**	**Noncases**	**Total**
Smoked throughout	36	370	406
Never smoked	168	3396	3564
Total	204	3766	3970

Source: A. M. Toschke, S. M. Montgomery, U. Pfeiffer, and R. von Kries, "Early Intrauterine Exposure to Tobacco-Inhaled Products and Obesity," *American Journal of Epidemiology*, *158* (2003), 1068–1074.

Compute the odds ratio to determine if smoking throughout pregnancy is related to premature birth. Use the chi-square test of independence to determine if one may conclude that there is an association between smoking throughout pregnancy and premature birth. Let $\alpha = .05$.

12.7.4 Sugiyama et al. (A-21) examined risk factors for allergic diseases among 13- and 14-year-old schoolchildren in Japan. One risk factor of interest was a family history of eating an unbalanced diet. The following table shows the cases and noncases of children exhibiting symptoms of rhinitis in the presence and absence of the risk factor.

	Rhinitis		
Family History	**Cases**	**Noncases**	**Total**
Unbalanced diet	656	1451	2107
Balanced diet	677	1662	2339
Total	1333	3113	4446

Source: Takako Sugiyama, Kumiya Sugiyama, Masao Toda, Tastuo Yukawa, Sohei Makino, and Takeshi Fukuda, "Risk Factors for Asthma and Allergic Diseases Among 13–14-Year-Old Schoolchildren in Japan," *Allergology International, 51* (2002), 139–150.

What is the estimated odds ratio of having rhinitis among subjects with a family history of an unbalanced diet compared to those eating a balanced diet? Compute the 95 percent confidence interval for the odds ratio.

12.7.5 According to Holben et al. (A-22), "Food insecurity implies a limited access to or availability of food or a limited/uncertain ability to acquire food in socially acceptable ways." These researchers

collected data on 297 families with a child in the Head Start nursery program in a rural area of Ohio near Appalachia. The main outcome variable of the study was household status relative to food security. Households that were not food secure are considered to be cases. The risk factor of interest was the absence of a garden from which a household was able to supplement its food supply. In the following table, the data are stratified by the head of household's employment status outside the home.

Stratum 1 (Employed Outside the Home)			
Risk Factor	**Cases**	**Noncases**	**Total**
No garden	40	37	77
Garden	13	38	51
Total	53	75	128
Stratum 2 (Not Employed Outside the Home)			
Risk Factor	**Cases**	**Noncases**	**Total**
No garden	75	38	113
Garden	15	33	48
Total	90	71	161

Source: Data provided courtesy of David H. Holben, Ph.D. and John P. Holcomb, Jr., Ph.D.

Compute the Mantel–Haenszel common odds ratio with stratification by employment status. Use the Mantel–Haenszel chi-square test statistic to determine if we can conclude that there is an association between the risk factor and food insecurity. Let $\alpha = .05$.

12.8 SUMMARY

In this chapter some uses of the versatile chi-square distribution are discussed. Chi-square goodness-of-fit tests applied to the normal, binomial, and Poisson distributions are presented. We see that the procedure consists of computing a statistic

$$X^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

that measures the discrepancy between the observed (O_i) and expected (E_i) frequencies of occurrence of values in certain discrete categories. When the appropriate null hypothesis is true, this quantity is distributed approximately as χ^2. When X^2 is greater than or equal to the tabulated value of χ^2 for some α, the null hypothesis is rejected at the α level of significance.

Tests of independence and tests of homogeneity are also discussed in this chapter. The tests are mathematically equivalent but conceptually different. Again, these tests essentially test the goodness-of-fit of observed data to expectation under hypotheses, respectively, of independence of two criteria of classifying the data and the homogeneity of proportions among two or more groups.

In addition, we discussed and illustrated in this chapter four other techniques for analyzing frequency data that can be presented in the form of a 2×2 contingency table: the Fisher exact test, the odds ratio, relative risk, and the Mantel–Haenszel procedure. Finally, we discussed the basic concepts of survival analysis and illustrated the computational procedures by means of two examples.

SUMMARY OF FORMULAS FOR CHAPTER 12

Formula Number	Name	Formula
12.2.1	Standard normal random variable	$z_i = \dfrac{y_i - \mu}{\sigma}$
12.2.2	Chi-square distribution with n degrees of freedom	$\chi^2_{(n)} = z_1^2 + z_2^2 + \cdots + z_n^2$
12.2.3	Chi-square probability density function	$f(u) = \dfrac{1}{\left(\dfrac{k}{2} - 1\right)!} \dfrac{1}{2^{k/2}} u^{(k/2)-1} e^{-(u/2)}$
12.2.4	Chi-square test statistic	$\chi^2 = \sum \left[\dfrac{(O_i - E_i)^2}{E_i} \right]$
12.4.1	Chi-square calculation formula for a 2×2 contingency table	$\chi^2 = \dfrac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$
12.4.2	Yates's corrected chi-square calculation for a 2×2 contingency table	$\chi^2_{\text{corrected}} = \dfrac{n(\lvert ad - bc \rvert - .5n)^2}{(a + c)(b + d)(a + b)(c + d)}$
12.6.1–12.6.2	Large-sample approximation to the chi-square	$z = \dfrac{(a/A) - (b/B)}{\sqrt{\hat{p}(1 - \hat{p})(1/A + 1/B)}}$ where $\hat{p} = (a + b)/(A + B)$
12.7.1	Relative risk estimate	$\widehat{RR} = \dfrac{a/(a + b)}{c/(c + d)}$
12.7.2	Confidence interval for the relative risk estimate	$100(1 - \alpha)\%CI = \widehat{RR}^{1 \pm \left(z_a / \sqrt{\chi^2}\right)}$
12.7.3	Odds ratio estimate	$\widehat{OR} = \dfrac{a/b}{c/d} = \dfrac{ad}{bc}$
12.7.4	Confidence interval for the odds ratio estimate	$100(1 - \alpha)\%CI = \widehat{OR}^{1 \pm \left(z_a / \sqrt{\chi^2}\right)}$

(*Continued*)

12.7.5	Expected frequency in the Mantel–Haenszel statistic	$e_i = \dfrac{(a_i + b_i)(a_i + c_i)}{n_i}$
12.7.6	Stratum expected frequency in the Mantel–Haenszel statistic	$v_i = \dfrac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)}$
12.7.7	Mantel–Haenszel test statistic	$\chi^2_{\text{MH}} = \dfrac{\left(\sum_{i=1}^{k} a_i - \sum_{i=1}^{k} e_i\right)}{\sum_{i=1}^{k} v_i}$
12.7.8	Mantel–Haenszel estimator of the common odds ratio	$\widehat{OR}_{\text{MH}} = \dfrac{\sum_{i=1}^{k} (a_i d_i / n_i)}{\sum_{i=1}^{k} (b_i c_i / n_i)}$
Symbol Key	• a, b, c, d = cell frequencies in a 2×2 contingency table • A, B = row totals in the 2×2 contingency table • β = regression coefficient • $\chi^2 \left(\text{or } X^2\right)$ = chi-square • e_i = expected frequency in the Mantel–Haenszel statistic • E_i = expected frequency • $E_{(y\|x)}$ = expected value of y at x • k = degrees of freedom in the chi-square distribution • μ = mean • O_i = observed frequency • $\widehat{OR}$ = odds ratio estimate • σ = standard deviation • $\widehat{RR}$ = relative risk estimate • v_i = stratum expected frequency in the Mantel–Haenszel statistic • y_i = data value at point i • z = normal variate	

REVIEW QUESTIONS AND EXERCISES

1. Explain how the chi-square distribution may be derived.
2. What are the mean and variance of the chi-square distribution?
3. Explain how the degrees of freedom are computed for the chi-square goodness-of-fit tests.
4. State Cochran's rule for small expected frequencies in goodness-of-fit tests.
5. How does one adjust for small expected frequencies?
6. What is a contingency table?
7. How are the degrees of freedom computed when an X^2 value is computed from a contingency table?

8. Explain the rationale behind the method of computing the expected frequencies in a test of independence.

9. Explain the difference between a test of independence and a test of homogeneity.

10. Explain the rationale behind the method of computing the expected frequencies in a test of homogeneity.

11. When do researchers use the Fisher exact test rather than the chi-square test?

12. Define the following:

(a) Observational study
(b) Risk factor
(c) Outcome
(d) Retrospective study
(e) Prospective study
(f) Relative risk
(g) Odds
(h) Odds ratio
(i) Confounding variable

13. Under what conditions is the Mantel–Haenszel test appropriate?

14. Explain how researchers interpret the following measures:

(a) Relative risk
(b) Odds ratio
(c) Mantel–Haenszel common odds ratio

15. In a study of violent victimization of women and men, Porcerelli et al. (A-23) collected information from 679 women and 345 men ages 18 to 64 years at several family practice centers in the metropolitan Detroit area. Patients filled out a health history questionnaire that included a question about victimization. The following table shows the sample subjects cross-classified by gender and the type of violent victimization reported. The victimization categories are defined as no victimization, partner victimization (and not by others), victimization by a person other than a partner (friend, family member, or stranger), and those who reported multiple victimization.

Gender	No Victimization	Partner	Nonpartner	Multiple	Total
Women	611	34	16	18	679
Men	308	10	17	10	345
Total	919	44	33	28	1024

Source: John H. Porcerelli, Rosemary Cogan, Patricia P. West, Edward A. Rose, Dawn Lambrecht, Karen E. Wilson, Richard K. Severson, and Dunia Karana, "Violent Victimization of Women and Men: Physical and Psychiatric Symptoms," *Journal of the American Board of Family Practice, 16* (2003), 32–39.

Can we conclude on the basis of these data that victimization status and gender are not independent? Let $\alpha = .05$.

16. Refer to Exercise 15. The following table shows data reported by Porcerelli et al. for 644 African-American and Caucasian women. May we conclude on the basis of these data that for women, race and victimization status are not independent? Let $\alpha = .05$.

	No Victimization	Partner	Nonpartner	Multiple	Total
Caucasian	356	20	3	9	388
African-American	226	11	10	9	256
Total	582	31	13	18	644

Source: John H. Porcerelli, Rosemary Cogan, Patricia P. West, Edward A. Rose, Dawn Lambrecht, Karen E. Wilson, Richard K. Severson, and Dunia Karana, "Violent Victimization of Women and Men: Physical and Psychiatric Symptoms," *Journal of the American Board of Family Practice, 16* (2003), 32–39.

17. A sample of 150 chronic carriers of a certain antigen and a sample of 500 noncarriers revealed the following blood group distributions:

Blood Group	Carriers	Noncarriers	Total
0	72	230	302
A	54	192	246
B	16	63	79
AB	8	15	23
Total	150	500	650

Can one conclude from these data that the two populations from which the samples were drawn differ with respect to blood group distribution? Let $\alpha = .05$. What is the p value for the test?

18. The following table shows 200 males classified according to social class and headache status:

	Social Class			
Headache Group	A	B	C	Total
No headache (in previous year)	6	30	22	58
Simple headache	11	35	17	63
Unilateral headache (nonmigraine)	4	19	14	37
Migraine	5	25	12	42
Total	26	109	65	200

Do these data provide sufficient evidence to indicate that headache status and social class are related? Let $\alpha = .05$. What is the p value for this test?

19. The following is the frequency distribution of scores made on an aptitude test by 175 applicants to a physical therapy training facility ($\bar{x} = 39.71$, $s = 12.92$).

Score	Number of Applicants	Score	Number of Applicants
10–14	3	40–44	28
15–19	8	45–49	20
20–24	13	50–54	18
25–29	17	55–59	12

(*Continued*)

Score	Number of Applicants	Score	Number of Applicants
30–34	19	60–64	8
35–39	25	65–69	4
Total			175

Do these data provide sufficient evidence to indicate that the population of scores is not normally distributed? Let $\alpha = .05$. What is the p value for this test?

20. A local health department sponsored a venereal disease (VD) information program that was open to high-school juniors and seniors who ranged in age from 16 to 19 years. The program director believed that each age level was equally interested in knowing more about VD. Since each age level was about equally represented in the area served, she felt that equal interest in VD would be reflected by equal age-level attendance at the program. The age breakdown of those attending was as follows:

Age	Number Attending
16	26
17	50
18	44
19	40

Are these data incompatible with the program director's belief that students in the four age levels are equally interested in VD? Let $\alpha = .05$. What is the p value for this test?

21. A survey of children under 15 years of age residing in the inner-city area of a large city were classified according to ethnic group and hemoglobin level. The results were as follows:

	Hemoglobin Level (g/100 ml)			
Ethnic Group	10.0 or Greater	9.0–9.9	< 9.0	Total
A	80	100	20	200
B	99	190	96	385
C	70	30	10	110
Total	249	320	126	695

Do these data provide sufficient evidence to indicate, at the .05 level of significance, that the two variables are related? What is the p value for this test?

22. A sample of reported cases of mumps in preschool children showed the following distribution by age:

Age (Years)	Number of Cases
Under 1	6
1	20
2	35
3	41
4	48
Total	150

Test the hypothesis that cases occur with equal frequency in the five age categories. Let $\alpha = .05$. What is the p value for this test?

23. Each of a sample of 250 men drawn from a population of suspected joint disease victims was asked which of three symptoms bother him most. The same question was asked of a sample of 300 suspected women joint disease victims. The results were as follows:

Most Bothersome Symptom	Men	Women
Morning stiffness	111	102
Nocturnal pain	59	73
Joint swelling	80	125
Total	250	300

Do these data provide sufficient evidence to indicate that the two populations are not homogeneous with respect to major symptoms? Let $\alpha = .05$. What is the p value for this test?

For each of the Exercises 24 through 34, indicate whether a null hypothesis of homogeneity or a null hypothesis of independence is appropriate.

24. A researcher wishes to compare the status of three communities with respect to immunity against polio in preschool children. A sample of preschool children was drawn from each of the three communities.

25. In a study of the relationship between smoking and respiratory illness, a random sample of adults were classified according to consumption of tobacco and extent of respiratory symptoms.

26. A physician who wished to know more about the relationship between smoking and birth defects studies the health records of a sample of mothers and their children, including stillbirths and spontaneously aborted fetuses where possible.

27. A health research team believes that the incidence of depression is higher among people with hypoglycemia than among people who do not suffer from this condition.

28. In a simple random sample of 200 patients undergoing therapy at a drug abuse treatment center, 60 percent belonged to ethnic group I. The remainder belonged to ethnic group II. In ethnic group I, 60 were being treated for alcohol abuse (A), 25 for marijuana abuse (B), and 20 for abuse of heroin, illegal methadone, or some other opioid (C). The remainder had abused barbiturates, cocaine, amphetamines, hallucinogens, or some other nonopioid besides marijuana (D). In ethnic group II the abused drug category and the numbers involved were as follows:

$$\text{A}(28) \qquad \text{B}(32) \quad \text{C}(13) \qquad \text{D (the remainder)}$$

Can one conclude from these data that there is a relationship between ethnic group and choice of drug to abuse? Let $\alpha = .05$ and find the p value.

29. Solar keratoses are skin lesions commonly found on the scalp, face, backs of hands, forearms, ears, scalp, and neck. They are caused by long-term sun exposure, but they are not skin cancers. Chen et al. (A-24) studied 39 subjects randomly assigned (with a 3 to 1 ratio) to imiquimod cream and a control cream. The criterion for effectiveness was having 75 percent or more of the lesion area cleared after 14 weeks of treatment. There were 21 successes among 29 imiquimod-treated subjects and three successes among 10 subjects using the control cream. The researchers used Fisher's exact test and obtained a p value of .027. What are the variables involved? Are the variables quantitative or qualitative? What null and alternative hypotheses are appropriate? What are your conclusions?

30. Janardhan et al. (A-25) examined 125 patients who underwent surgical or endovascular treatment for intracranial aneurysms. At 30 days postprocedure, 17 subjects experienced transient/persistent neurological deficits. The researchers performed logistic regression and found that the 95 percent confidence interval for the odds ratio for aneurysm size was .09–.96. Aneurysm size was dichotomized as less than 13 mm and greater than or equal to 13 mm. The larger tumors indicated higher odds of deficits. Describe the variables as to whether they are continuous, discrete, quantitative, or qualitative. What conclusions may be drawn from the given information?

31. In a study of smoking cessation by Gold et al. (A-26), 189 subjects self-selected into three treatments: nicotine patch only (NTP), Bupropion SR only (B), and nicotine patch with Bupropion SR (NTP + B). Subjects were grouped by age into younger than 50 years old, between 50 and 64, and 65 and older. There were 15 subjects younger than 50 years old who chose NTP, 26 who chose B, and 16 who chose NTP + B. In the 50–64 years category, six chose NTP, 54 chose B, and 40 chose NTP + B. In the oldest age category, six chose NTP, 21 chose B, and five chose NTP + B. What statistical technique studied in this chapter would be appropriate for analyzing these data? Describe the variables involved as to whether they are continuous, discrete, quantitative, or qualitative. What null and alternative hypotheses are appropriate? If you think you have sufficient information, conduct a complete hypothesis test. What are your conclusions?

32. Kozinszky and Bártai (A-27) examined contraceptive use by teenage girls requesting abortion in Szeged, Hungary. Subjects were classified as younger than 20 years old or 20 years old or older. Of the younger than 20-year-old women, 146 requested an abortion. Of the older group, 1054 requested an abortion. A control group consisted of visitors to the family planning center who did not request an abortion or persons accompanying women who requested an abortion. In the control group, there were 147 women under 20 years of age and 1053 who were 20 years or older. One of the outcome variables of interest was knowledge of emergency contraception. The researchers report that, "Emergency contraception was significantly [(Mantel–Haenszel) $p < .001$] less well known among the would-be aborter teenagers as compared to the older women requesting artificial abortion (OR = .07) than the relevant knowledge of the teenage controls (OR = .10)." Explain the meaning of the reported statistics. What are your conclusions based on the given information?

33. The goal of a study by Crosignani et al. (A-28) was to assess the effect of road traffic exhaust on the risk of childhood leukemia. They studied 120 children in Northern Italy identified through a population-based cancer registry (cases). Four controls per case, matched by age and gender, were sampled from population files. The researchers used a diffusion model of benzene to estimate exposure to traffic exhaust. Compared to children whose homes were not exposed to road traffic emissions, the rate of childhood leukemia was significantly higher for heavily exposed children. Characterize this study as to whether it is observational, prospective, or retrospective. Describe the variables as to whether they are continuous, discrete, quantitative, qualitative, a risk factor, or a confounding variable. Explain the meaning of the reported results. What are your conclusions based on the given information?

34. Gallagher et al. (A-29) conducted a descriptive study to identify factors that influence women's attendance at cardiac rehabilitation programs following a cardiac event. One outcome variable of interest was actual attendance at such a program. The researchers enrolled women discharged from four metropolitan hospitals in Sydney, Australia. Of 183 women, only 57 women actually attended programs. The authors reported odds ratios and confidence intervals on the following variables that significantly affected outcome: age-squared (1.72; 1.10–2.70). Women over the age of 70 had the lowest odds, while women ages 55–70 years had the highest odds.), perceived control (.92; .85–1.00), employment (.20; .07–.58), diagnosis (6.82, 1.84–25.21, odds ratio was higher for women who experienced coronary artery bypass grafting vs. myocardial infarction), and stressful event (.21, .06–.73). Characterize this study as to whether it is observational, prospective, or retrospective. Describe the

variables as to whether they are continuous, discrete, quantitative, qualitative, a risk factor, or a confounding variable. Explain the meaning of the reported odds ratios.

For each of the Exercises 35 through 51, do as many of the following as you think appropriate:

(a) Apply one or more of the techniques discussed in this chapter.
(b) Apply one or more of the techniques discussed in previous chapters.
(c) Construct graphs.
(d) Construct confidence intervals for population parameters.
(e) Formulate relevant hypotheses, perform the appropriate tests, and find p values.
(f) State the statistical decisions and clinical conclusions that the results of your hypothesis tests justify.
(g) Describe the population(s) to which you think your inferences are applicable.
(h) State the assumptions necessary for the validity of your analyses.

35. In a prospective, randomized, double-blind study, Stanley et al. (A-30) examined the relative efficacy and side effects of morphine and pethidine, drugs commonly used for patient-controlled analgesia (PCA). Subjects were 40 women, between the ages of 20 and 65 years, undergoing total abdominal hysterectomy. Patients were allocated randomly to receive morphine or pethidine by PCA. At the end of the study, subjects described their appreciation of nausea and vomiting, pain, and satisfaction by means of a three-point verbal scale. The results were as follows:

	Satisfaction			
Drug	**Unhappy/ Miserable**	**Moderately Happy**	**Happy/ Delighted**	**Total**
Pethidine	5	9	6	20
Morphine	9	9	2	20
Total	14	18	8	40

	Pain			
Drug	**Unbearable/ Severe**	**Moderate**	**Slight/ None**	**Total**
Pethidine	2	10	8	20
Morphine	2	8	10	20
Total	4	18	18	40

	Nausea			
Drug	**Unbearable/ Severe**	**Moderate**	**Slight/ None**	**Total**
Pethidine	5	9	6	20
Morphine	7	8	5	20
Total	12	17	11	40

Source: Data provided courtesy of Dr. Balraj L. Appadu.

36. Screening data from a statewide lead poisoning prevention program between April 1990 and March 1991 were examined by Sargent et al. (A-31) in an effort to learn more about community risk factors for iron deficiency in young children. Study subjects ranged in age between 6 and 59 months. Among 1860 children with Hispanic surnames, 338 had iron deficiency. Four-hundred-fifty-seven of 1139 with Southeast Asian surnames and 1034 of 8814 children with other surnames had iron deficiency.

37. To increase understanding of HIV-infection risk among patients with severe mental illness, Horwath et al. (A-32) conducted a study to identify predictors of injection drug use among patients who did not have a primary substance use disorder. Of 192 patients recruited from inpatient and outpatient public psychiatric facilities, 123 were males. Twenty-nine of the males and nine of the females were found to have a history of illicit-drug injection.

38. Skinner et al. (A-33) conducted a clinical trial to determine whether treatment with melphalan, prednisone, and colchicine (MPC) is superior to colchicine (C) alone. Subjects consisted of 100 patients with primary amyloidosis. Fifty were treated with C and 50 with MPC. Eighteen months after the last person was admitted and 6 years after the trial began, 44 of those receiving C and 36 of those receiving MPC had died.

39. The purpose of a study by Miyajima et al. (A-34) was to evaluate the changes of tumor cell contamination in bone marrow (BM) and peripheral blood (PB) during the clinical course of patients with advanced neuroblastoma. Their procedure involved detecting tyrosine hydroxylase (TH) mRNA to clarify the appropriate source and time for harvesting hematopoietic stem cells for transplantation. The authors used Fisher's exact test in the analysis of their data. If available, read their article and decide if you agree that Fisher's exact text was the appropriate technique to use. If you agree, duplicate their procedure and see if you get the same results. If you disagree, explain why.

40. Cohen et al. (A-35) investigated the relationship between HIV seropositivity and bacterial vaginosis in a population at high risk for sexual acquisition of HIV. Subjects were 144 female commercial sex workers in Thailand of whom 62 were HIV-positive and 109 had a history of sexually transmitted diseases (STD). In the HIV-negative group, 51 had a history of STD.

41. The purpose of a study by Lipschitz et al. (A-36) was to examine, using a questionnaire, the rates and characteristics of childhood abuse and adult assaults in a large general outpatient population. Subjects consisted of 120 psychiatric outpatients (86 females, 34 males) in treatment at a large hospital-based clinic in an inner-city area. Forty-seven females and six males reported incidents of childhood sexual abuse.

42. Subjects of a study by O'Brien et al. (A-37) consisted of 100 low-risk patients having well-dated pregnancies. The investigators wished to evaluate the efficacy of a more gradual method for promoting cervical change and delivery. Half of the patients were randomly assigned to receive a placebo, and the remainder received 2 mg of intravaginal prostaglandin E_2 (PGE_2) for 5 consecutive days. One of the infants born to mothers in the experimental group and four born to those in the control group had macrosomia.

43. The purposes of a study by Adra et al. (A-38) were to assess the influence of route of delivery on neonatal outcome in fetuses with gastroschisis and to correlate ultrasonographic appearance of the fetal bowel with immediate postnatal outcome. Among 27 cases of prenatally diagnosed gastroschisis the ultrasonograph appearance of the fetal bowel was normal in 15. Postoperative complications were observed in two of the 15 and in seven of the cases in which the ultrasonographic appearance was not normal.

44. Liu et al. (A-39) conducted household surveys in areas of Alabama under tornado warnings. In one of the surveys (survey 2) the mean age of the 193 interviewees was 54 years. Of these 56.0 percent were

women, 88.6 percent were white, and 83.4 percent had a high-school education or higher. Among the information collected were data on shelter-seeking activity and understanding of the term "tornado warning." One-hundred-twenty-eight respondents indicated that they usually seek shelter when made aware of a tornado warning. Of these, 118 understood the meaning of tornado warning. Forty-six of those who said they didn't usually seek shelter understood the meaning of the term.

45. The purposes of a study by Patel et al. (A-40) were to investigate the incidence of acute angle-closure glaucoma secondary to pupillary dilation and to identify screening methods for detecting angles at risk of occlusion. Of 5308 subjects studied, 1287 were 70 years of age or older. Seventeen of the older subjects and 21 of the younger subjects (40 through 69 years of age) were identified as having potentially occludable angles.

46. Voskuyl et al. (A-41) investigated those characteristics (including male gender) of patients with rheumatoid arthritis (RA) that are associated with the development of rheumatoid vasculitis (RV). Subjects consisted of 69 patients who had been diagnosed as having RV and 138 patients with RA who were not suspected to have vasculitis. There were 32 males in the RV group and 38 among the RA patients.

47. Harris et al. (A-42) conducted a study to compare the efficacy of anterior colporrhaphy and retropubic urethropexy performed for genuine stress urinary incontinence. The subjects were 76 women who had undergone one or the other surgery. Subjects in each group were comparable in age, social status, race, parity, and weight. In 22 of the 41 cases reported as cured the surgery had been performed by attending staff. In 10 of the failures, surgery had been performed by attending staff. All other surgeries had been performed by resident surgeons.

48. Kohashi et al. (A-43) conducted a study in which the subjects were patients with scoliosis. As part of the study, 21 patients treated with braces were divided into two groups, group A($n_A = 12$) and group B($n_B = 9$), on the basis of certain scoliosis progression factors. Two patients in group A and eight in group B exhibited evidence of progressive deformity, while the others did not.

49. In a study of patients with cervical intraepithelial neoplasia, Burger et al. (A-44) compared those who were human papillomavirus (HPV)-positive and those who were HPV-negative with respect to risk factors for HPV infection. Among their findings were 60 out of 91 nonsmokers with HPV infection and 44 HPV-positive patients out of 50 who smoked 21 or more cigarettes per day.

50. Thomas et al. (A-45) conducted a study to determine the correlates of compliance with follow-up appointments and prescription filling after an emergency department visit. Among 235 respondents, 158 kept their appointments. Of these, 98 were females. Of those who missed their appointments, 31 were males.

51. The subjects of a study conducted by O'Keefe and Lavan (A-46) were 60 patients with cognitive impairment who required parenteral fluids for at least 48 hours. The patients were randomly assigned to receive either intravenous (IV) or subcutaneous (SC) fluids. The mean age of the 30 patients in the SC group was 81 years with a standard deviation of 6. Fifty-seven percent were females. The mean age of the IV group was 84 years with a standard deviation of 7. Agitation related to the cannula or drip was observed in 11 of the SC patients and 24 of the IV patients.

Exercises for Use with the Large Data Sets Available on the Following Website: www.wiley.com/college/daniel

1. Refer to the data on smoking, alcohol consumption, blood pressure, and respiratory disease among 1200 adults (SMOKING). The variables are as follows:

Sex (A): 1 = male, 0 = female
Smoking status (B): 0 = nonsmoker, 1 = smoker
Drinking level (C): 0 = nondrinker
1 = light to moderate drinker
2 = heavy drinker
Symptoms of respiratory disease (D): 1 = present, 0 = absent
High blood pressure status (E): 1 = present, 0 = absent

Select a simple random sample of size 100 from this population and carry out an analysis to see if you can conclude that there is a relationship between smoking status and symptoms of respiratory disease. Let $\alpha = .05$ and determine the p value for your test. Compare your results with those of your classmates.

2. Refer to Exercise 1. Select a simple random sample of size 100 from the population and carry out a test to see if you can conclude that there is a relationship between drinking status and high blood pressure status in the population. Let $\alpha = .05$ and determine the p value. Compare your results with those of your classmates.

3. Refer to Exercise 1. Select a simple random sample of size 100 from the population and carry out a test to see if you can conclude that there is a relationship between gender and smoking status in the population. Let $\alpha = .05$ and determine the p value. Compare your results with those of your classmates.

4. Refer to Exercise 1. Select a simple random sample of size 100 from the population and carry out a test to see if you can conclude that there is a relationship between gender and drinking level in the population. Let $\alpha = .05$ and find the p value. Compare your results with those of your classmates.

REFERENCES

Methodology References

1. KARL PEARSON, "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such that It Can Be Reasonably Supposed to Have Arisen from Random Sampling," *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, Fifth Series, *50* (1900), 157–175. Reprinted in *Karl Pearson's Early Statistical Papers*, Cambridge University Press, 1948.
2. H. O. LANCASTER, *The Chi-Squared Distribution*, Wiley, New York, 1969.
3. MIKHAIL S. NIKULIN and PRISCILLA E. GREENWOOD, *A Guide to Chi-Squared Testing*, Wiley, New York, 1996.
4. WILLIAM G. COCHRAN, "The χ^2 Test of Goodness of Fit," *Annals of Mathematical Statistics*, *23* (1952), 315–345.
5. WILLIAM G. COCHRAN, "Some Methods for Strengthening the Common χ^2 Tests," *Biometrics*, *10* (1954), 417–451.
6. F. YATES, "Contingency Tables Involving Small Numbers and the χ^2 Tests," *Journal of the Royal Statistical Society, Supplement*, *1*, 1934 (Series B), 217–235.
7. R. A. FISHER, *Statistical Methods for Research Workers*, Fifth Edition, Oliver and Boyd, Edinburgh, 1934.
8. R. A. FISHER, "The Logic of Inductive Inference," *Journal of the Royal Statistical Society Series A*, *98* (1935), 39–54.
9. J. O. IRWIN, "Tests of Significance for Differences Between Percentages Based on Small Numbers," *Metron*, *12* (1935), 83–94.
10. F. YATES, "Contingency Tables Involving Small Numbers and the χ^2 Test," *Journal of the Royal Statistical Society, Supplement*, *1*, (1934), 217–235.
11. D. J. FINNEY, "The Fisher-Yates Test of Significance in 2×2 Contingency Tables," *Biometrika*, *35* (1948), 145–156.

12. R. Latscha, "Tests of Significance in a 2 × 2 Contingency Table: Extension of Finney's Table," *Biometrika*, *40* (1955), 74–86.
13. G. A. Barnard, "A New Test for 2 × 2 Tables," *Nature*, *156* (1945), 117.
14. G. A. Barnard, "A New Test for 2 × 2 Tables," *Nature*, *156* (1945), 783–784.
15. G. A. Barnard, "Significance Tests for 2 × 2 Tables," *Biometrika*, *34* (1947), 123–138.
16. R. A. Fisher, "A New Test for 2 × 2 Tables," *Nature*, *156* (1945), 388.
17. E. S. Pearson, "The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2 × 2 Table," *Biometrika*, *34* (1947), 139–167.
18. A. Sweetland, "A Comparison of the Chi-Square Test for 1 df and the Fisher Exact Test," Rand Corporation, Santa Monica, CA, 1972.
19. Wendell E. Carr, "Fisher's Exact Text Extended to More than Two Samples of Equal Size," *Technometrics*, *22* (1980), 269–270.
20. Henry R. Neave, "A New Look at an Old Test," *Bulletin of Applied Statistics*, *9* (1982), 165–178.
21. William D. Dupont, "Sensitivity of Fisher's Exact Text to Minor Perturbations in 2 × 2 Contingency Tables," *Statistics in Medicine*, *5* (1986), 629–635.
22. N. Mantel and W. Haenszel, "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *Journal of the National Cancer Institute*, *22* (1959), 719–748.
23. N. Mantel, "Chi-Square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure," *Journal of the American Statistical Association*, *58* (1963), 690–700.

Applications References

A-1. Carole W. Cranor and Dale B. Christensen, "The Asheville Project: Short-Term Outcomes of a Community Pharmacy Diabetes Care Program," *Journal of the American Pharmaceutical Association*, *43* (2003), 149–159.
A-2. Amy L. Byers, Heather Allore, Thomas M. Gill, and Peter N. Peduzzi, "Application of Negative Binomial Modeling for Discrete Outcomes: A Case Study in Aging Research," *Journal of Clinical Epidemiology*, *56* (2003), 559–564.
A-3. Kathleen M. Stepanuk, Jorge E. Tolosa, Dawneete Lewis, Victoria Meyers, Cynthia Royds, Juan Carlos Saogal, and Ron Librizzi, "Folic Acid Supplementation Use Among Women Who Contact a Teratology Information Service," *American Journal of Obstetrics and Gynecology*, *187* (2002), 964–967.
A-4. J. K. Silver and D. D. Aiello, "Polio Survivors: Falls and Subsequent Injuries," *American Journal of Physical Medicine and Rehabilitation*, *81* (2002), 567–570.
A-5. Cynthia G. Segal and Jacqueline J. Anderson, "Preoperative Skin Preparation of Cardiac Patients," *AORN Journal*, *76* (2002), 821–827.
A-6. Ralph Rothenberg and John P. Holcomb, "Guidelines for Monitoring of NSAIDs: Who Listened?," *Journal of Clinical Rheumatology*, *6* (2000), 258–265.
A-7. Sharon M. Boles and Patrick B. Johnson, "Gender, Weight Concerns, and Adolescent Smoking," *Journal of Addictive Diseases*, *20* (2001), 5–14.
A-8. The DMG Study Group, "Migraine and Idiopathic Narcolepsy—A Case-Control Study," *Cephalagia*, *23* (2003), 786–789.
A-9. Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research*, *37* (2003), 297–303.
A-10. Steven S. Coughlin, Robert J. Uhler, Thomas Richards, and Katherine M. Wilson, "Breast and Cervical Cancer Screening Practices Among Hispanic and Non-Hispanic Women Residing Near the United States–Mexico Border, 1999-2000," *Family and Community Health*, *26* (2003), 130–139.
A-11. Robert Swor, Scott Compton, Fern Vining, Lynn Ososky Farr, Sue Kokko, Rebecca Pascual, and Raymond E. Jackson, "A Randomized Controlled Trial of Chest Compression Only CPR for Older Adults: A Pilot Study," *Resuscitation*, *58* (2003), 177–185.
A-12. U. S. Justesen, A. M. Lervfing, A. Thomsen, J. A. Lindberg, C. Pedersen, and P. Tauris, "Low-Dose Indinavir in Combination with Low-Dose Ritonavir: Steady-State Pharmacokinetics and Long-Term Clinical Outcome Follow-Up," *HIV Medicine*, *4* (2003), 250–254.
A-13. J. F. Tahmassebi and M. E. J. Curzon, "The Cause of Drooling in Children with Cerebral Palsy—Hypersalivation or Swallowing Defect?" *International Journal of Paediatric Dentistry*, *13* (2003), 106–111.
A-14. Shu Dong Xiao and Tong Shi, "Is Cranberry Juice Effective in the Treatment and Prevention of *Helicobacter Pylori* Infection of Mice?," *Chinese Journal of Digestive Diseases*, *4* (2003), 136–139.

A-15. Gad Shaked, Oleg Kleiner, Robert Finally, Jacob Mordechai, Nitza Newman, and Zahavi Cohen, "Management of Blunt Pancreatic Injuries in Children," *European Journal of Trauma*, *29* (2003), 151–155.

A-16. Everett F. Magann, Sharon F. Evans, Beth Weitz, and John Newnham, "Antepartum, Intrapartum, and Neonatal Significance of Exercise on Healthy Low-Risk Pregnant Working Women," *Obstetrics and Gynecology*, *99* (2002), 466–472.

A-17. A. M. Toschke, S. M. Montgomery, U. Pfeiffer, and R.von Kries, "Early Intrauterine Exposure to Tobacco-Inhaled Products and Obesity," *American Journal of Epidemiology*, *158* (2003), 1068–1074.

A-18. Daniel H. LaMont, Matthew J. Budoff, David M. Shavelle, Robert Shavelle, Bruce H. Brundage, and James M. Hagar, "Coronary Calcium Scanning Adds Incremental Value to Patients with Positive Stress Tests," *American Heart Journal*, *143* (2002), 861–867.

A-19. Margaret L. J. Davy, Tom J. Dodd, Colin G. Luke, and David M. Roder, "Cervical Cancer: Effect of Glandular Cell Type on Prognosis, Treatment, and Survival," *Obstetrics and Gynecology*, *101* (2003), 38–45.

A-20. U. Stenestrand and L. Wallentin, "Early Revascularization and 1-Year Survival in 14-Day Survivors of Acute Myocardial Infarction," *Lancet*, *359* (2002), 1805–1811.

A-21. Takako Sugiyama, Kumiya Sugiyama, Masao Toda, Tastuo Yukawa, Sohei Makino, and Takeshi Fukuda, "Risk Factors for Asthma and Allergic Diseases Among 13-14-Year-Old Schoolchildren in Japan," *Allergology International*, *51* (2002), 139–150.

A-22. D. Holben, M. C. McClincy, J. P. Holcomb, and K. L. Dean, "Food Security Status of Households in Appalachian Ohio with Children in Head Start," *Journal of American Dietetic Association*, *104* (2004), 238–241.

A-23. John H. Porcerelli, Rosemary Cogan, Patricia P. West, Edward A. Rose, Dawn Lambrecht, Karen E. Wilson, Richard K. Severson, and Dunia Karana, "Violent Victimization of Women and Men: Physical and Psychiatric Symptoms," *Journal of the American Board of Family Practice*, *16* (2003), 32–39.

A-24. Keng Chen, Lee Mei Yap, Robin Marks, and Stephen Shumack, "Short-Course Therapy with Imiquimod 5% Cream for Solar Keratoses: A Randomized Controlled Trial," *Australasian Journal of Dermatology*, *44* (2003), 250–255.

A-25. Vallabh Janardhan, Robert Friedlander, Howard Riina, and Philip Edwin Stieg, "Identifying Patients at Risk for Postprocedural Morbidity After Treatment of Incidental Intracranial Aneurysms: The Role of Aneurysm Size and Location," *Neurosurgical Focus*, *13* (2002), 1–8.

A-26. Paul B. Gold, Robert N. Rubey, and Richard T. Harvey, "Naturalistic, Self-Assignment Comparative Trial of Bupropion SR, a Nicotine Patch, or Both for Smoking Cessation Treatment in Primary Care," *American Journal on Addictions*, *11* (2002), 315–331.

A-27. Zoltán Kozinszky and György Bártai, "Contraceptive Behavior of Teenagers Requesting Abortion," *European Journal of Obstetrics and Gynecology and Reproductive Biology*, *112* (2004), 80–83.

A-28. Paolo Crosignani, Andrea Tittarelli, Alessandro Borgini, Tiziana Codazzi, Adriano Rovelli, Emma Porro, Paolo Contiero, Nadia Bianchi, Giovanna Tagliabue, Rosaria Fissi, Francesco Rossitto, and Franco Berrino, "Childhood Leukemia and Road Traffic: A Population-Based Case-Control Study," *International Journal of Cancer*, *108* (2004), 596–599.

A-29. Robyn Gallagher, Sharon McKinley, and Kathleen Dracup, "Predictors of Women's Attendance at Cardiac Rehabilitation Programs," *Progress in Cardiovascular Nursing*, *18* (2003), 121–126.

A-30. G. Stanley, B. Appadu, M. Mead, and D. J. Rowbotham, "Dose Requirements, Efficacy and Side Effects of Morphine and Pethidine Delivered by Patient-Controlled Analgesia After Gynaecological Surgery," *British Journal of Anaesthesia*, *76* (1996), 484–486.

A-31. James D. Sargent, Therese A. Stukel, Madeline A. Dalton, Jean L. Freeman, and Mary Jean Brown, "Iron Deficiency in Massachusetts Communities: Socioeconomic and Demographic Risk Factors Among Children," *American Journal of Public Health*, *86* (1996), 544–550.

A-32. Ewald Horwath, Francine Cournos, Karen McKinnon, Jeannine R. Guido, and Richard Herman, "Illicit-Drug Injection Among Psychiatric Patients Without a Primary Substance Use Disorder," *Psychiatric Services*, *47* (1996), 181–185.

A-33. Martha Skinner, Jennifer J. Anderson, Robert Simms, Rodney Falk, Ming Wang, Caryn A. Libbey, Lee Anna Jones, and Alan S. Cohen, "Treatment of 100 Patients with Primary Amyloidosis: A Randomized Trial of Melphalan, Prednisone, and Colchicine Versus Colchicine Only," *American Journal of Medicine*, *100* (1996), 290–298.

A-34. Yuji Miyajima, Keizo Horibe, Minoru Fukuda, Kimikazu Matsumoto, Shin-ichiro Numata, Hiroshi Mori, and Koji Kato, "Sequential Detection of Tumor Cells in the Peripheral Blood and Bone Marrow of Patients with Stage IV Neuroblastoma by the Reverse Transcription-Polymerase Chain Reaction for Tyrosine Hydroxylase mRNA," *Cancer*, *77* (1996), 1214–1219.

A-35. Craig R. Cohen, Ann Duerr, Niwat Pruithithada, Sungwal Rugpao, Sharon Hillier, Patricia Garcia, and Kenrad Nelson, "Bacterial Vaginosis and HIV Seroprevalence Among Female Commercial Sex Workers in Chiang Mai, Thailand," *AIDS*, *9* (1995), 1093–1097.

A-36. Deborah S. Lipschitz, Margaret L. Kaplan, Jodie B. Sorkenn, Gianni L. Faedda, Peter Chorney, and Gregory M. Asnis, "Prevalence and Characteristics of Physical and Sexual Abuse Among Psychiatric Outpatients," *Psychiatric Services*, *47* (1996), 189–191.

A-37. John M. O'Brien, Brian M. Mercer, Nancy T. Cleary, and Baha M. Sibai, "Efficacy of Outpatient Induction with Low-Dose Intravaginal Prostaglandin E_2: A Randomized, Double-Blind, Placebo-Controlled Trial," *American Journal of Obstetrics and Gynecology*, *173* (1995), 1855–1859.

A-38. Abdallah M. Adra, Helain J. Landy, Jaime Nahmias, and Orlando Gómez-Marín, "The Fetus with Gastroschisis: Impact of Route of Delivery and Prenatal Ultrasonography," *American Journal of Obstetrics and Gynecology*, *174* (1996), 540–546.

A-39. Simin Liu, Lynn E. Quenemoen, Josephine Malilay, Eric Noji, Thomas Sinks, and James Mendlein, "Assessment of a Severe-Weather Warning System and Disaster Preparedness, Calhoun Country, Alabama, 1994," *American Journal of Public Health*, *86* (1996), 87–89.

A-40. Ketan H. Patel, Jonathan C. Javitt, James M. Tielsch, Debra A. Street, Joanne Katz, Harry A. Quigley, and Alfred Sommer, "Incidence of Acute Angle-Closure Glaucoma After Pharmacologic Mydriasis," *American Journal of Ophthalmology*, *120* (1995), 709–717.

A-41. Alexandre E. Voskuyl, Aeilko H. Zwinderman, Marie Louise Westedt, Jan P. Vandenbroucke, Ferdinand C. Breedveld, and Johanna M. W. Hazes, "Factors Associated with the Development of Vasculitis in Rheumatoid Arthritis: Results of a Case-Control Study," *Annals of the Rheumatic Diseases*, *55* (1996), 190–192.

A-42. Robert L. Harris, Christopher A. Yancey, Winfred L. Wiser, John C. Morrison, and G. Rodney Meeks, "Comparison of Anterior Colporrhaphy and Retropubic Urethropexy for Patients with Genuine Stress Urinary Incontinence," *American Journal of Obstetrics and Gynecology*, *173* (1995), 1671–1675.

A-43. Yoshihiro Kohashi, Masayoshi Oga, and Yoichi Sugioka, "A New Method Using Top Views of the Spine to Predict the Progression of Curves in Idiopathic Scoliosis During Growth," *Spine*, *21* (1996), 212–217.

A-44. M. P. M. Burger, H. Hollema, W. J. L. M. Pieters, F. P. Schröder, and W. G. V. Quint, "Epidemiological Evidence of Cervical Intraepithelial Neoplasia Without the Presence of Human Papillomavirus," *British Journal of Cancer*, *73* (1996), 831–836.

A-45. Eric J. Thomas, Helen R. Burstin, Anne C. O'Neil, E. John Orav, and Troyen A. Brennan, "Patient Noncompliance with Medical Advice After the Emergency Department Visit," *Annals of Emergency Medicine*, *27* (1996), 49–55.

A-46. S. T. O'Keefe and J. N. Lavan, "Subcutaneous Fluids in Elderly Hospital Patients with Cognitive Impairment," *Gerontology*, *42* (1996), 36–39.

CHAPTER 13

NONPARAMETRIC AND DISTRIBUTION-FREE STATISTICS

CHAPTER OVERVIEW

This chapter explores a wide variety of techniques that are useful when the underlying assumptions of traditional hypothesis tests are violated or one wishes to perform a test without making assumptions about the sampled population.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand the rank transformation and how nonparametric procedures can be used for weak measurement scales.

2. be able to calculate and interpret a wide variety of nonparametric tests commonly used in practice.
3. understand which nonparametric tests may be used in place of traditional parametric statistical tests when various test assumptions are violated.

13.1 INTRODUCTION

Most of the statistical inference procedures we have discussed up to this point are classified as *parametric statistics*. One exception is our use of chi-square—as a test of goodness-of-fit and as a test of independence. These uses of chi-square come under the heading of *nonparametric statistics*.

The obvious question now is, "What is the difference?" In answer, let us recall the nature of the inferential procedures that we have categorized as *parametric*. In each case, our interest was focused on estimating or testing a hypothesis about one or more population parameters. Furthermore, central to these procedures was a knowledge of the functional form of the population from which were drawn the samples providing the basis for the inference.

An example of a parametric statistical test is the widely used *t* test. The most common uses of this test are for testing a hypothesis about a single population mean or the difference between two population means. One of the assumptions underlying the valid use of this test is that the sampled population or populations are at least approximately normally distributed.

As we will learn, the procedures that we discuss in this chapter either are not concerned with population parameters or do not depend on knowledge of the sampled population. Strictly speaking, only those procedures that test hypotheses that are not statements about population parameters are classified as *nonparametric,* while those that make no assumption about the sampled population are called *distribution-free* procedures. Despite this distinction, it is customary to use the terms *nonparametric* and *distribution-free* interchangeably and to discuss the various procedures of both types under the heading *nonparametric statistics*. We will follow this convention.

The above discussion implies the following four advantages of nonparametric statistics.

1. They allow for the testing of hypotheses that are not statements about population parameter values. Some of the chi-square tests of goodness-of-fit and the tests of independence are examples of tests possessing this advantage.
2. Nonparametric tests may be used when the form of the sampled population is unknown.
3. Nonparametric procedures tend to be computationally easier and consequently more quickly applied than parametric procedures. This can be a desirable feature in certain cases, but when time is not at a premium, it merits a low priority as a criterion for choosing a nonparametric test. Indeed, most statistical software packages now include a wide variety of nonparametric analysis options, making considerations about computation speed unnecessary.
4. Nonparametric procedures may be applied when the data being analyzed consist merely of rankings or classifications. That is, the data may not be based on a

measurement scale strong enough to allow the arithmetic operations necessary for carrying out parametric procedures. The subject of measurement scales is discussed in more detail in the next section.

Although nonparametric statistics enjoy a number of advantages, their disadvantages must also be recognized.

1. The use of nonparametric procedures with data that can be handled with a parametric procedure results in a waste of data.
2. The application of some of the nonparametric tests may be laborious for large samples.

13.2 MEASUREMENT SCALES

As was pointed out in the previous section, one of the advantages of nonparametric statistical procedures is that they can be used with data that are based on a weak measurement scale. To understand fully the meaning of this statement, it is necessary to know and understand the meaning of measurement and the various measurement scales most frequently used. At this point the reader may wish to refer to the discussion of measurement scales in Chapter 1.

Many authorities are of the opinion that different statistical tests require different measurement scales. Although this idea appears to be followed in practice, there are alternative points of view.

Data based on ranks, as will be discussed in this chapter, are commonly encountered in statistics. We may, for example, simply note the order in which a sample of subjects complete an event instead of the actual time taken to complete it. More often, however, we use a *rank transformation* on the data by replacing, prior to analysis, the original data by their ranks. Although we usually lose some information by employing this procedure (for example, the ability to calculate the mean and variance), the transformed measurement scale allows the computation of most nonparametric statistical procedures. In fact, most of the commonly used nonparametric procedures, including most of those presented in this chapter, can be obtained by first applying the rank transformation and then using the standard parametric procedure on the transformed data instead of on the original data. For example, if we wish to determine whether two independent samples differ, we may employ the independent samples t test if the data are approximately normally distributed. If we cannot make the assumption of normal distributions, we may, as we shall see in the sections that follow, employ an appropriate nonparametric test. In lieu of these procedures, we could first apply the rank transformation on the data and then use the independent samples t test on the ranks. This will provide an equivalent test to the nonparametric test, and is a useful tool to employ if a desired nonparametric test is not available in your available statistical software package.

Readers should also keep in mind that other transformations (e.g., taking the logarithm of the original data) may sufficiently normalize the data such that standard parametric procedures can be used on the transformed data in lieu of using nonparametric methods.

13.3 THE SIGN TEST

The familiar t test is not strictly valid for testing (1) the null hypothesis that a population mean is equal to some particular value, or (2) the null hypothesis that the mean of a population of differences between pairs of measurements is equal to zero unless the relevant populations are at least approximately normally distributed. Case 2 will be recognized as a situation that was analyzed by the paired comparisons test in Chapter 7. When the normality assumptions cannot be made or when the data at hand are ranks rather than measurements on an interval or ratio scale, the investigator may wish for an optional procedure. Although the t test is known to be rather insensitive to violations of the normality assumption, there are times when an alternative test is desirable.

A frequently used nonparametric test that does not depend on the assumptions of the t test is the *sign test*. This test focuses on the median rather than the mean as a measure of central tendency or location. The median and mean will be equal in symmetric distributions. The only assumption underlying the test is that the distribution of the variable of interest is continuous. This assumption rules out the use of nominal data.

The sign test gets its name from the fact that pluses and minuses, rather than numerical values, provide the raw data used in the calculations. We illustrate the use of the sign test, first in the case of a single sample, and then by an example involving paired samples.

EXAMPLE 13.3.1

Researchers wished to know if instruction in personal care and grooming would improve the appearance of mentally retarded girls. In a school for the mentally retarded, 10 girls selected at random received special instruction in personal care and grooming. Two weeks after completion of the course of instruction the girls were interviewed by a nurse and a social worker who assigned each girl a score based on her general appearance. The investigators believed that the scores achieved the level of an ordinal scale. They felt that although a score of, say, 8 represented a better appearance than a score of 6, they were unwilling to say that the difference between scores of 6 and 8 was equal to the difference between, say, scores of 8 and 10; or that the difference between scores of 6 and 8 represented twice as much improvement as the difference between scores of 5 and 6. The scores are shown in Table 13.3.1. We wish to know if we can conclude that the median score of the population from which we assume this sample to have been drawn is different from 5.

TABLE 13.3.1 General Appearance Scores of 10 Mentally Retarded Girls

Girl	Score	Girl	Score
1	4	6	6
2	5	7	10
3	8	8	7
4	8	9	6
5	9	10	6

Solution:

1. **Data.** See problem statement.
2. **Assumptions.** We assume that the measurements are taken on a continuous variable.
3. **Hypotheses.**

   H_0 : The population median is 5.
   H_A : The population median is not 5.

   Let $\alpha = .05$.
4. **Test statistic.** The test statistic for the sign test is either the observed number of plus signs or the observed number of minus signs. The nature of the alternative hypothesis determines which of these test statistics is appropriate. In a given test, any one of the following alternative hypotheses is possible:

   $$\begin{array}{ll} H_A : P(+) > (-) & \text{one-sided alternative} \\ H_A : P(+) < (-) & \text{one-sided alternative} \\ H_A : P(+) \neq 1(-) & \text{two-sided alternative} \end{array}$$

   If the alternative hypothesis is

   $$H_A : P(+) > P(-)$$

   a sufficiently small number of minus signs causes rejection of H_0. The test statistic is the number of minus signs. Similarly, if the alternative hypothesis is

   $$H_A : P(+) < P(-)$$

   a sufficiently small number of plus signs causes rejection of H_0. The test statistic is the number of plus signs. If the alternative hypothesis is

   $$H_A : P(+) \neq P(-)$$

   either a sufficiently small number of plus signs or a sufficiently small number of minus signs causes rejection of the null hypothesis. We may take as the test statistic the less frequently occurring sign.
5. **Distribution of test statistic.** As a first step in determining the nature of the test statistic, let us examine the data in Table 13.3.1 to determine which scores lie above and which ones lie below the hypothesized median of 5. If we assign a plus sign to those scores that lie above the hypothesized median and a minus to those that fall below, we have the results shown in Table 13.3.2.

   If the null hypothesis were true, that is, if the median were, in fact, 5, we would expect the numbers of scores falling above and below 5 to be

TABLE 13.3.2 Scores Above (+) and Below (−) the Hypothesized Median Based on Data of Example 13.3.1

Girl	1	2	3	4	5	6	7	8	9	10
Score relative to hypothesized median	−	0	+	+	+	+	+	+	+	+

approximately equal. This line of reasoning suggests an alternative way in which we could have stated the null hypothesis, namely, that the probability of a plus is equal to the probability of a minus, and these probabilities are equal to .5. Stated symbolically, the hypothesis would be

$$H_0 : P(+) = P(-) = .5$$

In other words, we would expect about the same number of plus signs as minus signs in Table 13.3.2 when H_0 is true. A look at Table 13.3.2 reveals a preponderance of pluses; specifically, we observe eight pluses, one minus, and one zero, which was assigned to the score that fell exactly on the median. The usual procedure for handling zeros is to eliminate them from the analysis and reduce n, the sample size, accordingly. If we follow this procedure, our problem reduces to one consisting of nine observations of which eight are plus and one is minus.

Since the number of pluses and minuses is not the same, we wonder if the distribution of signs is sufficiently disproportionate to cast doubt on our hypothesis. Stated another way, we wonder if this small a number of minuses could have come about by chance alone when the null hypothesis is true, or if the number is so small that something other than chance (that is, a false null hypothesis) is responsible for the results.

Based on what we learned in Chapter 4, it seems reasonable to conclude that the observations in Table 13.3.2 constitute a set of n independent random variables from the Bernoulli population with parameter p. If we let $k =$ the test statistic, the sampling distribution of k is the binomial probability distribution with parameter $p = .5$ if the null hypothesis is true.

6. **Decision rule.** The decision rule depends on the alternative hypothesis.
   For $H_A : P(+) > P(-)$, reject H_0 if, when H_0 is true, the probability of observing k or fewer minus signs is less than or equal to α.
   For $H_A : P(+) < P(-)$, reject H_0 if the probability of observing, when H_0 is true, k or fewer plus signs is equal to or less than α.
   For $H_A : P(+) \neq P(-)$, reject H_0 if (given that H_0 is true) the probability of obtaining a value of k as extreme as or more extreme than was actually computed is equal to or less than $\alpha/2$.

   For this example the decision rule is: Reject H_0 if the p value for the computed test statistic is less than or equal to .05.

7. **Calculation of test statistic.** We may determine the probability of observing x or fewer minus signs when given a sample of size n and parameter p by evaluating the following expression:

$$P(k \leq x \mid n, p) = \sum_{k=0}^{x} {}_nC_k p^k q^{n-k} \tag{13.3.1}$$

For our example we would compute

$${}_9C_0(.5)^0(.5)^{9-0} + {}_9C_1(.5)^1(.5)^{9-1} = .00195 + .01758 = .0195$$

8. **Statistical decision.** In Appendix Table B we find

$$P(k \leq 1|9, .5) = .0195$$

With a two-sided test either a sufficiently small number of minuses or a sufficiently small number of pluses would cause rejection of the null hypothesis. Since, in our example, there are fewer minuses, we focus our attention on minuses rather than pluses. By setting α equal to .05, we are saying that if the number of minuses is so small that the probability of observing this few or fewer is less than .025 (half of α), we will reject the null hypothesis. The probability we have computed, .0195, is less than .025. We, therefore, reject the null hypothesis.

9. **Conclusion.** We conclude that the median score is not 5.
10. ***p* value.** The p value for this test is $2(.0195) = .0390$. ■

Sign Test: Paired Data When the data to be analyzed consist of observations in matched pairs and the assumptions underlying the t test are not met, or the measurement scale is weak, the sign test may be employed to test the null hypothesis that the median difference is 0. An alternative way of stating the null hypothesis is

$$P(X_i > Y_i) = P(X_i < Y_i) = .5$$

One of the matched scores, say, Y_i, is subtracted from the other score, X_i. If Y_i is less than X_i, the sign of the difference is $+$, and if Y_i is greater than X_i, the sign of the difference is $-$. If the median difference is 0, we would expect a pair picked at random to be just as likely to yield a $+$ as a $-$ when the subtraction is performed. We may state the null hypothesis, then, as

$$H_0 : P(+) = P(-) = .5$$

In a random sample of matched pairs, we would expect the number of $+$'s and $-$'s to be about equal. If there are more $+$'s or more $-$'s than can be accounted for by chance alone when the null hypothesis is true, we will entertain some doubt about the truth of our null hypothesis. By means of the sign test, we can decide how many of one sign constitutes more than can be accounted for by chance alone.

EXAMPLE 13.3.2

A dental research team wished to know if teaching people how to brush their teeth would be beneficial. Twelve pairs of patients seen in a dental clinic were obtained by carefully matching on such factors as age, sex, intelligence, and initial oral hygiene scores. One member of each pair received instruction on how to brush his or her teeth and on other oral hygiene matters. Six months later all 24 subjects were examined and assigned an oral hygiene score by a dental hygienist unaware of which subjects had received the instruction. A low score indicates a high level of oral hygiene. The results are shown in Table 13.3.3.

Solution:

1. **Data.** See problem statement.
2. **Assumptions.** We assume that the population of differences between pairs of scores is a continuous variable.
3. **Hypotheses.** If the instruction produces a beneficial effect, this fact would be reflected in the scores assigned to the members of each pair. If we take the differences $X_i - Y_i$, we would expect to observe more $-$'s than $+$'s if instruction had been beneficial, since a low score indicates a higher level of oral hygiene. If, in fact, instruction is beneficial, the median of the hypothetical population of all such differences would be less than 0, that is, negative. If, on the other hand, instruction has no effect, the median of this population would be zero. The null and alternate hypotheses, then, are:

TABLE 13.3.3 Oral Hygiene Scores of 12 Subjects Receiving Oral Hygiene Instruction (X_i) and 12 Subjects Not Receiving Instruction (Y_i)

	Score	
Pair Number	Instructed (X_i)	Not Instructed (Y_i)
1	1.5	2.0
2	2.0	2.0
3	3.5	4.0
4	3.0	2.5
5	3.5	4.0
6	2.5	3.0
7	2.0	3.5
8	1.5	3.0
9	1.5	2.5
10	2.0	2.5
11	3.0	2.5
12	2.0	2.5

TABLE 13.3.4 Signs of Differences $(X_i - Y_i)$ in Oral Hygiene Scores of 12 Subjects Instructed (X_i) and 12 Matched Subjects Not Instructed (Y_i)

Pair	1	2	3	4	5	6	7	8	9	10	11	12
Sign of score differences	−	0	−	+	−	−	−	−	−	−	+	−

H_0: The median of the differences is zero $[P(+) = P(-)]$.
H_A: The median of the differences is negative $[P(+) < P(-)]$.

Let α be .05.

4. **Test statistic.** The test statistic is the number of plus signs.
5. **Distribution of test statistic.** The sampling distribution of k is the binomial distribution with parameters n and .5 if H_0 is true.
6. **Decision rule.** Reject H_0 if $P(k \leq 2 \,|11, .5) \leq .05$.
7. **Calculation of test statistic.** As will be seen, the procedure here is identical to the single sample procedure once the score differences have been obtained for each pair. Performing the subtractions and observing signs yields the results shown in Table 13.3.4.

   The nature of the hypothesis indicates a one-sided test so that all of $\alpha = .05$ is associated with the rejection region, which consists of all values of k (where k is equal to the number of + signs) for which the probability of obtaining that many or fewer pluses due to chance alone when H_0 is true is equal to or less than .05. We see in Table 13.3.4 that the experiment yielded one zero, two pluses, and nine minuses. When we eliminate the zero, the effective sample size is $n = 11$ with two pluses and nine minuses. In other words, since a "small" number of plus signs will cause rejection of the null hypothesis, the value of our test statistic is $k = 2$.
8. **Statistical decision.** We want to know the probability of obtaining no more than two pluses out of 11 tries when the null hypothesis is true. As we have seen, the answer is obtained by evaluating the appropriate binomial expression. In this example we find

$$P(k \leq 2 \,|11, .5) = \sum_{k=0}^{2} {}_{11}C_k (.5)^k (.5)^{11-k}$$

   By consulting Appendix Table B, we find this probability to be .0327. Since .0327 is less than .05, we must reject H_0.
9. **Conclusion.** We conclude that the median difference is negative. That is, we conclude that the instruction was beneficial.
10. ***p* value.** For this test, $p = .0327$.

■

Sign Test with "Greater Than" Tables As has been demonstrated, the sign test may be used with a single sample or with two samples in which each member of

one sample is matched with a member of the other sample to form a sample of matched pairs. We have also seen that the alternative hypothesis may lead to either a one-sided or a two-sided test. In either case we concentrate on the less frequently occurring sign and calculate the probability of obtaining that few or fewer of that sign.

We use the least frequently occurring sign as our test statistic because the binomial probabilities in Appendix Table B are "less than or equal to" probabilities. By using the least frequently occurring sign, we can obtain the probability we need directly from Table B without having to do any subtracting. If the probabilities in Table B were "greater than or equal to" probabilities, which are often found in tables of the binomial distribution, we would use the more frequently occurring sign as our test statistic in order to take advantage of the convenience of obtaining the desired probability directly from the table without having to do any subtracting. In fact, we could, in our present examples, use the more frequently occurring sign as our test statistic, but because Table B contains "less than or equal to" probabilities we would have to perform a subtraction operation to obtain the desired probability. As an illustration, consider the last example. If we use as our test statistic the most frequently occurring sign, it is 9, the number of minuses. The desired probability, then, is the probability of nine or more minuses, when $n = 11$ and $p = .5$. That is, we want

$$P(k = 9 \mid 11,\ .5)$$

However, since Table B contains "less than or equal to" probabilities, we must obtain this probability by subtraction. That is,

$$\begin{aligned} P(k \geq 9 \mid 11,\ .5) &= 1 - P(k \leq 8 \mid 11,\ .5) \\ &= 1 - .9673 \\ &= .0327 \end{aligned}$$

which is the result obtained previously.

Sample Size We saw in Chapter 5 that when the sample size is large and when p is close to .5, the binomial distribution may be approximated by the normal distribution. The rule of thumb used was that the normal approximation is appropriate when both np and nq are greater than 5. When $p = .5$, as was hypothesized in our two examples, a sample of size 12 would satisfy the rule of thumb. Following this guideline, one could use the normal approximation when the sign test is used to test the null hypothesis that the median or median difference is 0 and n is equal to or greater than 12. Since the procedure involves approximating a continuous distribution by a discrete distribution, the continuity correction of .5 is generally used. The test statistic then is

$$z = \frac{(k \pm .5) - .5n}{.5\sqrt{n}} \tag{13.3.2}$$

which is compared with the value of z from the standard normal distribution corresponding to the chosen level of significance. In Equation 13.3.2, $k + .5$ is used when $k < n/2$ and $k - .5$ is used when $k \geq n/2$.

Data:

```
C1:  4  5  8  8  9  6  10  7  6  6
```

Dialog box:

Stat ➤ Nonparametrics ➤ 1-Sample Sign

Type *C1* in **Variables.** Choose **Test median** and type *5* in the text box. Click **OK.**

Session command:

```
MTB > STest  5  C1;
SUBC>   Alternative  0.
```

Output:

Sign Test for Median: C1

```
Sign  test  of  median  =  5.00  versus  N.E.  5.000
                N  BELOW  EQUAL  ABOVE  P-VALUE  MEDIAN
C1             10      1      1      8   0.0391   6.500
```

FIGURE 13.3.1 MINITAB procedure and output for Example 13.3.1.

Computer Analysis Many statistics software packages will perform the sign test. For example, if we use MINITAB to perform the test for Example 13.3.1 in which the data are stored in Column 1, the procedure and output would be as shown in Figure 13.3.1.

EXERCISES

13.3.1 A random sample of 15 student nurses was given a test to measure their level of authoritarianism with the following results:

Student Number	Authoritarianism Score	Student Number	Authoritarianism Score
1	75	9	82
2	90	10	104
3	85	11	88
4	110	12	124
5	115	13	110
6	95	14	76
7	132	15	98
8	74		

Test at the .05 level of significance, the null hypothesis that the median score for the sampled population is 100. Determine the p value.

13.3.2 Determining the effects of grapefruit juice on pharmacokinetics of oral digoxin (a drug often prescribed for heart ailments) was the goal of a study by Parker et al. (A-1). Seven healthy nonsmoking volunteers participated in the study. Subjects took digoxin with water for 2 weeks, no digoxin for 2 weeks, and digoxin with grapefruit juice for 2 weeks. The average peak plasma digoxin concentration (Cmax) when subjects took digoxin with water is given in the first column of the following table. The second column gives the Cmax concentration when subjects took digoxin with grapefruit juice. May we conclude on the basis of these data that the Cmax concentration is higher when digoxin is taken with grapefruit juice? Let $\alpha = .5$.

	Cmax	
Subject	**H_2O**	**GFJ**
1	2.34	3.03
2	2.46	3.46
3	1.87	1.97
4	3.09	3.81
5	5.59	3.07
6	4.05	2.62
7	6.21	3.44

Source: Data provided courtesy of Robert B. Parker, Pharm.D.

13.3.3 A sample of 15 patients suffering from asthma participated in an experiment to study the effect of a new treatment on pulmonary function. Among the various measurements recorded were those of forced expiratory volume (liters) in 1 second (FEV_1) before and after application of the treatment. The results were as follows:

Subject	**Before**	**After**	**Subject**	**Before**	**After**
1	1.69	1.69	9	2.58	2.44
2	2.77	2.22	10	1.84	4.17
3	1.00	3.07	11	1.89	2.42
4	1.66	3.35	12	1.91	2.94
5	3.00	3.00	13	1.75	3.04
6	.85	2.74	14	2.46	4.62
7	1.42	3.61	15	2.35	4.42
8	2.82	5.14			

On the basis of these data, can one conclude that the treatment is effective in increasing the FEV_1 level? Let $\alpha = .05$ and find the p value.

13.4 THE WILCOXON SIGNED-RANK TEST FOR LOCATION

Sometimes we wish to test a null hypothesis about a population mean, but for some reason neither z nor t is an appropriate test statistic. If we have a small sample $(n < 30)$ from a population that is known to be grossly nonnormally distributed, and the central limit theorem is not applicable, the z statistic is ruled out. The t statistic is not appropriate

because the sampled population does not sufficiently approximate a normal distribution. When confronted with such a situation we usually look for an appropriate nonparametric statistical procedure. As we have seen, the sign test may be used when our data consist of a single sample or when we have paired data. If, however, the data for analysis are measured on at least an interval scale, the sign test may be undesirable because it would not make full use of the information contained in the data. A more appropriate procedure might be the Wilcoxon (1) signed-rank test, which makes use of the magnitudes of the differences between measurements and a hypothesized location parameter rather than just the signs of the differences.

Assumptions The Wilcoxon test for location is based on the following assumptions about the data.

1. The sample is random.
2. The variable is continuous.
3. The population is symmetrically distributed about its mean μ.
4. The measurement scale is at least interval.

Hypotheses The following are the null hypotheses (along with their alternatives) that may be tested about some unknown population mean μ_0.

(a) $H_0 : \mu = \mu_0$
$H_A : \mu \neq \mu_0$

(b) $H_0 : \mu \geq \mu_0$
$H_A : \mu < \mu_0$

(c) $H_0 : \mu \leq \mu_0$
$H_A : \mu > \mu_0$

When we use the Wilcoxon procedure, we perform the following calculations.

1. Subtract the hypothesized mean μ_0 from each observation x_i, to obtain

$$d_i = x_i - \mu_0$$

If any x_i is equal to the mean, so that $d_i = 0$, eliminate that d_i from the calculations and reduce n accordingly.

2. Rank the usable d_i from the smallest to the largest without regard to the sign of d_i. That is, consider only the absolute value of the d_i, designated $|d_i|$, when ranking them. If two or more of the $|d_i|$ are equal, assign each tied value the mean of the rank positions the tied values occupy. If, for example, the three smallest $|d_i|$ are all equal, place them in rank positions 1, 2, and 3, but assign each a rank of $(1 + 2 + 3)/3 = 2$.
3. Assign each rank the sign of the d_i that yields that rank.
4. Find T_+, the sum of the ranks with positive signs, and T_-, the sum of the ranks with negative signs.

The Test Statistic The Wilcoxon test statistic is either T_+ or T_-, depending on the nature of the alternative hypothesis. If the null hypothesis is true, that is, if the true population mean is equal to the hypothesized mean, and if the assumptions are met, the probability of observing a positive difference $d_i = x_i - \mu_0$ of a given magnitude is equal to

the probability of observing a negative difference of the same magnitude. Then, in repeated sampling, when the null hypothesis is true and the assumptions are met, the expected value of T_+ is equal to the expected value of T_-. We do not expect T_+ and T_- computed from a given sample to be equal. However, when H_0 is true, we do not expect a large difference in their values. Consequently, a sufficiently small value of T_+ or a sufficiently small value of T_- will cause rejection of H_0.

When the alternative hypothesis is two-sided ($\mu \neq \mu_0$), either a sufficiently small value of T_+ or a sufficiently small value of T_- will cause us to reject $H_0 : \mu = \mu_0$. The test statistic, then, is T_+ or T_-, whichever is smaller. To simplify notation, we call the smaller of the two T.

When $H_0 : \mu \geq \mu_0$ is true, we expect our sample to yield a large value of T_+. Therefore, when the one-sided alternative hypothesis states that the true population mean is less than the hypothesized mean ($\mu < \mu_0$), a sufficiently small value of T_+ will cause rejection of H_0, and T_+ is the test statistic.

When $H_0 : \mu \leq \mu_0$ is true, we expect our sample to yield a large value of T_-. Therefore, for the one-sided alternative $H_A : \mu > \mu_0$, a sufficiently small value of T_- will cause rejection of H_0 and T_- is the test statistic.

Critical Values Critical values of the Wilcoxon test statistic are given in Appendix Table K. Exact probability levels (P) are given to four decimal places for all possible rank totals (T) that yield a different probability level at the fourth decimal place from .0001 up through .5000. The rank totals (T) are tabulated for all sample sizes from $n = 5$ through $n = 30$. The following are the decision rules for the three possible alternative hypotheses:

(a) $H_A : \mu \neq \mu_0$. Reject H_0 at the α level of significance if the calculated T is smaller than or equal to the tabulated T for n and preselected $\alpha/2$. Alternatively, we may enter Table K with n and our calculated value of T to see whether the tabulated P associated with the calculated T is less than or equal to our stated level of significance. If so, we may reject H_0.

(b) $H_A : \mu < \mu_0$. Reject H_0 at the α level of significance if T_+ is less than or equal to the tabulated T for n and preselected α.

(c) $H_A : \mu > \mu_0$. Reject H_0 at the α level of significance if T_- is less than or equal to the tabulated T for n and preselected α.

EXAMPLE 13.4.1

Cardiac output (liters/minute) was measured by thermodilution in a simple random sample of 15 postcardiac surgical patients in the left lateral position. The results were as follows:

4.91	4.10	6.74	7.27	7.42	7.50	6.56	4.64
5.98	3.14	3.23	5.80	6.17	5.39	5.77	

We wish to know if we can conclude on the basis of these data that the population mean is different from 5.05.

Solution:

1. **Data.** See statement of example.
2. **Assumptions.** We assume that the requirements for the application of the Wilcoxon signed-ranks test are met.
3. **Hypotheses.**

   $H_0 : \mu = 5.05$
   $H_A : \mu \neq 5.05$

   Let $\alpha = 0.05$.
4. **Test statistic.** The test statistic will be T_+ or T_-, whichever is smaller. We will call the test statistic T.
5. **Distribution of test statistic.** Critical values of the test statistic are given in Table K of the Appendix.
6. **Decision rule.** We will reject H_0 if the computed value of T is less than or equal to 25, the critical value for $n = 15$, and $\alpha/2 = .0240$, the closest value to .0250 in Table K.
7. **Calculation of test statistic.** The calculation of the test statistic is shown in Table 13.4.1.
8. **Statistical decision.** Since 34 is greater than 25, we are unable to reject H_0.
9. **Conclusion.** We conclude that the population mean may be 5.05.
10. ***p* value.** From Table K we see that $p = 2(.0757) = .1514$.

TABLE 13.4.1 Calculation of the Test Statistic for Example 13.4.1

Cardiac Output	$d_i = x_i - 5.05$	Rank of $\lvert d_i \rvert$	Signed Rank of $\lvert d_i \rvert$
4.91	−.14	1	−1
4.10	−.95	7	−7
6.74	+1.69	10	+10
7.27	+2.22	13	+13
7.42	+2.37	14	+14
7.50	+2.45	15	+15
6.56	+1.51	9	+9
4.64	−.41	3	−3
5.98	+.93	6	+6
3.14	−1.91	12	−12
3.23	−1.82	11	−11
5.80	+.75	5	+5
6.17	+1.12	8	+8
5.39	+.34	2	+2
5.77	+.72	4	+4
			$T_+ = 86,\ T_- = 34,\ T = 34$

■

Dialog box:

Stat ➤ Nonparametrics ➤ 1-Sample Wilcoxon

Type *C1* in **Variables.** Choose **Test median.** Type 5.05 in the text box. Click **OK.**

Session command:

```
MTB > WTEST 5.05 C1;
SUBC>   Alternative 0.
```

Output:

Wilcoxon Signed Rank Test: C1

```
TEST OF MEDIAN = 5.050 VERSUS MEDIAN N.E. 5.050

                   N FOR WILCOXON          ESTIMATED
                N   TEST STATISTIC P-VALUE    MEDIAN
C1             15     15      86.0   0.148     5.747
```

FIGURE 13.4.1 MINITAB procedure and output for Example 13.4.1.

Wilcoxon Matched-Pairs Signed-Ranks Test The Wilcoxon test may be used with paired data under circumstances in which it is not appropriate to use the paired-comparisons t test described in Chapter 7. In such cases obtain each of the n d_i values, the difference between each of the n pairs of measurements. If we let $\mu_D =$ the mean of a population of such differences, we may follow the procedure described above to test any one of the following null hypotheses: $H_0 : \mu_D = 0$, $H_0 : \mu_D \geq 0$, and $H_0 : \mu_D \leq 0$.

Computer Analysis Many statistics software packages will perform the Wilcoxon signed-rank test. If, for example, the data of Example 13.4.1 are stored in Column 1, we could use MINITAB to perform the test as shown in Figure 13.4.1.

EXERCISES

13.4.1 Sixteen laboratory animals were fed a special diet from birth through age 12 weeks. Their weight gains (in grams) were as follows:

63 68 79 65 64 63 65 64 76 74 66 66 67 73 69 76

Can we conclude from these data that the diet results in a mean weight gain of less than 70 grams? Let $\alpha = .05$, and find the p value.

13.4.2 Amateur and professional singers were the subjects of a study by Grape et al. (A-2). The researchers investigated the possible beneficial effects of singing on well-being during a single singing lesson. One of the variables of interest was the change in cortisol as a result of the signing lesson. Use the data in the following table to determine if, in general, cortisol (nmol/L) increases after a singing lesson. Let $\alpha = .05$. Find the p value.

Subject	1	2	3	4	5	6	7	8
Before	214	362	202	158	403	219	307	331
After	232	276	224	412	562	203	340	313

Source: Data provided courtesy of Christina Grape, M.P.H., Licensed Nurse.

13.4.3 In a study by Zuckerman and Heneghan (A-3), hemodynamic stresses were measured on subjects undergoing laparoscopic cholecystectomy. An outcome variable of interest was the ventricular end diastolic volume (LVEDV) measured in milliliters. A portion of the data appear in the following table. Baseline refers to a measurement taken 5 minutes after induction of anesthesia, and the term "5 minutes" refers to a measurement taken 5 minutes after baseline.

	LVEDV (ml)	
Subject	Baseline	5 Minutes
1	51.7	49.3
2	79.0	72.0
3	78.7	87.3
4	80.3	88.3
5	72.0	103.3
6	85.0	94.0
7	69.7	94.7
8	71.3	46.3
9	55.7	71.7
10	56.3	72.3

Source: Data provided courtesy of R. S. Zuckerman, MD.

May we conclude, on the basis of these data, that among subjects undergoing laparoscopic cholecystectomy, the average LVEDV levels change? Let $\alpha = .01$.

13.5 THE MEDIAN TEST

A nonparametric procedure that may be used to test the null hypothesis that two independent samples have been drawn from populations with equal medians is the median test. The test, attributed mainly to Mood (2) and Westenberg (3), is also discussed by Brown and Mood (4).

We illustrate the procedure by means of an example.

EXAMPLE 13.5.1

Do urban and rural male junior high school students differ with respect to their level of mental health?

Solution:

1. **Data.** Members of a random sample of 12 male students from a rural junior high school and an independent random sample of 16 male

TABLE 13.5.1 Level of Mental Health Scores of Junior High Boys

School			
Urban	Rural	Urban	Rural
35	29	25	50
26	50	27	37
27	43	45	34
21	22	46	31
27	42	33	
38	47	26	
23	42	46	
25	32	41	

students from an urban junior high school were given a test to measure their level of mental health. The results are shown in Table 13.5.1.

To determine if we can conclude that there is a difference, we perform a hypothesis test that makes use of the median test. Suppose we choose a .05 level of significance.

2. **Assumptions.** The assumptions underlying the test are (a) the samples are selected independently and at random from their respective populations; (b) the populations are of the same form, differing only in location; and (c) the variable of interest is continuous. The level of measurement must be, at least, ordinal. The two samples do not have to be of equal size.
3. **Hypotheses.**

$$H_0 : M_U = M_R$$
$$H_A : M_U \neq M_R$$

M_U is the median score of the sampled population of urban students, and M_R is the median score of the sampled population of rural students. Let $\alpha = .05$.

4. **Test statistic.** As will be shown in the discussion that follows, the test statistic is X^2 as computed, for example, by Equation 12.4.1 for a 2×2 contingency table.
5. **Distribution of test statistic.** When H_0 is true and the assumptions are met, X^2 is distributed approximately as χ^2 with 1 degree of freedom.
6. **Decision rule.** Reject H_0 if the computed value of X^2 is ≥ 3.841 (since $\alpha = .05$).
7. **Calculation of test statistic.** The first step in calculating the test statistic is to compute the common median of the two samples combined. This is done by arranging the observations in ascending order

TABLE 13.5.2 Level of Mental Health Scores of Junior High School Boys

	Urban	Rural	Total
Number of scores above median	6	8	14
Number of scores below median	10	4	14
Total	16	12	28

and, because the total number of observations is even, obtaining the mean of the two middle numbers. For our example the median is $(33 + 34)/2 = 33.5$.

We now determine for each group the number of observations falling above and below the common median. The resulting frequencies are arranged in a 2×2 table. For the present example we construct Table 13.5.2.

If the two samples are, in fact, from populations with the same median, we would expect about one-half the scores in each sample to be above the combined median and about one-half to be below. If the conditions relative to sample size and expected frequencies for a 2×2 contingency table as discussed in Chapter 12 are met, the chi-square test with 1 degree of freedom may be used to test the null hypothesis of equal population medians. For our examples we have, by Formula 12.4.1,

$$X^2 = \frac{28[(6)(4) - (8)(10)]^2}{(16)(12)(14)(14)} = 2.33$$

8. **Statistical decision.** Since $2.33 < 3.841$, the critical value of χ^2 with $\alpha = .05$ and 1 degree of freedom, we are unable to reject the null hypothesis on the basis of these data.
9. **Conclusion.** We conclude that the two samples may have been drawn from populations with equal medians.
10. ***p* value.** Since $2.33 < 2.706$, we have $p > .10$. ■

Handling Values Equal to the Median Sometimes one or more observed values will be exactly equal to the common median and, hence, will fall neither above nor below it. We note that if $n_1 + n_2$ is odd, at least one value will always be exactly equal to the median. This raises the question of what to do with observations of this kind. One solution is to drop them from the analysis if $n_1 + n_2$ is large and there are only a few values that fall at the combined median. Or we may dichotomize the scores into those that exceed the median and those that do not, in which case the observations that equal the median will be counted in the second category.

Median Test Extension The median test extends logically to the case where it is desired to test the null hypothesis that $k \geq 3$ samples are from populations with equal medians. For this test a $2 \times k$ contingency table may be constructed by using the

Dialog box: **Session command:**

Stat ➤ Nonparametrics ➤ Mood's Median Test `MTB > Mood C1 C2.`

Type *C1* in **Response** and *C2* in **Factor.** Click **OK.**

Output:

Mood Median Test: C1 versus C2

```
Mood median test of C1

Chisquare = 2.33 df = 1 p = 0.127

                                        Individual 95.0% CIs
  C2 N<= N> Median   Q3-Q1 --------+---------+---------+--------
   1  10  6   27.0    15.0   (-+--------------------)
   2   4  8   39.5    14.8             (-------------+----------)
                             --------+---------+---------+--------
                                  30.0      36.0      42.0
Overall median = 33.5

A 95.0% C.I. for median (1) - median(2): (-17.1,3.1)
```

FIGURE 13.5.1 MINITAB procedure and output for Example 13.5.1.

frequencies that fall above and below the median computed from combined samples. If conditions as to sample size and expected frequencies are met, X^2 may be computed and compared with the critical χ^2 with $k - 1$ degrees of freedom.

Computer Analysis The median test calculations may be carried out using MINITAB. To illustrate using the data of Example 13.5.1 we first store the measurements in MINITAB Column 1. In MINITAB Column 2 we store codes that identify the observations as to whether they are for an urban (1) or rural (2) subject. The MINITAB procedure and output are shown in Figure 13.5.1.

EXERCISES

13.5.1 Fifteen patient records from each of two hospitals were reviewed and assigned a score designed to measure level of care. The scores were as follows:

Hospital A:	99, 85, 73, 98, 83, 88, 99, 80, 74, 91, 80, 94, 94, 98, 80
Hospital B:	78, 74, 69, 79, 57, 78, 79, 68, 59, 91, 89, 55, 60, 55, 79

Would you conclude, at the .05 level of significance, that the two population medians are different? Determine the p value.

13.5.2 The following serum albumin values were obtained from 17 normal and 13 hospitalized subjects:

Serum Albumin (g/100 ml)		Serum Albumin (g/100 ml)	
Normal Subjects	**Hospitalized Subjects**	**Normal Subjects**	**Hospitalized Subjects**
2.4 3.0	1.5 3.1	3.4 4.0	3.8 1.5
3.5 3.2	2.0 1.3	4.5 3.5	3.5
3.1 3.5	3.4 1.5	5.0 3.6	
4.0 3.8	1.7 1.8	2.9	
4.2 3.9	2.0 2.0		

Would you conclude at the .05 level of significance that the medians of the two populations sampled are different? Determine the p value.

13.6 THE MANN–WHITNEY TEST

The median test discussed in the preceding section does not make full use of all the information present in the two samples when the variable of interest is measured on at least an ordinal scale. Reducing an observation's information content to merely that of whether or not it falls above or below the common median is a waste of information. If, for testing the desired hypothesis, there is available a procedure that makes use of more of the information inherent in the data, that procedure should be used if possible. Such a nonparametric procedure that can often be used instead of the median test is the Mann–Whitney test (5), sometimes called the Mann–Whitney–Wilcoxon test. Since this test is based on the ranks of the observations, it utilizes more information than does the median test.

Assumptions The assumptions underlying the Mann–Whitney test are as follows:

1. The two samples, of size n and m, respectively, available for analysis have been independently and randomly drawn from their respective populations.
2. The measurement scale is at least ordinal.
3. The variable of interest is continuous.
4. If the populations differ at all, they differ only with respect to their medians.

Hypotheses When these assumptions are met we may test the null hypothesis that the two populations have equal medians against either of the three possible alternatives: (1) the populations do not have equal medians (two-sided test), (2) the median of population 1 is larger than the median of population 2 (one-sided test), or (3) the median of population 1 is smaller than the median of population 2 (one-sided test). If the two populations are symmetric, so that within each population the mean and median are the same, the conclusions we reach regarding the two population medians will also apply to the two population means. The following example illustrates the use of the Mann–Whitney test.

EXAMPLE 13.6.1

A researcher designed an experiment to assess the effects of prolonged inhalation of cadmium oxide. Fifteen laboratory animals served as experimental subjects, while 10 similar animals served as controls. The variable of interest was hemoglobin level following the experiment. The results are shown in Table 13.6.1. We wish to know if we can conclude that prolonged inhalation of cadmium oxide reduces hemoglobin level.

Solution:

1. **Data.** See Table 13.6.1.
2. **Assumptions.** We assume that the assumptions of the Mann–Whitney test are met.
3. **Hypotheses.** The null and alternative hypotheses are as follows:

$$H_0 : M_X \geq M_Y$$
$$H_A : M_X < M_Y$$

where M_X is the median of a population of animals exposed to cadmium oxide and M_Y is the median of a population of animals not exposed to the substance. Suppose we let $\alpha = .05$.

4. **Test statistic.** To compute the test statistic we combine the two samples and rank all observations from smallest to largest while keeping track of the sample to which each observation belongs. Tied observations are assigned a rank equal to the mean of the rank positions for which they are tied. The results of this step are shown in Table 13.6.2.

TABLE 13.6.1 Hemoglobin Determinations (grams) for 25 Laboratory Animals

Exposed Animals (X)	Unexposed Animals (Y)
14.4	17.4
14.2	16.2
13.8	17.1
16.5	17.5
14.1	15.0
16.6	16.0
15.9	16.9
15.6	15.0
14.1	16.3
15.3	16.8
15.7	
16.7	
13.7	
15.3	
14.0	

TABLE 13.6.2 Original Data and Ranks, Example 13.6.1

X	Rank	*Y*	Rank
13.7	1		
13.8	2		
14.0	3		
14.1	4.5		
14.1	4.5		
14.2	6		
14.4	7		
		15.0	8.5
		15.0	8.5
15.3	10.5		
15.3	10.5		
15.6	12		
15.7	13		
15.9	14		
		16.0	15
		16.2	16
		16.3	17
16.5	18		
16.6	19		
16.7	20		
		16.8	21
		16.9	22
		17.1	23
		17.4	24
		17.5	25
Total	145		

The test statistic is

$$T = S - \frac{n(n+1)}{2} \tag{13.6.1}$$

where n is the number of sample X observations and S is the sum of the ranks assigned to the sample observations from the population of X values. The choice of which sample's values we label X is arbitrary.

5. **Distribution of test statistic.** Critical values from the distribution of the test statistic are given in Appendix Table L for various levels of α.
6. **Decision rule.** If the median of the X population is, in fact, smaller than the median of the Y population, as specified in the alternative hypothesis, we would expect (for equal sample sizes) the sum of the ranks assigned

to the observations from the X population to be smaller than the sum of the ranks assigned to the observations from the Y population. The test statistic is based on this rationale in such a way that a sufficiently small value of T will cause rejection of $H_0 : M_X \geq M_Y$. In general, for one-sided tests of the type illustrated here the decision rule is:

Reject $H_0 : M_X = M_Y$ if the computed T is less than w_α, where w_α is the critical value of T obtained by entering Appendix Table L with n, the number of X observations; m, the number of Y observations; and α, the chosen level of significance.

If we use the Mann–Whitney procedure to test

$$H_0 : M_X \leq M_Y$$

against

$$H_A : M_X > M_Y$$

sufficiently large values of T will cause rejection so that the decision rule is:

Reject $H_0 : M_X \leq M_Y$ if computed T is greater than $w_{1-\alpha}$, where $w_{1-\alpha} = nm - w_\alpha$.

For the two-sided test situation with

$$H_0 : M_X = M_Y$$
$$H_A : M_X \neq M_Y$$

computed values of T that are either sufficiently large or sufficiently small will cause rejection of H_0. The decision rule for this case, then, is:

Reject $H_0 : M_X = M_Y$ if the computed value of T is either less than $w_{\alpha/2}$ or greater than $w_{1-(\alpha/2)}$ where $w_{\alpha/2}$ is the critical value of T for n, m, and $\alpha/2$ given in Appendix Table L, and $w_{1-(\alpha/2)} = nm - w_{\alpha/2}$.

For this example the decision rule is:

Reject H_0 if the computed value of T is smaller than 45, the critical value of the test statistic for $n = 15$, $m = 10$, and $\alpha = .05$ found in Table L.

The rejection regions for each set of hypotheses are shown in Figure 13.6.1.

7. **Calculation of test statistic.** For our present example we have, as shown in Table 13.6.2, $S = 145$, so that

$$T = 145 - \frac{15(15+1)}{2} = 25$$

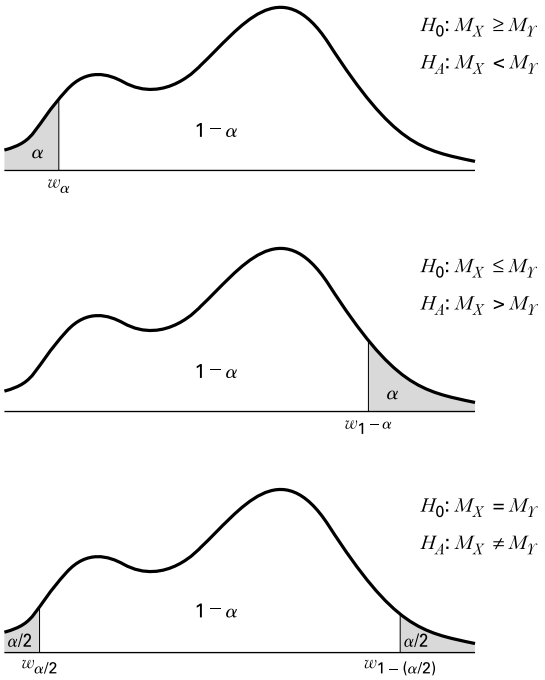


FIGURE 13.6.1 Mann–Whitney test rejection regions for three sets of hypotheses.

8. **Statistical decision.** When we enter Table L with $n = 15$, $m = 10$, and $\alpha = .05$, we find the critical value of w_α to be 45. Since $25 < 45$, we reject H_0.
9. **Conclusion.** We conclude that M_X is smaller than M_Y. This leads to the conclusion that prolonged inhalation of cadmium oxide does reduce the hemoglobin level.
10. ***p* value.** Since $22 < 25 < 30$, we have for this test $.005 > p > .001$.

■

Large-Sample Approximation When either n or m is greater than 20 we cannot use Appendix Table L to obtain critical values for the Mann–Whitney test. When this is the case we may compute

$$z = \frac{T - mn/2}{\sqrt{nm(n + m + 1|)/12}} \tag{13.6.2}$$

and compare the result, for significance, with critical values of the standard normal distribution.

Mann–Whitney Statistic and the Wilcoxon Statistic As was noted at the beginning of this section, the Mann–Whitney test is sometimes referred to as the

Dialog box:

Stat ➤ Nonparametrics ➤ Mann–Whitney

Type *C1* in **First Sample** and *C2* in **Second Sample.**
At **Alternative** choose less than.
Click **OK.**

Session command:

```
MTB > Mann-Whitney 95.0
C1 C2;
SUBC > Alternative −1.
```

Output:

Mann–Whitney Test and CI: C1, C2

```
C1      N = 15      Median =        15.300
C2      N = 10      Median =        16.550
Point estimate for ETA1 − ETA2 is −1.300
95.1 Percent C.I. for ETA1 − ETA2 is (−2.300,−0.600)
W = 145.0
Test of ETA1 = ETA2 vs. ETA1 < ETA2 is significant at 0.0030
The test is significant at 0.0030 (adjusted for ties)
```

FIGURE 13.6.2 MINITAB procedure and output for Example 13.6.1.

Ranks

	y	N	Mean Rank	Sum of Rank
x	1.000000	15	9.67	146.00
	2.000000	10	18.00	180.00
	Total	25		

Test Statistic[b]

	x
Mann-Whitney U	25.000
Wilcoxon W	145.000
Z	−2.775
Asymp. Sig. (2-tailed)	.006
Exact Sig. [2*(1-tailed Sig.)]	.004[a]

a. Not corrected for ties
b. Grouping Variable: y

FIGURE 13.6.3 SPSS output for Example 13.6.1.

Mann–Whitney-Wilcoxon test. Indeed, many computer packages give the test value of both the Mann–Whitney test (U) and the Wilcoxon test (W). These two tests are algebraically equivalent tests, and are related by the following equality when there are no ties in the data:

$$U + W = \frac{m(m + 2n + 1)}{2} \tag{13.6.3}$$

Computer Analysis Many statistics software packages will perform the Mann–Whitney test. With the data of two samples stored in Columns 1 and 2, for example, MINITAB will perform a one-sided or two-sided test. The MINITAB procedure and output for Example 13.6.1 are shown in Figure 13.6.2.

The SPSS output for Example 13.6.1 is shown in Figure 13.6.3. As we see this output provides the Mann–Whitney test, the Wilcoxon test, and large-sample z approximation.

EXERCISES

13.6.1 Cranor and Christensen (A-4) studied diabetics insured by two employers. Group 1 subjects were employed by the City of Asheville, North Carolina, and group 2 subjects were employed by Mission–St. Joseph's Health System. At the start of the study, the researchers performed the Mann–Whitney test to determine if a significant difference in weight existed between the two study groups. The data are displayed in the following table.

Weight (Pounds)					
Group 1			Group 2		
252	215	240	185	195	220
240	190	302	310	210	295
205	270	312	212	190	202
200	159	126	238	172	268
170	204	268	184	190	220
170	215	215	136	140	311
320	254	183	200	280	164
148	164	287	270	264	206
214	288	210	200	270	170
270	138	225	212	210	190
265	240	258	182	192	
203	217	221	225	126	

Source: Data provided courtesy of Carole W. Carnor, Ph.D.

May we conclude, on the basis of these data, that patients in the two groups differ significantly with respect to weight? Let $\alpha = .05$.

13.6.2 One of the purposes of a study by Liu et al. (A-5) was to determine the effects of MRZ 2/579 (a receptor antagonist shown to provide neuroprotective activity in vivo and in vitro) on neurological

deficit in Sprague–Dawley rats. In the study, 10 rats were to receive MRZ 2/579 and nine rats were to receive regular saline. Prior to treatment, researchers studied the blood gas levels in the two groups of rats. The following table shows the pO_2 levels for the two groups.

Saline (mmHg)	MRZ 2/579 (mmHg)
112.5	133.3
106.3	106.4
99.5	113.1
98.3	117.2
103.4	126.4
109.4	98.1
108.9	113.4
107.4	116.8
	116.5

Source: Data provided courtesy of Ludmila Belayev, M.D.

May we conclude, on the basis of these data, that, in general, subjects on saline have, on average, lower pO_2 levels at baseline? Let $\alpha = .01$.

13.6.3 The purpose of a study by researchers at the Cleveland (Ohio) Clinic (A-6) was to determine if the use of Flomax® reduced the urinary side effects commonly experienced by patients following brachytherapy (permanent radioactive seed implant) treatment for prostate cancer. The following table shows the American Urological Association (AUA) symptom index scores for two groups of subjects after 8 weeks of treatment. The higher the AUA index, the more severe the urinary obstruction and irritation.

AUA Index (Flomax®)			AUA Index (Placebo)		
1	5	11	1	6	12
1	5	11	1	6	12
2	6	11	2	6	13
2	6	11	2	6	14
2	7	12	2	6	17
2	7	12	3	7	18
3	7	13	3	8	19
3	7	14	3	8	20
3	8	16	3	9	23
4	8	16	4	9	23
4	8	18	4	10	
4	8	21	4	10	
4	9	31	5	11	
4	9		5	11	
4	10		5	12	

Source: Data provided courtesy of Chandana Reddy, M.S.

May we conclude, on the basis of these data, that the median AUA index in the Flomax® group differs significantly from the median AUA index of the placebo group? Let $\alpha = .05$.

13.7 THE KOLMOGOROV–SMIRNOV GOODNESS-OF-FIT TEST

When one wishes to know how well the distribution of sample data conforms to some theoretical distribution, a test known as the Kolmogorov–Smirnov goodness-of-fit test provides an alternative to the chi-square goodness-of-fit test discussed in Chapter 12. The test gets its name from A. Kolmogorov and N. V. Smirnov, two Russian mathematicians who introduced two closely related tests in the 1930s.

Kolmogorov's work (6) is concerned with the one-sample case as discussed here. Smirnov's work (7) deals with the case involving two samples in which interest centers on testing the hypothesis that the distributions of the two-parent populations are identical. The test for the first situation is frequently referred to as the Kolmogorov–Smirnov one-sample test. The test for the two-sample case, commonly referred to as the Kolmogorov–Smirnov two-sample test, will not be discussed here.

The Test Statistic In using the Kolmogorov–Smirnov goodness-of-fit test, a comparison is made between some theoretical cumulative distribution function, $F_T(x)$, and a sample cumulative distribution function, $F_S(x)$. The sample is a random sample from a population with unknown cumulative distribution function $F(x)$. It will be recalled (Section 4.2) that a cumulative distribution function gives the probability that X is equal to or less than a particular value, x. That is, by means of the sample cumulative distribution function, $F_S(x)$, we may estimate $P(X \leq x)$. If there is close agreement between the theoretical and sample cumulative distributions, the hypothesis that the sample was drawn from the population with the specified cumulative distribution function, $F_T(x)$, is supported. If, however, there is a discrepancy between the theoretical and observed cumulative distribution functions too great to be attributed to chance alone, when H_0 is true, the hypothesis is rejected.

The difference between the theoretical cumulative distribution function, $F_T(x)$, and the sample cumulative distribution function, $F_S(x)$, is measured by the statistic D, which is the greatest vertical distance between $F_S(x)$ and $F_T(x)$. When a two-sided test is appropriate, that is, when the hypotheses are

$$H_0 : F(x) = F_T(x) \quad \text{for all } x \quad \text{from } -\infty \text{ to } +\infty$$

$$H_A : F(x) \neq F_T(x) \quad \text{for at least one } x$$

the test statistic is

$$D = \sup_x |F_S(x) - F_T(x)| \tag{13.7.1}$$

which is read, "D equals the supremum (greatest), over all x, of the absolute value of the difference $F_S(X)$ minus $F_T(X)$."

The null hypothesis is rejected at the α level of significance if the computed value of D exceeds the value shown in Appendix Table M for $1 - \alpha$ (two-sided) and the sample size n.

Assumptions The assumptions underlying the Kolmogorov–Smirnov test include the following:

1. The sample is a random sample.
2. The hypothesized distribution $F_T(x)$ is continuous.

When values of D are based on a discrete theoretical distribution, the test is conservative. When the test is used with discrete data, then, the investigator should bear in mind that the true probability of committing a type I error is at most equal to α, the stated level of significance. The test is also conservative if one or more parameters have to be estimated from sample data.

EXAMPLE 13.7.1

Fasting blood glucose determinations made on 36 nonobese, apparently healthy, adult males are shown in Table 13.7.1. We wish to know if we may conclude that these data are not from a normally distributed population with a mean of 80 and a standard deviation of 6.

Solution:

1. **Data.** See Table 13.7.1.
2. **Assumptions.** The sample available is a simple random sample from a continuous population distribution.
3. **Hypotheses.** The appropriate hypotheses are

$$H_0 : F(x) = F_T(x) \quad \text{for all } x \text{ from } -\infty \text{ to } +\infty$$
$$H_A : F(x) \neq F_T(x) \quad \text{for at least one } x$$

   Let $a = .05$.
4. **Test statistic.** See Equation 13.7.1.
5. **Distribution of test statistic.** Critical values of the test statistic for selected values of α are given in Appendix Table M.
6. **Decision rule.** Reject H_0 if the computed value of D exceeds .221, the critical value of D for $n = 36$ and $\alpha = .05$.
7. **Calculation of test statistic.** Our first step is to compute values of $F_S(x)$ as shown in Table 13.7.2.

TABLE 13.7.1 Fasting Blood Glucose Values (mg/100 ml) for 36 Nonobese, Apparently Healthy, Adult Males

75	92	80	80	84	72
84	77	81	77	75	81
80	92	72	77	78	76
77	86	77	92	80	78
68	78	92	68	80	81
87	76	80	87	77	86

TABLE 13.7.2 Values of $F_S(x)$ for Example 13.7.1

x	Frequency	Cumulative Frequency	$F_S(x)$
68	2	2	.0556
72	2	4	.1111
75	2	6	.1667
76	2	8	.2222
77	6	14	.3889
78	3	17	.4722
80	6	23	.6389
81	3	26	.7222
84	2	28	.7778
86	2	30	.8333
87	2	32	.8889
92	4	36	1.0000
	36		

Each value of $F_S(x)$ is obtained by dividing the corresponding cumulative frequency by the sample size. For example, the first value of $F_S(x) = 2/36 = .0556$.

We obtain values of $F_T(x)$ by first converting each observed value of x to a value of the standard normal variable, z. From Appendix Table D we then find the area between $-\infty$ and z. From these areas we are able to compute values of $F_T(x)$. The procedure, which is similar to that used to obtain expected relative frequencies in the chi-square goodness-of-fit test, is summarized in Table 13.7.3.

TABLE 13.7.3 Steps in Calculation of $F_T(x)$ for Example 13.7.1

x	$z = (x - 80)/6$	$F_T(x)$
68	−2.00	.0228
72	−1.33	.0918
75	−.83	.2033
76	−.67	.2514
77	−.50	.3085
78	−.33	.3707
80	.00	.5000
81	.17	.5675
84	.67	.7486
86	1.00	.8413
87	1.17	.8790
92	2.00	.9772

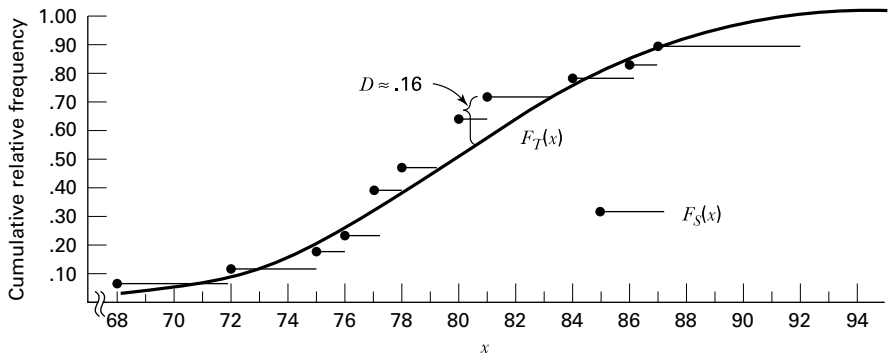


FIGURE 13.7.1 $F_S(x)$ and $F_T(x)$ for Example 13.7.1.

The test statistic D may be computed algebraically, or it may be determined graphically by actually measuring the largest vertical distance between the curves of $F_S(x)$ and $F_T(x)$ on a graph. The graphs of the two distributions are shown in Figure 13.7.1.

Examination of the graphs of $F_S(x)$ and $F_T(x)$ reveals that $D \approx .16 = (.72 - .56)$. Now let us compute the value of D algebraically. The possible values of $|F_S(x) - F_T(x)|$ are shown in Table 13.7.4. This table shows that the exact value of D is .1547.

8. **Statistical decision.** Reference to Table M reveals that a computed D of .1547 is not significant at any reasonable level. Therefore, we are not willing to reject H_0.
9. **Conclusion.** The sample may have come from the specified distribution.
10. ***p* value.** Since we have a two-sided test, and since $.1547 < .174$, we have $p > .20$.

TABLE 13.7.4 Calculation of $|F_s(x) - F_T(x)|$ for Example 13.7.1

x	$F_s(x)$	$F_T(x)$	$\lvert F_s(x) - F_T(x)\rvert$
68	.0556	.0228	.0328
72	.1111	.0918	.0193
75	.1667	.2033	.0366
76	.2222	.2514	.0292
77	.3889	.3085	.0804
78	.4722	.3707	.1015
80	.6389	.5000	.1389
81	.7222	.5675	.1547
84	.7778	.7486	.0292
86	.8333	.8413	.0080
87	.8889	.8790	.0099
92	1.0000	.9772	.0228

■

Kolmogorov–Smirnov One-Sample Test

kolmogorov (response = glucose, method = asymp, di = no (mean = 80, stddev = 6), time_limit = none);

Data File:

Column Variable: Glucose

Sample Size: 36

Summary of the Test Statistic:

	Type	Mean	Std. Dev
Hypothesized distribution F(X)	Normal	80	6

Let S(X) be the empirical distribution.

Inference:

	Statistic		
Item	**Sup{|S(X) - F(X)|}**	**Sup{S(X) - F(X)}**	**Sup{F(X) - S(X)}**
Observed Statistic	0.156	0.156	0.09122
Stand. Statistic	0.9362	0.9362	0.5473
Asymptotic p-value	0.3447	0.1732	0.5493

FIGURE 13.7.2 StatXact output for Example 13.7.1

StatXact is often used for nonparametric statistical analysis. This particular software program has a nonparametric module that contains nearly all of the commonly used nonparametric tests, and many less common, but useful, procedures as well. Computer analysis using StatXact for the data in Example 13.7.1 is shown in Figure 13.7.2. Note that it provides the test statistic of $D = 0.156$ and the exact two-sided p value of .3447.

A Precaution The reader should be aware that in determining the value of D, *it is not always sufficient to compute and choose from the possible values of* $|F_S(x) - F_T(x)|$. *The largest vertical distance between* $F_S(x)$ *and* $F_T(x)$ *may not occur at an observed value,* x, *but at some other value of* X. Such a situation is illustrated in Figure 13.7.3. We see that if only values of $|F_S(x) - F_T(x)|$ at the left endpoints of the horizontal bars are considered, we would incorrectly compute D as $|.2 - .4| = .2$. One can see by examining the graph, however, that the largest vertical distance between $F_S(x)$ and $F_T(x)$ occurs at the right endpoint of the horizontal bar originating at the point corresponding to $x = .4$, and the correct value of D is $|.5 - .2| = .3$.

One can determine the correct value of D algebraically by computing, in addition to the differences $|F_S(x) - F_T(x)|$, the differences $|F_S(x_{i-1}) - F_T(x_i)|$ for all values of $i = 1, 2, \ldots, r + 1$, where $r =$ the number of different values of x and $F_S(x_0) = 0$. The correct value of the test statistic will then be

$$D = \underset{1 \le i \le r}{\text{maximum}} \{\text{maximum}[|F_S(x_i) - F_T(x_i)|, |F_S(x_{i-1}) - F_T(x_i)|]\} \tag{13.7.2}$$

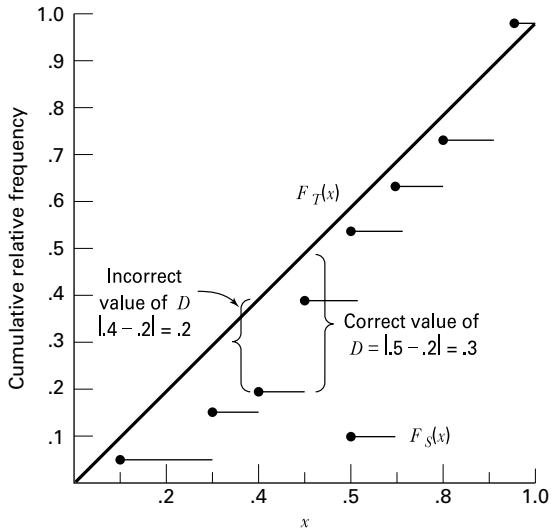


FIGURE 13.7.3 Graph of fictitious data showing correct calculation of D.

Advantages and Disadvantages The following are some important points of comparison between the Kolmogorov–Smirnov and the chi-square goodness-of-fit tests.

1. The Kolmogorov–Smirnov test does not require that the observations be grouped as is the case with the chi-square test. The consequence of this difference is that the Kolmogorov–Smirnov test makes use of all the information present in a set of data.
2. The Kolmogorov–Smirnov test can be used with any size sample. It will be recalled that certain minimum sample sizes are required for the use of the chi-square test.
3. As has been noted, the Kolmogorov–Smirnov test is not applicable when parameters have to be estimated from the sample. The chi-square test may be used in these situations by reducing the degrees of freedom by 1 for each parameter estimated.
4. The problem of the assumption of a continuous theoretical distribution has already been mentioned.

EXERCISES

13.7.1 The weights at autopsy of the brains of 25 adults suffering from a certain disease were as follows:

Weight of Brain (grams)				
859	1073	1041	1166	1117
962	1051	1064	1141	1202
973	1001	1016	1168	1255
904	1012	1002	1146	1233
920	1039	1086	1140	1348

Can one conclude from these data that the sampled population is not normally distributed with a mean of 1050 and a standard deviation of 50? Determine the p value for this test.

13.7.2 IQs of a sample of 30 adolescents arrested for drug abuse in a certain metropolitan jurisdiction were as follows:

IQ					
95	100	91	106	109	110
98	104	97	100	107	119
92	106	103	106	105	112
101	91	105	102	101	110
101	95	102	104	107	118

Do these data provide sufficient evidence that the sampled population of IQ scores is not normally distributed with a mean of 105 and a standard deviation of 10? Determine the p value.

13.7.3 For a sample of apparently normal subjects who served as controls in an experiment, the following systolic blood pressure readings were recorded at the beginning of the experiment:

162	177	151	167
130	154	179	146
147	157	141	157
153	157	134	143
141	137	151	161

Can one conclude on the basis of these data that the population of blood pressures from which the sample was drawn is not normally distributed with $\mu = 150$ and $\sigma = 12$? Determine the p value.

13.8 THE KRUSKAL–WALLIS ONE-WAY ANALYSIS OF VARIANCE BY RANKS

In Chapter 8 we discuss how one-way analysis of variance may be used to test the null hypothesis that several population means are equal. When the assumptions underlying this technique are not met, that is, when the populations from which the samples are drawn are not normally distributed with equal variances, or when the data for analysis consist only of ranks, a nonparametric alternative to the one-way analysis of variance may be used to test the hypothesis of equal location parameters. As was pointed out in Section 13.5, the median test may be extended to accommodate the situation involving more than two samples. A deficiency of this test, however, is the fact that it uses only a small amount of the information available. The test uses only information as to whether or not the observations are above or below a single number, the median of the combined samples. The test does not directly use measurements of known quantity. Several nonparametric analogs to analysis of variance are available that use more information by taking into account the magnitude of

each observation relative to the magnitude of every other observation. Perhaps the best known of these procedures is the Kruskal–Wallis one-way analysis of variance by ranks (8).

The Kruskal–Wallis Procedure The application of the test involves the following steps.

1. The $n_1, n_2, \ldots, n_k$ observations from the k samples are combined into a single series of size n and arranged in order of magnitude from smallest to largest. The observations are then replaced by ranks from 1, which is assigned to the smallest observation, to n, which is assigned to the largest observation. When two or more observations have the same value, each observation is given the mean of the ranks for which it is tied.
2. The ranks assigned to observations in each of the k groups are added separately to give k rank sums.
3. The test statistic

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1) \tag{13.8.1}$$

is computed. In Equation 13.8.1,

k = the number of samples
n_j = the number of observations in the jth *sample*
n = the number of observations in all samples combined
R_j = the sum of the ranks in the jth *sample*

4. When there are three samples and five or fewer observations in each sample, the significance of the computed H is determined by consulting Appendix Table N. When there are more than five observations in one or more of the samples, H is compared with tabulated values of χ^2 with $k - 1$ degrees of freedom.

EXAMPLE 13.8.1

In a study of pulmonary effects on guinea pigs, Lacroix et al. (A-7) exposed ovalbumin (OA)-sensitized guinea pigs to regular air, benzaldehyde, or acetaldehyde. At the end of exposure, the guinea pigs were anesthetized and allergic responses were assessed in bronchoalveolar lavage (BAL). One of the outcome variables examined was the count of eosinophil cells, a type of white blood cell that can increase with allergies. Table 13.8.1 gives the eosinophil cell count $\left(\times 10^6\right)$ for the three treatment groups.

Can we conclude that the three populations represented by the three samples differ with respect to eosinophil cell count? We can so conclude if we can reject the null hypothesis that the three populations do not differ in eosinophil cell count.

TABLE 13.8.1 Eosinophil Count for Ovalbumin-Sensitized Guinea Pigs

Eosinophil Cell Count ($\times 10^6$)		
Air	**Benzaldehyde**	**Acetaldehyde**
12.22	3.68	54.36
28.44	4.05	27.87
28.13	6.47	66.81
38.69	21.12	46.27
54.91	3.33	30.19

Source: Data provided courtesy of G. Lacroix.

Solution:

1. **Data.** See Table 13.8.1.
2. **Assumptions.** The samples are independent random samples from their respective populations. The measurement scale employed is at least ordinal. The distributions of the values in the sampled populations are identical except for the possibility that one or more of the populations are composed of values that tend to be larger than those of the other populations.
3. **Hypotheses.**

   H_0: The population centers are all equal.

   H_A: At least one of the populations tends to exhibit larger values than at least one of the other populations.

   Let $\alpha = .01$.
4. **Test statistic.** See Equation 13.8.1.
5. **Distribution of test statistic.** Critical values of H for various sample sizes and α levels are given in Appendix Table N.
6. **Decision rule.** The null hypothesis will be rejected if the computed value of H is so large that the probability of obtaining a value that large or larger when H_0 is true is equal to or less than the chosen significance level, α.
7. **Calculation of test statistic.** When the three samples are combined into a single series and ranked, the table of ranks shown in Table 13.8.2 may be constructed.

   The null hypothesis implies that the observations in the three samples constitute a single sample of size 15 from a single population. If this is true, we would expect the ranks to be well distributed among the three groups. Consequently, we would expect the total sum of ranks to be divided among the three groups in proportion to group size.

TABLE 13.8.2 The Data of Table 13.8.1 Replaced by Ranks

Air	Benzaldehyde	Acetaldehyde
5	2	13
9	3	7
8	4	15
11	6	12
14	1	10
$R_1 = 47$	$R_2 = 16$	$R_3 = 57$

Departures from these conditions are reflected in the magnitude of the test statistics H.

From the data in Table 13.8.2 and Equation 13.8.1, we obtain

$$H = \frac{12}{15(16)}\left[\frac{(47)^2}{5} + \frac{(16)^2}{5} + \frac{(57)^2}{5}\right] - 3(15+1) = 9.14$$

8. **Statistical decision.** Table N shows that when the n_j are 5, 5, and 5, the probability of obtaining a value of $H = 9.14$ is less than .009. The null hypothesis can be rejected at the .01 level of significance.
9. **Conclusion.** We conclude that there is a difference in the average eosinophil cell count among the three populations.
10. ***p* value.** For this test, $p < .009$.

■

Ties When ties occur among the observations, we may adjust the value of H by dividing it by

$$1 - \frac{\sum T}{n^3 - n} \tag{13.8.2}$$

where $T = t^3 - t$. The letter t is used to designate the number of tied observations in a group of tied values. In our example there are no groups of tied values but, in general, there may be several groups of tied values resulting in several values of T.

The effect of the adjustment for ties is usually negligible. Note also that the effect of the adjustment is to increase H, so that if the unadjusted H is significant at the chosen level, there is no need to apply the adjustment.

More than Three Samples/Large Samples Now let us illustrate the procedure when there are more than three samples and at least one of the n_j is greater than 5.

TABLE 13.8.3 Net Book Value of Equipment per Bed by Hospital Type

Type Hospital				
A	B	C	D	E
$1735(11)	$5260(35)	$2790(20)	$3475(26)	$6090(40)
1520(2)	4455(28)	2400(12)	3115(22)	6000(38)
1476(1)	4480(29)	2655(16)	3050(21)	5894(37)
1688(7)	4325(27)	2500(13)	3125(23)	5705(36)
1702(10)	5075(32)	2755(19)	3275(24)	6050(39)
2667(17)	5225(34)	2592(14)	3300(25)	6150(41)
1575(4)	4613(30)	2601(15)	2730(18)	5110(33)
1602(5)	4887(31)	1648(6)		
1530(3)		1700(9)		
1698(8)				
$R_1 = 68$	$R_2 = 246$	$R_3 = 124$	$R_4 = 159$	$R_5 = 264$

EXAMPLE 13.8.2

Table 13.8.3 shows the net book value of equipment capital per bed for a sample of hospitals from each of five types of hospitals. We wish to determine, by means of the Kruskal–Wallis test, if we can conclude that the average net book value of equipment capital per bed differs among the five types of hospitals. The ranks of the 41 values, along with the sum of ranks for each sample, are shown in the table.

Solution: From the sums of the ranks we compute

$$H = \frac{12}{41(41+1)}\left[\frac{(68)^2}{10} + \frac{(246)^2}{8} + \frac{(124)^2}{9} + \frac{(159)^2}{7} + \frac{(264)^2}{7}\right] - 3(41+1)$$
$$= 36.39$$

Reference to Appendix Table F with $k - 1 = 4$ degrees of freedom indicates that the probability of obtaining a value of H as large as or larger than 36.39, due to chance alone, when there is no difference among the populations, is less than .005. We conclude, then, that there is a difference among the five populations with respect to the average value of the variable of interest. ■

Computer Analysis The MINITAB software package computes the Kruskal–Wallis test statistic and provides additional information. After we enter the eosinophil counts in Table 13.8.1 into Column 1 and the group codes into Column 2, the MINITAB procedure and output are as shown in Figure 13.8.1.

Data:

```
C1: 12.22 28.44 28.13 38.69 54.91 3.68 4.05 6.47 21.12 3.33 54.36 27.87 66.81 46.27 30.19
C2: 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
```

Dialog box:

Stat ➤ Nonparametrics ➤ Kruskal–Wallis
Type *C1* in **Response** and *C2* in **Factor.** Click **OK.**

Session command:

```
MTB > Kruskal–Wallis  C1  C2.
```

Output:

Kruskal–Wallis Test: C1 versus C2

```
Kruskal–Wallis  Test  on  C1

C2          N     Median    Ave  Rank        Z
1           5     28.440          9.4     0.86
2           5      4.050          3.2    -2.94
3           5     46.270         11.4     2.08
Overall    15                     8.0

H  =  9.14   DF  =  2   P  =  0.010
```

FIGURE 13.8.1 MINITAB procedure and output, Kruskal–Wallis test of eosinophil count data in Table 13.8.1.

EXERCISES

For the following exercises, perform the test at the indicated level of significance and determine the *p* value.

13.8.1 In a study of healthy subjects grouped by age (Younger: 19–50 years, Seniors: 65–75 years, and Longeval: 85–102 years), Herrmann et al. (A-8) measured their vitamin B-12 levels (ng/L). All elderly subjects were living at home and able to carry out normal day-to-day activities. The following table shows vitamin B-12 levels for 50 subjects in the young group, 92 seniors, and 90 subjects in the longeval group.

Young (19–50 Years)		Senior (65–75 Years)				Longeval (85–102 Years)			
230	241	319	371	566	170	148	149	631	198
477	442	190	460	290	542	1941	409	305	321
561	491	461	440	271	282	128	229	393	2772
347	279	163	520	308	194	145	183	282	428
566	334	377	256	440	445	174	193	273	259
260	247	190	335	238	921	495	161	157	111

(*Continued*)

Young (19–50 Years)		Senior (65–75 Years)				Longeval (85–102 Years)			
300	314	375	137	525	1192	460	400	1270	262
230	254	229	452	298	748	548	348	252	161
215	419	193	437	153	187	198	175	262	1113
260	335	294	236	323	350	165	540	381	409
349	455	740	432	205	1365	226	293	162	378
315	297	194	411	248	232	557	196	340	203
257	456	780	268	371	509	166	632	370	221
536	668	245	703	668	357	218	438	483	917
582	240	258	282	197	201	186	368	222	244
293	320	419	290	260	177	346	262	277	
569	562	372	286	198	872	239	190	226	
325	360	413	143	336		240	241	203	
275	357	685	310	421		136	195	369	
172	609	136	352	712		359	220	162	
2000	740	441	262	461		715	164	95	
240	430	423	404	631		252	279	178	
235	645	617	380	1247		414	297	530	
284	395	985	322	1033		372	474	334	
883	302	170	340	285		236	375	521	

Source: Data provided courtesy of W. Herrmann and H. Schorr.

May we conclude, on the basis of these data, that the populations represented by these samples differ with respect to vitamin B-12 levels? Let $\alpha = .01$.

13.8.2 The following are outpatient charges $(-\$100)$ made to patients for a certain surgical procedure by samples of hospitals located in three different areas of the country:

Area		
I	**II**	**III**
\$80.75	\$58.63	\$84.21
78.15	72.70	101.76
85.40	64.20	107.74
71.94	62.50	115.30
82.05	63.24	126.15

Can we conclude at the .05 level of significance that the three areas differ with respect to the charges?

13.8.3 A study of young children by Flexer et al. (A-9) published in the *Hearing Journal* examines the effectiveness of an FM sound field when teaching phonics to children. In the study, children in a classroom with no phonological or phonemic awareness training (control) were compared to a class with phonological and phonemic awareness (PPA) and to a class that utilized phonological and phonemic awareness training and the FM sound field (PPA/FM). A total of 53 students from three separate preschool classrooms participated in this study. Students were given a measure of phonemic awareness in preschool and then at the end of the first semester of kindergarten. The improvement scores are listed in the following table as measured by the Yopp–Singer Test of Phonemic Segmentation.

Improvement (Control)		Improvement PPA	Improvement PPA/FM	
0	1	2	1	19
−1	1	3	3	20
0	2	15	7	21
1	2	18	9	21
4	3	19	11	22
5	6	20	17	22
9	7	5	17	15
9	8		17	17
13	9		18	17
18	18		18	19
0	20		19	22
0			19	

Source: Data provided courtesy of John P. Holcomb, Jr., Ph.D.

Test for a significant difference among the three groups. Let $\alpha = .05$.

13.8.4 Refer to Example 13.8.1. Another variable of interest to Lacroix et al. (A-7) was the number of alveolar cells in three groups of subjects exposed to air, benzaldehyde, or acetaldehyde. The following table gives the information for six guinea pigs in each of the three treatment groups.

Number of Alveolar Cells ($\times 10^6$)		
Air	**Benzaldehyde**	**Acetaldehyde**
0.55	0.81	0.65
0.48	0.56	13.69
7.8	1.11	17.11
8.72	0.74	7.43
0.65	0.77	5.48
1.51	0.83	0.99
0.55	0.81	0.65

Source: Data provided courtesy of G. Lacroix.

May we conclude, on the basis of these data, that the number of alveolar cells in ovalbumin-sensitized guinea pigs differs with type of exposure? Let $\alpha = .05$.

13.8.5 The following table shows the pesticide residue levels (ppb) in blood samples from four populations of human subjects. Use the Kruskal–Wallis test to test at the .05 level of significance the null hypothesis that there is no difference among the populations with respect to average level of pesticide residue.

Population				Population			
A	**B**	**C**	**D**	**A**	**B**	**C**	**D**
10	4	15	7	44	11	9	4
37	35	5	11	12	7	11	5
12	32	10	10	15	32	9	2

(*Continued*)

Population				Population			
A	B	C	D	A	B	C	D
31	19	12	8	42	17	14	6
11	33	6	2	23	8	15	3
9	18	6	5				

13.8.6 Hepatic γ-glutamyl transpeptidase (GGTP) activity was measured in 22 patients undergoing percutaneous liver biopsy. The results were as follows:

Subject	Diagnosis	Hepatic GGTP Level
1	Normal liver	27.7
2	Primary biliary cirrhosis	45.9
3	Alcoholic liver disease	85.3
4	Primary biliary cirrhosis	39.0
5	Normal liver	25.8
6	Persistent hepatitis	39.6
7	Chronic active hepatitis	41.8
8	Alcoholic liver disease	64.1
9	Persistent hepatitis	41.1
10	Persistent hepatitis	35.3
11	Alcoholic liver disease	71.5
12	Primary biliary cirrhosis	40.9
13	Normal liver	38.1
14	Primary biliary cirrhosis	40.4
15	Primary biliary cirrhosis	34.0
16	Alcoholic liver disease	74.4
17	Alcoholic liver disease	78.2
18	Persistent hepatitis	32.6
19	Chronic active hepatitis	46.3
20	Normal liver	39.6
21	Chronic active hepatitis	52.7
22	Chronic active hepatitis	57.2

Can we conclude from these sample data that the average population GGTP level differs among the five diagnostic groups? Let $\alpha = .05$ and find the p value.

13.9 THE FRIEDMAN TWO-WAY ANALYSIS OF VARIANCE BY RANKS

Just as we may on occasion have need of a nonparametric analog to the parametric one-way analysis of variance, we may also find it necessary to analyze the data in a two-way classification by nonparametric methods analogous to the two-way analysis of variance. Such a need may arise because the assumptions necessary for parametric analysis of variance are not met, because the measurement scale employed is weak, or because results

are needed in a hurry. A test frequently employed under these circumstances is the Friedman two-way analysis of variance by ranks (9,10). This test is appropriate whenever the data are measured on, at least, an ordinal scale and can be meaningfully arranged in a two-way classification as is given for the randomized block experiment discussed in Chapter 8. The following example illustrates this procedure.

EXAMPLE 13.9.1

A physical therapist conducted a study to compare three models of low-volt electrical stimulators. Nine other physical therapists were asked to rank the stimulators in order of preference. A rank of 1 indicates first preference. The results are shown in Table 13.9.1. We wish to know if we can conclude that the models are not preferred equally.

Solution:

1. **Data.** See Table 13.9.1.
2. **Assumptions.** The observations appearing in a given block are independent of the observations appearing in each of the other blocks, and within each block measurement on at least an ordinal scale is achieved.
3. **Hypothesis.** In general, the hypotheses are:

   H_0: The treatments all have identical effects.

   H_A: At least one treatment tends to yield larger observations than at least one of the other treatments.

   For our present example we state the hypotheses as follows:

   H_0: The three models are equally preferred.

   H_A: The three models are not equally preferred.

   Let $\alpha = .05$.

TABLE 13.9.1 Physical Therapists' Rankings of Three Models of Low-Volt Electrical Stimulators

	Model		
Therapist	A	B	C
1	2	3	1
2	2	3	1
3	2	3	1
4	1	3	2
5	3	2	1
6	1	2	3
7	2	3	1
8	1	3	2
9	1	3	2
R_j	15	25	14

4. **Test statistic.** By means of the Friedman test we will be able to determine if it is reasonable to assume that the columns of ranks have been drawn from the same population. If the null hypothesis is true we would expect the observed distribution of ranks within any column to be the result of chance factors and, hence, we would expect the numbers 1, 2, and 3 to occur with approximately the same frequency in each column. If, on the other hand, the null hypothesis is false (that is, the models are not equally preferred), we would expect a preponderance of relatively high (or low) ranks in at least one column. This condition would be reflected in the sums of the ranks. The Friedman test will tell us whether or not the observed sums of ranks are so discrepant that it is not likely they are a result of chance when H_0 is true.

   Since the data already consist of rankings within blocks (rows), our first step is to sum the ranks within each column (treatment). These sums are the R_j shown in Table 13.9.1. A test statistic, denoted by Friedman as χ_r^2, is computed as follows:

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(R_j\right)^2 - 3n(k+1) \tag{13.9.1}$$

   where n = the number of rows (blocks) and k = the number of columns (treatments).

5. **Distribution of test statistic.** Critical values for various values of n and k are given in Appendix Table O.

6. **Decision rule.** Reject H_0 if the probability of obtaining (when H_0 is true) a value of χ_r^2 as large as or larger than actually computed is less than or equal to α.

7. **Calculation of test statistic.** Using the data in Table 13.9.1 and Equations 13.9.1, we compute

$$\chi_r^2 = \frac{12}{9(3)(3+1)} \left[(15)^2 + (25)^2 + (14)^2\right] - 3(9)(3+1) = 8.222$$

8. **Statistical decision.** When we consult Appendix Table Oa, we find that the probability of obtaining a value of χ_r^2 as large as 8.222 due to chance alone, when the null hypothesis is true, is .016. We are able, therefore, to reject the null hypothesis.

9. **Conclusion.** We conclude that the three models of low-volt electrical stimulator are not equally preferred.

10. ***p* value.** For this test, $p = .016$. ■

Ties When the original data consist of measurements on an interval or a ratio scale instead of ranks, the measurements are assigned ranks based on their relative magnitudes within blocks. If ties occur, each value is assigned the mean of the ranks for which it is tied.

Large Samples When the values of k and/or n exceed those given in Table O, the critical value of χ_r^2 is obtained by consulting the χ^2 table (Table F) with the chosen α and $k - 1$ degrees of freedom.

EXAMPLE 13.9.2

Table 13.9.2 shows the responses, in percent decrease in salivary flow, of 16 experimental animals following different dose levels of atropine. The ranks (in parentheses) and the sum of the ranks are also given in the table. We wish to see if we may conclude that the different dose levels produce different responses. That is, we wish to test the null hypothesis of no difference in response among the four dose levels.

Solution: From the data, we compute

$$\chi_r^2 = \frac{12}{16(4)(4+1)}\left[(20)^2 + (36.5)^2 + (44)^2 + (59.5)^2\right] - 3(16)(4+1) = 30.32$$

Reference to Table F indicates that with $k - 1 = 3$ degrees of freedom the probability of getting a value of χ_r^2 as large as 30.32 due to chance alone is, when H_0 is true, less than .005. We reject the null hypothesis and conclude that the different dose levels do produce different responses.

TABLE 13.9.2 Percent Decrease in Salivary Flow of Experimental Animals Following Different Dose Levels of Atropine

	Dose Level			
Animal Number	A	B	C	D
1	29(1)	48(2)	75(3)	100(4)
2	72(2)	30(1)	100(3.5)	100(3.5)
3	70(1)	100(4)	86(2)	96(3)
4	54(2)	35(1)	90(3)	99(4)
5	5(1)	43(3)	32(2)	81(4)
6	17(1)	40(2)	76(3)	81(4)
7	74(1)	100(3)	100(3)	100(3)
8	6(1)	34(2)	60(3)	81(4)
9	16(1)	39(2)	73(3)	79(4)
10	52(2)	34(1)	88(3)	96(4)
11	8(1)	42(3)	31(2)	79(4)
12	29(1)	47(2)	72(3)	99(4)
13	71(1)	100(3.5)	97(2)	100(3.5)
14	7(1)	33(2)	58(3)	79(4)
15	68(1)	99(4)	84(2)	93(3)
16	70(2)	30(1)	99(3.5)	99(3.5)
R_j	20	36.5	44	59.5

■

Dialog box:

Stat ➤ Nonparametrics ➤ Friedman

Type *C3* in **Response,** *C1* in **Treatment** and *C2* in **Blocks.** Click **OK.**

Session command:

```
MTB > FRIEDMAN C3 C1 C2
```

Output:

Friedman Test: C3 versus C1 blocked by C2

```
S = 8.22 d.f. = 2 p = 0.017

                    Est.    Sum of
     C1     N      Median    RANKS
      1     9      2.0000     15.0
      2     9      2.6667     25.0
      3     9      1.3333     14.0

Grand median = 2.0000
```

FIGURE 13.9.1 MINITAB procedure and output for Example 13.9.1.

Computer Analysis Many statistics software packages, including MINITAB, will perform the Friedman test. To use MINITAB we form three columns of data. We may, for example, set up the columns so that Column 1 contains numbers that indicate the treatment to which the observations belong, Column 2 contains numbers indicating the blocks to which the observations belong, and Column 3 contains the observations. If we do this for Example 13.9.1, the MINITAB procedure and output are as shown in Figure 13.9.1.

EXERCISES

For the following exercises perform the test at the indicated level of significance and determine the *p* value.

13.9.1 The following table shows the scores made by nine randomly selected student nurses on final examinations in three subject areas:

Student Number	Subject Area		
	Fundamentals	**Physiology**	**Anatomy**
1	98	95	77
2	95	71	79

(*Continued*)

Student Number	Subject Area		
	Fundamentals	Physiology	Anatomy
3	76	80	91
4	95	81	84
5	83	77	80
6	99	70	93
7	82	80	87
8	75	72	81
9	88	81	83

Test the null hypothesis that student nurses constituting the population from which the above sample was drawn perform equally well in all three subject areas against the alternative hypothesis that they perform better in, at least, one area. Let $\alpha = .05$.

13.9.2 Fifteen randomly selected physical therapy students were given the following instructions: "Assume that you will marry a person with one of the following handicaps (the handicaps were listed and designated by the letters A to J). Rank these handicaps from 1 to 10 according to your first, second, third (and so on) choice of a handicap for your marriage partner." The results are shown in the following table.

Student Number	Handicap									
	A	B	C	D	E	F	G	H	I	J
1	1	3	5	9	8	2	4	6	7	10
2	1	4	5	7	8	2	3	6	9	10
3	2	3	7	8	9	1	4	6	5	10
4	1	4	7	8	9	2	3	6	5	10
5	1	4	7	8	10	2	3	6	5	9
6	2	3	7	9	8	1	4	5	6	10
7	2	4	6	9	8	1	3	7	5	10
8	1	5	7	9	10	2	3	4	6	8
9	1	4	5	7	8	2	3	6	9	10
10	2	3	6	8	9	1	4	7	5	10
11	2	4	5	8	9	1	3	7	6	10
12	2	3	6	8	10	1	4	5	7	9
13	3	2	6	9	8	1	4	7	5	10
14	2	5	7	8	9	1	3	4	6	10
15	2	3	6	7	8	1	5	4	9	10

Test the null hypothesis of no preference for handicaps against the alternative that some handicaps are preferred over others. Let $\alpha = .05$.

13.9.3 Ten subjects with exercise-induced asthma participated in an experiment to compare the protective effect of a drug administered in four dose levels. Saline was used as a control. The variable of interest was change in FEV_1 after administration of the drug or saline. The results were as follows:

		Dose Level of Drug (mg/ml)			
Subject	**Saline**	**2**	**10**	**20**	**40**
1	−.68	−.32	−.14	−.21	−.32
2	−1.55	−.56	−.31	−.21	−.16
3	−1.41	−.28	−.11	−.08	−.83
4	−.76	−.56	−.24	−.41	−.08
5	−.48	−.25	−.17	−.04	−.18
6	−3.12	−1.99	−1.22	−.55	−.75
7	−1.16	−.88	−.87	−.54	−.84
8	−1.15	−.31	−.18	−.07	−.09
9	−.78	−.24	−.39	−.11	−.51
10	−2.12	−.35	−.28	+.11	−.41

Can one conclude on the basis of these data that different dose levels have different effects? Let $\alpha = .05$ and find the p value.

13.10 THE SPEARMAN RANK CORRELATION COEFFICIENT

Several nonparametric measures of correlation are available to the researcher. Of these a frequently used procedure that is attractive because of the simplicity of the calculations involved is due to Spearman (11). The measure of correlation computed by this method is called the Spearman rank correlation coefficient and is designated by r_s. This procedure makes use of the two sets of ranks that may be assigned to the sample values of X and Y, the independent and continuous variables of a bivariate distribution.

Hypotheses The usually tested hypotheses and their alternatives are as follows:

(a) H_0: X and Y are mutually independent.
H_A: X and Y are not mutually independent.

(b) H_0: X and Y are mutually independent.
H_A: There is a tendency for large values of X and large values of Y to be paired together.

(c) H_0: X and Y are mutually independent.
H_A: There is a tendency for large values of X to be paired with small values of Y.

The hypotheses specified in (a) lead to a two-sided test and are used when it is desired to detect any departure from independence. The one-sided tests indicated by (b) and (c) are used, respectively, when investigators wish to know if they can conclude that the variables are directly or inversely correlated.

The Procedure The hypothesis-testing procedure involves the following steps.

1. Rank the values of X from 1 to n (numbers of pairs of values of X and Y in the sample). Rank the values of Y from 1 to n.

2. Compute d_i for each pair of observations by subtracting the rank of Y_i from the rank of X_i.
3. Square each d_i and compute $\sum d_i^2$, the sum of the squared values.
4. Compute

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{13.10.1}$$

5. If n is between 4 and 30, compare the computed value of r_s with the critical values, r_s^*, of Appendix Table P. For the two-sided test, H_0 is rejected at the α significance level if r_s is greater than r_s^* or less than $-r_s^*$, where r_s^* is at the intersection of the column headed $\alpha/2$ and the row corresponding to n. For the one-sided test with H_A specifying direct correlation, H_0 is rejected at the α significance level if r_s is greater than r_s^* for α and n. The null hypothesis is rejected at the α significance level in the other one-sided test if r_s is less than $-r_s^*$ for α and n.
6. If n is greater than 30, one may compute

$$z = r_s\sqrt{n-1} \tag{13.10.2}$$

and use Appendix Table D to obtain critical values.

7. Tied observations present a problem. The use of Table P is strictly valid only when the data do not contain any ties (unless some random procedure for breaking ties is employed). In practice, however, the table is frequently used after some other method for handling ties has been employed. If the number of ties is large, the following correction for ties may be employed:

$$T = \frac{t^3 - t}{12} \tag{13.10.3}$$

where t = the number of observations that are tied for some particular rank. When this correction factor is used, r_s is computed from

$$r_s = \frac{\sum x^2 + \sum y^2 - \sum d_i^2}{2\sqrt{\sum x^2 \sum y^2}} \tag{13.10.4}$$

instead of from Equation 13.10.1.

In Equation 13.10.4,

$$\sum x^2 = \frac{n^3 - n}{12} - \sum T_x$$
$$\sum y^2 = \frac{n^3 - n}{12} - \sum T_y$$

T_x = the sum of the values of T for the various tied ranks in X
T_y = the sum of the values of T for the various tied ranks in Y

Most authorities agree that unless the number of ties is excessive, the correction makes very little difference in the value of r_s. When the number of ties is small, we can follow the

usual procedure of assigning the tied observations the mean of the ranks for which they are tied and proceed with steps 2 to 6.

EXAMPLE 13.10.1

In a study of the relationship between age and the EEG, data were collected on 20 subjects between ages 20 and 60 years. Table 13.10.1 shows the age and a particular EEG output value for each of the 20 subjects. The investigator wishes to know if it can be concluded that this particular EEG output is inversely correlated with age.

Solution:

1. **Data.** See Table 13.10.1.
2. **Assumptions.** We assume that the sample available for analysis is a simple random sample and that both X and Y are measured on at least the ordinal scale.
3. **Hypotheses.**

   H_0: This EEG output and age are mutually independent.

   H_A: There is a tendency for this EEG output to decrease with age.

   Suppose we let $\alpha = .05$.

TABLE 13.10.1 Age and EEG Output Value for 20 Subjects

Subject Number	Age (X)	EEG Output Value (Y)
1	20	98
2	21	75
3	22	95
4	24	100
5	27	99
6	30	65
7	31	64
8	33	70
9	35	85
10	38	74
11	40	68
12	42	66
13	44	71
14	46	62
15	48	69
16	51	54
17	53	63
18	55	52
19	58	67
20	60	55

4. **Test statistic.** See Equation 13.10.1.
5. **Distribution of test statistic.** Critical values of the test statistic are given in Appendix Table P.
6. **Decision rule.** For the present test we will reject H_0 if the computed value of r_s is less than $-.3789$.
7. **Calculation of test statistic.** When the X and Y values are ranked, we have the results shown in Table 13.10.2. The d_i, d_i^2, and $\sum d_i^2$ are shown in the same table.

   Substitution of the data from Table 13.10.2 into Equation 13.10.1 gives

$$r_s = 1 - \frac{6(2340)}{20[(20)^2 - 1]} = -.76$$

8. **Statistical decision.** Since our computed $r_s = -.76$ is less than the critical r_s^*, we reject the null hypothesis.
9. **Conclusion.** We conclude that the two variables are inversely related.
10. ***p* value.** Since $-.76 < -0.6586$, we have for this test $p < .001$.

TABLE 13.10.2 Ranks for Data of Example 13.10.1

Subject Number	Rank (X)	Rank (Y)	d_i	d_i^2
1	1	18	−17	289
2	2	15	−13	169
3	3	17	−14	196
4	4	20	−16	256
5	5	19	−14	196
6	6	7	−1	1
7	7	6	1	1
8	8	12	−4	16
9	9	16	−7	49
10	10	14	−4	16
11	11	10	1	1
12	12	8	4	16
13	13	13	0	0
14	14	4	10	100
15	15	11	4	16
16	16	2	14	196
17	17	5	12	144
18	18	1	17	289
19	19	9	10	100
20	20	3	17	289
				$\sum d_i^2 = 2340$

■

Let us now illustrate the procedure for a sample with $n > 30$ and some tied observations.

EXAMPLE 13.10.2

In Table 13.10.3 are shown the ages and concentrations (ppm) of a certain mineral in the tissue of 35 subjects on whom autopsies were performed as part of a large research project.

The ranks, d_i, d_i^2, and $\sum d_i^2$ are shown in Table 13.10.4. Let us test, at the .05 level of significance, the null hypothesis that X and Y are mutually independent against the two-sided alternative that they are not mutually independent.

Solution: From the data in Table 13.10.4 we compute

$$r_s = 1 - \frac{6(1788.5)}{35\left[(35)^2 - 1\right]} = .75$$

To test the significance of r_s we compute

$$z = .75\sqrt{35 - 1} = 4.37$$

TABLE 13.10.3 Age and Mineral Concentration (ppm) in Tissue of 35 Subjects

Subject Number	Age (X)	Mineral Concentration (Y)	Subject Number	Age (X)	Mineral Concentration (Y)
1	82	169.62	19	50	4.48
2	85	48.94	20	71	46.93
3	83	41.16	21	54	30.91
4	64	63.95	22	62	34.27
5	82	21.09	23	47	41.44
6	53	5.40	24	66	109.88
7	26	6.33	25	34	2.78
8	47	4.26	26	46	4.17
9	37	3.62	27	27	6.57
10	49	4.82	28	54	61.73
11	65	108.22	29	72	47.59
12	40	10.20	30	41	10.46
13	32	2.69	31	35	3.06
14	50	6.16	32	75	49.57
15	62	23.87	33	50	5.55
16	33	2.70	34	76	50.23
17	36	3.15	35	28	6.81
18	53	60.59			

TABLE 13.10.4 Ranks for Data of Example 13.10.2

Subject Number	Rank (X)	Rank (Y)	d_i	d_i^2	Subject Number	Rank (X)	Rank (Y)	d_i	d_i^2
1	32.5	35	−2.5	6.25	19	17	9	8	64.00
2	35	27	8	64.00	20	28	25	3	9.00
3	34	23	11	121.00	21	21.5	21	.5	.25
4	25	32	−7	49.00	22	23.5	22	1.5	2.25
5	32.5	19	13.5	182.25	23	13.5	24	−10.5	110.25
6	19.5	11	8.5	72.25	24	27	34	−7	49.00
7	1	14	−13	169.00	25	6	3	3	9.00
8	13.5	8	5.5	30.25	26	12	7	5	25.00
9	9	6	3	9.00	27	2	15	−13	169.00
10	15	10	5	25.00	28	21.5	31	−9.5	90.25
11	26	33	−7	49.00	29	29	26	3	9.00
12	10	17	−7	49.00	30	11	18	−7	49.00
13	4	1	3	9.00	31	7	4	3	9.00
14	17	13	4	16.00	32	30	28	2	4.00
15	23.5	20	3.5	12.25	33	17	12	5	25.00
16	5	2	3	9.00	34	31	29	2	4.00
17	8	5	3	9.00	35	3	16	−13	169.00
18	19.5	30	−10.5	110.25					
									$\sum d_i^2 = 1788.5$

Since 4.37 is greater than $z = 3.89$, $p < 2(.0001) = .0002$, and we reject H_0 and conclude that the two variables under study are not mutually independent.

For comparative purposes let us correct for ties using Equation 13.10.3 and then compute r_s by Equation 13.10.4.

In the rankings of X we had six groups of ties that were broken by assigning the values 13.5, 17, 19.5, 21.5, 23.5, and 32.5. In five of the groups two observations tied, and in one group three observations tied. We, therefore, compute five values of

$$T_x = \frac{2^3 - 2}{12} = \frac{6}{12} = .5$$

and one value of

$$T_x = \frac{3^3 - 3}{12} = \frac{24}{12} = 2$$

From these computations, we have $\sum T_x = 5(.5) + 2 = 4.5$, so that

$$\sum x^2 = \frac{35^2 - 35}{12} - 4.5 = 3565.5$$

Dialog box:

Stat ➤ Basic Statistics ➤ Correlation

Type *C3–C4* in **Variables.** Click **OK.**

Session command:

```
MTB > CORRELATION C3 C4
```

Output:

Correlations (Pearson)

```
Correlation of (X)Rank and (Y)Rank = -0.759
```

FIGURE 13.10.1 MINITAB procedure and output for computing Spearman rank correlation coefficient, Example 13.10.1.

Since no ties occurred in the Y rankings, we have $\sum T_y = 0$ and

$$\sum y^2 = \frac{35^3 - 35}{12} - 0 = 3570.0$$

From Table 13.10.4 we have $\sum d_i^2 = 1788.5$. From these data we may now compute by Equation 13.10.4

$$r_s = \frac{3565.5 + 3570.0 - 1788.5}{2\sqrt{(3565.5)(3570)}} = .75$$

We see that in this case the correction for ties does not make any difference in the value of r_s. ■

Computer Analysis We may use MINITAB, as well as many other statistical software packages, to compute the Spearman correlation coefficient. To use MINITAB, we must first have MINITAB rank the observations and store the ranks in separate columns, one for the X ranks and one for the Y ranks. If we rank the X and Y values of Example 13.10.1 and store them in Columns 3 and 4, we may obtain the Spearman rank correlation coefficient with the procedure shown in Figure 13.10.1. Other software packages such as SAS® and SPSS, for example, automatically rank the measurements before computing the coefficient, thereby eliminating an extra step in the procedure.

EXERCISES

For the following exercises perform the test at the indicated level of significance and determine the p value.

13.10.1 The following table shows 15 randomly selected geographic areas ranked by population density and age-adjusted death rate. Can we conclude at the .05 level of significance that population density and age-adjusted death rate are not mutually independent?

Area	Rank by Population Density (X)	Rank by Age-Adjusted Death Rate (Y)	Area	Rank by Population Density (X)	Rank by Age-Adjusted Death Rate (Y)
1	8	10	9	6	8
2	2	14	10	14	5
3	12	4	11	7	6
4	4	15	12	1	2
5	9	11	13	13	9
6	3	1	14	15	3
7	10	12	15	11	13
8	5	7			

13.10.2 The following table shows 10 communities ranked by decayed, missing, or filled (DMF) teeth per 100 children and fluoride concentration in ppm in the public water supply:

Community	Rank by DMF Teeth per 100 Children (X)	Rank by Fluoride Concentration (Y)	Community	Rank by DMF Teeth per 100 Children (X)	Rank by Fluoride Concentration (Y)
1	8	1	6	4	7
2	9	3	7	1	10
3	7	4	8	5	6
4	3	9	9	6	5
5	2	8	10	10	2

Do these data provide sufficient evidence to indicate that the number of DMF teeth per 100 children tends to decrease as fluoride concentration increases? Let $\alpha = .05$.

13.10.3 The purpose of a study by Nozawa et al. (A-10) was to evaluate the outcome of surgical repair of pars interarticularis defect by segmental wire fixation in young adults with lumbar spondylolysis. The authors cite literature indicating that segmental wire fixation has been successful in the treatment of nonathletes with spondylolysis and point out that no information existed on the results of this type of surgery in athletes. In a retrospective study of subjects having surgery between 1993 and 2000, the authors found 20 subjects who had undergone the surgery. The following table shows the age (years) at surgery and duration (months) of follow-up care for these subjects.

Duration of Follow-Up (Months)	Age (Years)	Duration of Follow-Up (Months)	Age (Years)
103	37	38	27
68	27	36	31
62	12	34	24
60	18	30	23
60	18	19	14

(*Continued*)

Duration of Follow-Up (Months)	Age (Years)	Duration of Follow-Up (Months)	Age (Years)
54	28	19	23
49	25	19	18
44	20	19	29
42	18	17	24
41	30	16	27

Source: Satoshi Nozawa, Katsuji Shimizu, Kei Miyamoto, and Mizuo Tanaka, "Repair of Pars Interarticularis Defect by Segmental Wire Fixation in Young Athletes with Spondylolysis," *American Journal of Sports Medicine*, *31* (2003), pp. 359–364.

May we conclude, on the basis of these data, that in a population of similar subjects there is an association between age and duration of follow-up? Let $\alpha = .05$.

13.10.4 Refer to Exercise 13.10.3. Nozawa et al. (A-10) also calculated the Japanese Orthopaedic Association score for measuring back pain (JOA). The results for the 20 subjects along with the duration of follow-up are shown in the following table. The higher the number, the lesser the degree of pain.

Duration of Follow-Up (Months)	JOA	Duration of Follow-Up (Months)	JOA
103	21	38	13
68	14	36	24
62	26	34	21
60	24	30	22
60	13	19	25
54	24	19	23
49	22	19	20
44	23	19	21
42	18	17	25
41	24	16	21

Source: Satoshi Nozawa, Katsuji Shimizu, Kei Miyamoto, and Mizuo Tanaka, "Repair of Pars Interarticularis Defect by Segmental Wire Fixation in Young Athletes with Spondylolysis," *American Journal of Sports Medicine*, *31* (2003), pp. 359–364.

Can we conclude from these data that in general there is a relationship between length of follow-up and JOA score at the time of the operation? Let $\alpha = .05$.

13.10.5 Butz et al. (A-11) studied the use of noninvasive positive-pressure ventilation by patients with amyotrophic lateral sclerosis. They evaluated the benefit of the procedure on patients' symptoms, quality of life, and survival. Two variables of interest are $PaCO_2$, partial pressure of arterial carbon dioxide, and PaO_2, partial pressure of arterial oxygen. The following table shows, for 30 subjects, values of these variables (mm Hg) obtained from baseline arterial blood gas analyses.

$PaCO_2$	PaO_2	$PaCO_2$	PaO_2	$PaCO_2$	PaO_2
40	101	54.5	80	34.5	86.5
47	69	54	72	40.1	74.7

(*Continued*)

$PaCO_2$	PaO_2	$PaCO_2$	PaO_2	$PaCO_2$	PaO_2
34	132	43	105	33	94
42	65	44.3	113	59.9	60.4
54	72	53.9	69.2	62.6	52.5
48	76	41.8	66.7	54.1	76.9
53.6	67.2	33	67	45.7	65.3
56.9	70.9	43.1	77.5	40.6	80.3
58	73	52.4	65.1	56.6	53.2
45	66	37.9	71	59	71.9

Source: M. Butz, K. H. Wollinsky, U. Widemuth-Catrinescu, A. Sperfeld, S. Winter, H. H. Mehrkens, A. C. Ludolph, and H. Schreiber, "Longitudinal Effects of Noninvasive Positive-Pressure Ventilation in Patients with Amyotrophic Lateral Sclerosis," *American Journal of Medical Rehabilitation, 82* (2003) 597–604.

On the basis of these data may we conclude that there is an association between $PaCO_2$ and PaO_2 values? Let $\alpha = .05$.

13.10.6 Seventeen patients with a history of congestive heart failure participated in a study to assess the effects of exercise on various bodily functions. During a period of exercise the following data were collected on the percent change in plasma norepinephrine (Y) and the percent change in oxygen consumption (X):

Subject	X	Y	Subject	X	Y
1	500	525	10	50	60
2	475	130	11	175	105
3	390	325	12	130	148
4	325	190	13	76	75
5	325	90	14	200	250
6	205	295	15	174	102
7	200	180	16	201	151
8	75	74	17	125	130
9	230	420			

On the basis of these data can one conclude that there is an association between the two variables? Let $\alpha = .05$.

13.11 NONPARAMETRIC REGRESSION ANALYSIS

When the assumptions underlying simple linear regression analysis as discussed in Chapter 9 are not met, we may employ nonparametric procedures. In this section we present estimators of the slope and intercept that are easy-to-calculate alternatives to the least-squares estimators described in Chapter 9.

Theil's Slope Estimator Theil (12) proposes a method for obtaining a point estimate of the slope coefficient β. We assume that the data conform to the classic regression model

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i, \quad i = 1, \ldots, n$$

where the x_i are known constants, β_0 and β_1 are unknown parameters, and Y_i is an observed value of the continuous random variable Y at x_i. For each value of x_i, we assume a subpopulation of Y values, and the ε_i are mutually independent. The x_i are all distinct (no ties), and we take $x_1 < x_2 < \cdots < x_n$.

The data consist of n pairs of sample observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where the ith pair represents measurements taken on the ith unit of association.

To obtain Theil's estimator of β_1 we first form all possible sample slopes $S_{ij} = (y_j - y_i)/(x_j - x_i)$, where $i < j$. There will be $N = {}_nC_2$ values of S_{ij}. The estimator of β_1 which we designate by $\hat{\beta}_1$ is the median of S_{ij} values. That is,

$$\hat{\beta}_1 = \text{median}\{S_{ij}\} \tag{13.11.1}$$

The following example illustrates the calculation of $\hat{\beta}_1$.

EXAMPLE 13.11.1

In Table 13.11.1 are the plasma testosterone (ng/ml) levels (Y) and seminal citric acid (mg/ml) levels in a sample of eight adult males. We wish to compute the estimate of the population regression slope coefficient by Theil's method.

Solution: The $N = {}_8C_2 = 28$ ordered values of S_{ij} are shown in Table 13.11.2.

If we let $i = 1$ and $j = 2$, the indicators of the first and second values of Y and X in Table 13.11.1, we may compute S_{12} as follows:

$$S_{12} = (175 - 230)/(278 - 421) = -.3846$$

When all the slopes are computed in a similar manner and ordered as in Table 13.11.2, $-.3846$ winds up as the tenth value in the ordered array.

The median of the S_{ij} values is .4878. Consequently, our estimate of the population slope coefficient $\hat{\beta}_1 = .4878$.

TABLE 13.11.1 Plasma Testosterone and Seminal Citric Acid Levels in Adult Males

Testosterone:	230	175	315	290	275	150	360	425
Citric acid:	421	278	618	482	465	105	550	750

TABLE 13.11.2 Ordered Values of S_{ij} for Example 13.11.1

−.6618	.5037
.1445	.5263
.1838	.5297
.2532	.5348
.2614	.5637
.3216	.5927
.3250	.6801
.3472	.8333
.3714	.8824
.3846	.9836
.4118	1.0000
.4264	1.0078
.4315	1.0227
.4719	1.0294

■

An Estimator of the Intercept Coefficient Dietz (13) recommends two intercept estimators. The first, designated $\left(\hat{\beta}_0\right)_{1,M}$ is the median of the n terms $y_i - \hat{\beta}_1 x_i$ in which $\hat{\beta}_1$ is the Theil estimator. It is recommended when the researcher is not willing to assume that the error terms are symmetric about 0. If the researcher is willing to assume a symmetric distribution of error terms, Dietz recommends the estimator $\left(\hat{\beta}_0\right)_{2,M}$ which is the median of the $n(n+1)/2$ pairwise averages of the $y_i - \hat{\beta}_1 x_i$ terms. We illustrate the calculation of each in the following example.

EXAMPLE 13.11.2

Refer to Example 13.11.1. Let us compute $\hat{\alpha}_{1,M}$ and $\hat{\alpha}_{2,M}$ from the data on testosterone and citric acid levels.

Solution: The ordered $y_i - .4878x_i$ terms are: 13.5396, 24.6362, 39.3916, 48.1730, 54.8804, 59.1500, 91.7100, and 98.7810. The median, 51.5267, is the estimator $\left(\hat{\beta}_0\right)_{1,M}$.

The $8(8+1)/2 = 36$ ordered pairwise averages of the $y_i - .4878x_i$ are

13.5396	49.2708	75.43
19.0879	51.5267	76.8307
24.6362	52.6248	78.9655
26.4656	53.6615	91.71
30.8563	54.8804	95.2455
32.0139	56.1603	98.781
34.21	57.0152	
36.3448	58.1731	

(*Continued*)

36.4046	59.15
39.3916	61.7086
39.7583	65.5508
41.8931	69.0863
43.7823	69.9415
47.136	73.2952
48.173	73.477

The median of these averages, 53.1432, is the estimator $\hat{\alpha}_{2,M}$. The estimating equation, then, is $y_i = 53.1432 + .4878x_i$ if we are willing to assume that the distribution of error terms is symmetric about 0. If we are not willing to make the assumption of symmetry, the estimating equation is $y_i = 51.5267 + .4878x_i$. ■

EXERCISES

13.11.1 The following are the heart rates (HR: beats/minute) and oxygen consumption values (VO_2: cal/kg/24 h) for nine infants with chronic congestive heart failure:

HR(X):	163	164	156	151	152	167	165	153	155
VO_2(Y):	53.9	57.4	41.0	40.0	42.0	64.4	59.1	49.9	43.2

Compute $\hat{\beta}_1$, $\left(\hat{\beta}_0\right)_{1,M}$, and $\left(\hat{\beta}_0\right)_{2,M}$

13.11.2 The following are the body weights (grams) and total surface area (cm^2) of nine laboratory animals:

Body weight (X):	660.2	706.0	924.0	936.0	992.1	888.9	999.4	890.3	841.2
Surface area (Y):	781.7	888.7	1038.1	1040.0	1120.0	1071.5	1134.5	965.3	925.0

Compute the slope estimator and two intercept estimators.

13.12 SUMMARY

This chapter is concerned with nonparametric statistical tests. These tests may be used either when the assumptions underlying the parametric tests are not realized or when the data to be analyzed are measured on a scale too weak for the arithmetic procedures necessary for the parametric tests.

Nine nonparametric tests are described and illustrated. Except for the Kolmogorov–Smirnov goodness-of-fit test, each test provides a nonparametric alternative to a well-known parametric test. There are a number of other nonparametric tests available. The interested reader is referred to the many books devoted to nonparametric methods, including those by Gibbons (14) and Pett (15).

SUMMARY OF FORMULAS FOR CHAPTER 13

Formula Number	Name	Formula
13.3.1	Sign test statistic	$P(k \le x \mid n,p) = \sum_{k=0}^{x} {}_nC_k\, p^k q^{n-k}$
13.3.2	Large-sample approximation of the sign test	$z = \dfrac{(k+0.5)+0.5n}{0.5\sqrt{n}}, \quad \text{if} \quad k < \dfrac{n}{2}$ $z = \dfrac{(k-0.5)-0.5n}{0.5\sqrt{n}}, \quad \text{if} \quad k \ge \dfrac{n}{2}$
13.6.1	Mann–Whitney test statistic	$T = S - \dfrac{n(n+1)}{2}$
13.6.2	Large-sample approximation of the Mann–Whitney test	$Z = \dfrac{T - mn/2}{\sqrt{nm(n+m+1)/12}}$
13.6.3	Equivalence of the Mann–Whitney and Wilcoxon two-sample statistics	$U + W = \dfrac{m(m+2n+1)}{2}$
13.7.1–13.7.2	Kolmogorov–Smirnov test statistic	$D = \sup_x \lvert F_s(x) - F_T(x)\rvert$ $= \max_{1\le i\le r} \{\max[\lvert F_s(x_i) - F_T(x_i)\rvert, \lvert F_s(x_{i-1}) - F_T(x_{i-1})\rvert]\}$
13.8.1	Kruskal–Wallis test statistic	$H = \dfrac{12}{n(n+1)} \sum_{j=1}^{k} \dfrac{R_j^2}{n_j} - 3(n+1)$
13.8.2	Kruskal–Wallis test statistic adjustment for ties	$1 - \dfrac{\sum T}{n^3 - n}$
13.9.2	Friedman test statistic	$\chi_r^2 = \dfrac{12}{nk(k+1)} \sum_{j=1}^{k} (R_j)^2 - 3n(k+1)$
13.10.1	Spearman rank correlation test statistic	$r_s = 1 - \dfrac{6\sum d_i^2}{n(n^2-1)}$

13.10.2	Large-sample approximation of the Spearman rank correlation	$z = r_s\sqrt{n-1}$
13.10.3–13.10.4	Correction for tied observations in the Spearman rank correlation	$T = \dfrac{t^3 - t}{12}$ with $r_s = \dfrac{\sum x^2 + \sum y^2 - \sum d_i^2}{2\sqrt{\sum x^2 \sum y^2}}$
13.11.1	Theil's estimator of β	$\hat{\beta} = median\{S_i\}$
Symbol Key	• $\hat{\beta}$ = Theil's estimator of β • $\chi^2 (\text{or} X^2)$ = chi-square • D = Kolmogorov − Smirnov test statistic • $F_i(x)$ = distribution function of i • H = Friedman test ststictic • k = sign test statistic and the number of columns in the Friedman test • m = sample size of the smaller of two samples • n = sample size of the larger of two samples • p = probability of success • $q = 1 - p$ = probability of failure • R = rank • r_s = Spearman rank correlation coefficient • S = sum of ranks • S_{ij} = slope between point i and j • sup = supremum (greatest) • t = number of tied observations • T = correction for tied observations • x and y = data value for variables x and y • U = Mann−Whitney test ststistic • W = Wilcoxon test ststistic • z = normal variate	

REVIEW QUESTIONS AND EXERCISES

1. Define nonparametric statistics.
2. What is meant by the term *distribution-free statistical tests*?
3. What are some of the advantages of using nonparametric statistical tests?
4. What are some of the disadvantages of the nonparametric tests?

5. Describe a situation in your particular area of interest where each of the following tests could be used. Use real or realistic data and test an appropriate hypothesis using each test.
   (a) The sign test
   (b) The median test
   (c) The Wilcoxon test
   (d) The Mann–Whitney test
   (e) The Kolmogorov–Smirnov goodness-of-fit test
   (f) The Kruskal–Wallis one-way analysis of variance by ranks
   (g) The Friedman two-way analysis of variance by ranks
   (h) The Spearman rank correlation coefficient
   (i) Nonparametric regression analysis

6. The following are the ranks of the ages (X) of 20 surgical patients and the dose (Y) of an analgesic agent required to block one spinal segment.

Rank of Age in Years (X)	Rank of Dose Requirement (Y)	Rank of Age in Years (X)	Rank of Dose Requirement (Y)
1	1	11	13
2	7	12	5
3	2	13	11
4	4	14	16
5	6	15	20
6	8	16	18
7	3	17	19
8	15	18	17
9	9	19	10
10	12	20	14

   Compute r_s and test (two-sided) for significance. Let $\alpha = .05$. Determine the p value for this test.

7. Otani and Kishi (A-12) studied seven subjects with diabetic macular edema. They measured the foveal thickness (μm) in seven eyes pre- and post-unilateral vitrectomy surgery. The results are shown in the following table:

Subject	Pre-op Foveal Thickness (μm)	Post-op Foveal Thickness (μm)
1	690	200
2	840	280
3	470	230
4	690	200
5	730	560
6	500	210
7	440	200

Source: Data provided courtesy of Tomohiro Otani, M.D.

   Use the Wilcoxon signed-rank test to determine whether one should conclude that the surgery is effective in reducing foveal thickness. Let $\alpha = .05$. What is the p value?

8. The subjects of a study by J. Jose and S. R. Ell (A-13) were 303 healthy volunteers who self-assessed their own nasal flow status by indicating whether their nasal airway was (1) totally clear, (2) not very clear, (3) very blocked, or (4) totally blocked. Following the self-assessment, an In-Check meter was used to measure peak inspiratory nasal flow rate (PINFR, L/min). Data on 175 subjects in three of the self-assessment categories are displayed in the following table. The authors performed a Kruskal–Wallis test to determine if these data provide sufficient evidence to indicate a difference in population centers of PINFR among these three response groups. Let $\alpha = .01$. What is the test statistic value for this test?

Peak Inspiratory Nasal Flow Rate (L/min)							
Totally Clear				Not Very Clear			Partially Blocked
180	105	150	120	160	190	130	100
150	150	110	95	200	95	110	100
200	240	130	140	70	130	110	100
130	120	100	135	75	240	130	105
200	90	170	100	150	180	125	95
120	135	80	130	80	140	100	85
150	110	125	180	130	150	230	50
150	155	115	155	160	130	110	105
160	105	140	130	180	90	270	200
150	140	140	140	90	115	180	
110	200	95	120	180	130	130	
190	170	110	290	140	210	125	
150	150	160	170	230	190	90	
120	120	90	280	220	135	210	
180	170	135	150	130	130	140	
140	200	110	185	180	210	125	
130	160	130	150	140	90	210	
230	180	170	150	140	125	120	
200	170	130	170	120	140	115	
140	160	115	210	140	160	100	
150	150	145	140	150	230	130	
170	100	130	140	190	100	130	
180	100	170	160	210	120	110	
160	180	160	120	130	120	150	
200	130	90	230	190	150	110	
90	200	110	100	220	110	90	
130	120	130	190	160	150	120	
140	145	130	90	105	130	115	
200	130	120	100	120	150	140	
220	100	130	125	140	130	130	
200	130	180	180	130	145	160	
120	160	140	200	115	160	110	
310	125	175	160	115	120	165	
160	100	185	170	100	220	120	
115	140	190	85	150	145	150	

(*Continued*)

Peak Inspiratory Nasal Flow Rate (L/min)							
Totally Clear				Not Very Clear			Partially Blocked
170	185	130	150	130	150	170	
130	180	160	280	130	120	110	
220	115	160	140	170	155	120	
250	260	130	100	130	100	85	
160	160	135	140	145	140		
130	170	130	90				
130	115	120	190				
150	150	190	130				
160	130	170					

Source: Data provided courtesy of J. Jose, MS, FRCS.

9. Ten subjects with bronchial asthma participated in an experiment to evaluate the relative effectiveness of three drugs. The following table shows the change in FEV_1 (forced expired volume in 1 second) values (expressed as liters) 2 hours after drug administration:

	Drug				Drug		
Subject	A	B	C	Subject	A	B	C
1	.00	.13	.26	6	.03	.18	.25
2	.04	.17	.23	7	.05	.21	.32
3	.02	.20	.21	8	.02	.23	.38
4	.02	.27	.19	9	.00	.24	.30
5	.04	.11	.36	10	.12	.08	.30

Are these data sufficient to indicate a difference in drug effectiveness? Let $\alpha = .05$. What is the p value for this test?

10. One facet of the nursing curriculum at Wright State University requires that students use mathematics to perform appropriate dosage calculations. In a study by Wendy Gantt (A-14), undergraduate nursing students were given a standardized mathematics test to determine their mathematical aptitude (scale: 0–100). The students were divided into two groups: traditional college age (18–24 years, 26 observations) and nontraditional (25+, eight observations). Scores on the mathematics test appear in the following table:

Traditional Students' Scores			Nontraditional Students' Scores
70	6	88	77
57	79	68	72
85	14	88	54
55	82	92	87
87	45	85	85

(*Continued*)

Traditional Students' Scores			Nontraditional Students' Scores
84	57	56	62
56	91	31	77
68	76	80	86
94	60		

Source: Data provided courtesy of Wendy Gantt and the Wright State University Statistical Consulting Center.

Do these data provide sufficient evidence to indicate a difference in population medians? Let $\alpha = .05$. What is the p value for this test? Use both the median test and the Mann–Whitney test and compare the results.

11. The following are the $PaCO_2$ (mm Hg) values in 16 patients with bronchopulmonary disease:

39, 40, 45, 48, 49, 56, 60, 75, 42, 48, 32, 37, 32, 33, 33, 36

Use the Kolmogorov–Smirnov test to test the null hypothesis that $PaCO_2$ values in the sampled population are normally distributed with $\mu = 44$ and $\sigma = 12$.

12. The following table shows the caloric intake (cal/day/kg) and oxygen consumption VO_2 (ml/min/kg) in 10 infants:

Calorie Intake (X)	VO_2(Y)	Calorie Intake (X)	VO_2(Y)
50	7.0	100	10.8
70	8.0	150	12.0
90	10.5	110	10.0
120	11.0	75	9.5
40	9.0	160	11.9

Test the null hypothesis that the two variables are mutually independent against the alternative that they are directly related. Let $\alpha = 0.5$. What is the p value for this test?

13. Mary White (A-15) surveyed physicians to measure their opinions regarding the importance of ethics in medical practice. The measurement tool utilized a scale from 1 to 5 in which a higher value indicated higher opinion of the importance of ethics. The ages and scores of the study subjects are shown in the following table. Can one conclude on the basis of these results that age and ethics score are directly related? Let the probability of committing a type I error be .05. What is the p value?

Age	Ethics	Age	Ethics	Age	Ethics
25	4.00	26	4.50	26	4.50
34	4.00	29	4.75	27	5.00
30	4.25	30	4.25	22	3.75
31	3.50	26	4.50	22	4.25
25	4.75	30	4.25	24	4.50

(*Continued*)

Age	Ethics	Age	Ethics	Age	Ethics
25	3.75	25	3.75	22	4.25
25	4.75	24	4.75	24	3.75
29	4.50	24	4.00	38	4.50
29	4.50	25	4.50	22	4.50
26	3.75	25	4.00	22	4.50
25	3.25	26	4.75	25	4.00
29	4.50	34	3.25	23	3.75
27	3.75	23	4.50	22	4.25
29	4.25	26	3.25	23	4.00
25	3.75	23	5.00	22	4.25
25	4.50	24	4.25	25	3.50
25	4.00	45	3.25	26	4.25
26	4.25	23	3.75	25	4.25
26	4.00	25	3.75	27	4.75
24	4.00	25	3.75	23	3.75
25	4.00	23	3.75	22	4.00
22	3.75	23	4.75	26	4.75
26	4.50	26	4.00	22	4.25
				23	4.00

Source: Data provided courtesy of Mary White, Ph.D. and Wright State University Statistical Consulting Center.

14. Dominic Sprott (A-16) conducted an experiment with rabbits in which the outcome variable was the fatty infiltration in the shoulder mass (PFI, measured as a percent). At baseline, 15 rabbits had a randomly chosen shoulder muscle detached. The shoulder was then reattached. Six weeks later, five randomly chosen rabbits were sacrificed and the differences in the PFI between the reattached shoulder and the nondetached shoulder were recorded (group A). Six months later, the 10 remaining rabbits were sacrificed and again the differences in the PFI between the reattached shoulder and the nondetached shoulder were recorded (group B).

Percent Fatty Infiltration Difference (Nondetached–Reattached)		
Group A	**Group B**	
2.55	1.04	1.38
0.9	3.29	0.75
0.2	0.99	0.36
−0.29	1.79	0.74
1.11	−0.85	0.3

Source: Data provided courtesy of Dominic Sprott, M.D. and the Wright State University Statistical Consulting Center.

Can we conclude, at the .05 level of significance, that the treatments have a differential effect on PFI between the two shoulder muscles? What is the p value for the test?

In each of the Exercises 15 through 29, do one or more of the following that you think are appropriate:

(a) Apply one or more of the techniques discussed in this chapter.

(b) Apply one or more of the techniques discussed in previous chapters.

(c) Formulate relevant hypotheses, perform the appropriate tests, and find p values.

(d) State the statistical decisions and clinical conclusions that the results of your hypothesis tests justify.

(e) Describe the population(s) to which you think your inferences are applicable.

(f) State the assumptions necessary for the validity of your analyses.

15. The purpose of a study by Damm et al. (A-17) was to investigate insulin sensitivity and insulin secretion in women with previous gestational diabetes (GDM). Subjects were 12 normal-weight glucose-tolerant women (mean age, 36.6 years; standard deviation, 4.16) with previous gestational diabetes and 11 controls (mean age, 35 years; standard deviation, 3.3). Among the data collected were the following fasting plasma insulin values (mmol/L). Use the Mann–Whitney test to determine if you can conclude on the basis of these data that the two populations represented differ with respect to average fasting plasma insulin level.

Controls	Previous GDM	Controls	Previous GDM
46.25	30.00	40.00	31.25
40.00	41.25	30.00	56.25
31.25	56.25	51.25	61.25
38.75	45.00	32.50	50.00
41.25	46.25	43.75	53.75
38.75	46.25		62.50

Source: Data provided courtesy of Dr. Peter Damm.

16. Gutin et al. (A-18) compared three measures of body composition, including dual-energy x-ray absorptiometry (DXA). Subjects were apparently healthy children (21 boys and 22 girls) between the ages of 9 and 11 years. Among the data collected were the following measurements of body-composition compartments by DXA. The investigators were interested in the correlation between all possible pairs of these variables.

Percent Fat	Fat Mass	Fat-Free Mass	Bone Mineral Content	Fat-Free Soft Tissue
11.35	3.8314	29.9440	1.19745	28.7465
22.90	6.4398	21.6805	0.79250	20.8880
12.70	4.0072	27.6290	0.95620	26.6728
42.20	24.0329	32.9164	1.45740	31.4590
24.85	9.4303	28.5009	1.32505	27.1758
26.25	9.4292	26.4344	1.17412	25.2603
23.80	8.4171	26.9938	1.11230	25.8815
37.40	20.2313	33.8573	1.40790	32.4494
14.00	3.9892	24.4939	0.95505	23.5388
19.35	7.2981	30.3707	1.45545	28.9153
29.35	11.1863	26.8933	1.17775	25.7156

(*Continued*)

Percent Fat	Fat Mass	Fat-Free Mass	Bone Mineral Content	Fat-Free Soft Tissue
18.05	5.8449	26.5341	1.13820	25.3959
13.95	4.6777	28.9144	1.23730	27.6771
32.85	13.2474	27.0849	1.17515	25.9097
11.40	3.7912	29.5245	1.42780	28.0967
9.60	3.2831	30.8228	1.14840	29.6744
20.90	7.2277	27.3302	1.24890	26.0813
44.70	25.7246	31.8461	1.51800	30.3281
17.10	5.1219	24.8233	0.84985	23.9734
16.50	5.0749	25.7040	1.09240	24.6116
14.35	5.0341	30.0228	1.40080	28.6220
15.45	4.8695	26.6403	1.07285	25.5674
28.15	10.6715	27.2746	1.24320	26.0314
18.35	5.3847	23.9875	0.94965	23.0379
15.10	5.6724	31.9637	1.32300	30.6407
37.75	25.8342	42.6004	1.88340	40.7170
39.05	19.6950	30.7579	1.50540	29.2525
22.25	7.2755	25.4560	0.88025	24.5757
15.50	4.4964	24.4888	0.96500	23.5238
14.10	4.3088	26.2401	1.17000	25.0701
26.65	11.3263	31.2088	1.48685	29.7219
20.25	8.0265	31.5657	1.50715	30.0586
23.55	10.1197	32.8385	1.34090	31.4976
46.65	24.7954	28.3651	1.22575	27.1394
30.55	10.0462	22.8647	1.01055	21.8541
26.80	9.5499	26.0645	1.05615	25.0083
28.10	9.4096	24.1042	0.97540	23.1288
24.55	14.5113	44.6181	2.17690	42.4412
17.85	6.6987	30.8043	1.23525	29.5690
20.90	6.5967	24.9693	0.97875	23.9905
33.00	12.3689	25.1049	0.96725	24.1377
44.00	26.1997	33.3471	1.42985	31.9172
19.00	5.0785	21.6926	0.78090	20.9117

Source: Data provided courtesy of Dr. Mark Litaker.

17. The concern of a study by Crim et al. (A-19) was the potential role of flow cytometric analysis of bronchoalveolar lavage fluid (BALF) in diagnosing acute lung rejection. The investigators note that previous studies suggested an association of acute lung rejection with increases in CD8+ lymphocytes, and increased expression of human lymphocyte antigen (HLA)-DR antigen and interleukin-2 receptor (IL-2R). Subjects consisted of lung transplant (LT) recipients who had no histologic evidence of rejection or infection, normal human volunteers (NORM), healthy heart transplant (HT) recipient volunteers, and lung transplant recipients who were experiencing acute lung rejection (AR). Among the data collected were the following percentages of BALF CD8+ lymphocytes that also express IL-2R observed in the four groups of subjects.

Norm	HT	LT	AR	
0	0	1	6	12
2	0	0	6	0
1	5	5	8	9
0	4	0	16	7
0	6	0	24	2
2	0	5	5	6
3	0	18	3	14
0	4	2	22	10
0	8	2	10	3
1	8	8	0	0
		0	8	0
		7	3	1
		2	4	1
		5	4	0
		1	18	0
			0	4

Source: Data provided courtesy of Dr. Courtney Crim.

18. Ichinose et al. (A-20) studied the involvement of endogenous tachykinins in exercise-induced airway narrowing in patients with asthma by means of a selective neurokinin 1-receptor antagonist, FK-888. Nine subjects (eight male, one female) ages 18 to 43 years with at least a 40 percent fall in the specific airway conductance participated in the study. The following are the oxygen consumption (ml/min) data for the subjects at rest and during exercise while under treatment with a placebo and FK-888:

Placebo		FK-888	
At Rest	Exercise	At Rest	Exercise
303	2578	255	2406
288	2452	348	2214
285	2768	383	3134
280	2356	328	2536
295	2112	321	1942
270	2716	234	2652
274	2614	387	2824
185	1524	198	1448
364	2538	312	2454

Source: Data provided courtesy of Dr. Kunio Shirato.

19. Transforming growth factor α (TGFα), according to Tomiya and Fujiwara (A-21), is alleged to play a role in malignant progression as well as normal cell growth in an autocrine manner, and its serum levels have been reported to increase during this progression. The present investigators have developed an enzyme-linked immunosorbent assay (ELISA) for measuring serum TGFα levels in the diagnosis of hepatocellular carcinoma (HCC) complicating cirrhosis. In a study in which they evaluated the significance of serum TGFα levels for diagnostic purposes, they collected the following

measurements on the liver function tests, TGFα (pg/ml), and serum α-fetoprotein (AFP) (ng/ml) from HCC patients:

TGFα	AFP	TGFα	AFP	TGFα	AFP	TGFα	AFP
32.0	12866	44.0	23077	100.0	479	15.0	921
65.9	9	75.0	371	12.0	47	34.0	118
25.0	124.3	36.0	291	32.0	177	100.0	6.2
30.0	9	65.0	700	98.0	9	26.0	19
22.0	610	44.0	40	20.0	1063	53.0	594
40.0	238	56.0	9538	20.0	21	140.0	10
52.0	153	34.0	19	9.0	206	24.0	292
28.0	23	300.0	11	58.0	32	20.0	11
11.0	28	39.0	42246	39.0	628	35.0	37
45.0	240	82.0	12571			52.0	35
29.0	66	85.0	20			50.0	742
45.0	83	24.0	29			95.0	10
21.0	4	40.0	310			18.0	291
38.0	214	9.0	19				

Source: Data provided courtesy of Dr. Kenji Fujiwara.

20. The objective of a study by Sakhaee et al. (A-22) was to ascertain body content of aluminum (A1) noninvasively using the increment in serum and urinary Al following the intravenous administration of deferoxamine (DFO) in patients with kidney stones and osteoporotic women undergoing long-term treatment with potassium citrate (K_3Cit) or tricalcium dicitrate (Ca_3Cit_2), respectively. Subjects consisted of 10 patients with calcium nephrolithiasis and five patients with osteoporosis who were maintained on potassium citrate or calcium citrate for 2–8 years, respectively, plus 16 normal volunteers without a history of regular aluminum-containing antacid use. Among the data collected were the following 24-hour urinary aluminum excretion measurements (μg/day) before (PRE) and after (POST) 2-hour infusion of DFO.

Group	PRE	POST	Group	PRE	POST
Control	41.04	135.00	Control	9.39	12.32
Control	70.00	95.20	Control	10.72	13.42
Control	42.60	74.00	Control	16.48	17.40
Control	15.48	42.24	Control	10.20	14.20
Control	26.90	104.30	Control	11.40	20.32
Control	16.32	66.90	Control	8.16	12.80
Control	12.80	10.68	Control	14.80	62.00
Control	68.88	46.48	Patient	15.20	27.15
Control	25.50	73.80	Patient	8.70	38.72
Patient	0.00	14.16	Patient	5.52	7.84
Patient	2.00	20.72	Patient	13.28	31.70
Patient	4.89	15.72	Patient	3.26	17.04
Patient	25.90	52.40	Patient	29.92	151.36
Patient	19.35	35.70	Patient	15.00	61.38

(*Continued*)

Group	PRE	POST	Group	PRE	POST
Patient	4.88	70.20	Patient	36.80	142.45
Patient	42.75	86.25			

Source: Data provided courtesy of Dr. Khashayar Sakhaee.

21. The purpose of a study by Dubuis et al. (A-23) was to determine whether neuropsychological deficit of children with the severe form of congenital hypothyroidism can be avoided by earlier onset of therapy and higher doses of levothyroxine. Subjects consisted of 10 infants (ages 3 to 24 days) with severe and 35 infants (ages 2 to 10 days) with moderate congenital hypothyroidism. Among the data collected were the following measurements on plasma T_4 (nmol/L) levels at screening:

Severe Cases		Moderate Cases			
Sex	T_4 (nmol/L)	Sex	T_4 (nmol/L)	Sex	T_4 (nmol/L)
M	16	F	20	F	62
M	57	F	34	M	50
M	40	F	188	F	40
F	50	F	69	F	116
F	57	F	162	F	80
F	38	F	148	F	97
F	51	F	108	F	51
F	38	F	54	F	84
M	*	F	96	F	51
F	60	M	76	F	94
		M	122	M	158
		M	43	F	*
		F	40	M	47
		F	29	M	143
		F	83	M	128
		F	62	M	112
				M	111
				F	84
				M	55

* = Missing data.
Source: Data provided courtesy of Dr. Guy Van Vliet.

22. Kuna et al. (A-24) conducted a study concerned with chemokines in seasonal allergic rhinitis. Subjects included 18 atopic individuals with seasonal allergic rhinitis caused by ragweed pollen. Among the data collected on these subjects were the following eosinophil cationic protein (ECP) and histamine measurements:

ECP (ng/ml)	Histamine (ng/ml)	ECP (ng/ml)	Histamine (ng/ml)
511.0	31.2	25.3	5.6
388.0	106.0	31.1	62.7
14.1	37.0	325.0	138.0
314.0	90.0	437.0	116.0

(*Continued*)

ECP (ng/ml)	Histamine (ng/ml)	ECP (ng/ml)	Histamine (ng/ml)
74.1	29.0	277.0	70.6
8.8	87.0	602.0	184.0
144.0	45.0	33.0	8.6
56.0	151.8	661.0	264.0
205.0	86.0	162.0	92.0

Source: Data provided courtesy of Dr. Allen P. Kaplan.

23. The purpose of a study by Kim et al. (A-25) was to investigate the serial changes in Lp(a) lipoprotein levels with the loss of female sex hormones by surgical menopause and with estrogen replacement therapy in the same women. Subjects were 44 premenopausal women who underwent a transabdominal hysterectomy (TAH). Thirty-one of the women had a TAH and unilateral salpingo-oophorectomy (USO), and 13 had a TAH and bilateral salpingo-oophorectomy (BSO). The women ranged in age from 30 to 53 years. Subjects in the BSO group received .625 mg of conjugated equine estrogen daily 2 months after the operation. The following were the subjects' total cholesterol levels before (TC0), 2 months after (TC2), and 4 months after (TC4) the surgical procedure and hormone replacement therapy.

USO			
Subject	**TC0**	**TC2**	**TC4**
1	202	203	196
2	204	183	203
3	206	199	192
4	166	180	176
5	150	171	154
6	137	134	129
7	164	168	171
8	207	249	223
9	126	121	140
10	131	141	167
11	133	159	149
12	142	152	140
13	225	193	180
14	158	182	179
15	184	177	182
16	223	244	234
17	154	178	187
18	176	137	162
19	205	253	288
20	167	156	136
21	164	176	191
22	177	168	185
23	140	175	167
24	167	186	195

USO			
Subject	**TC0**	**TC2**	**TC4**
25	134	131	135
26	163	190	185
27	196	183	192
28	181	194	208
29	160	162	181
30	188	200	181
31	172	188	189

BSO			
Subject	**TC0**	**TC2**	**TC4**
32	224	218	239
33	202	196	231
34	181	182	208
35	191	230	208
36	248	284	279
37	224	228	199
38	229	318	272
39	147	199	194
40	248	258	302
41	160	218	229
42	175	187	166
43	262	260	247
44	189	199	181

Source: Data provided courtesy of Dr. Chee Jeong Kim.

24. Velthuis et al. (A-26) conducted a study to evaluate whether the combination of passively immobilized heparin-coating and standard heparization can reduce complement activation in patients undergoing cardiac surgical intervention. The investigators note that heparin-coated extracorporeal circuits reduce complement activation during cardiac operations, but that little in vivo information is available on the reduction in alternative and classic pathway activation. Complement activation initiates a systemic inflammatory response during and after cardiac operations and is associated with pathophysiologic events such as postoperative cardiac depression, pulmonary capillary leakage, and hemolysis. Subjects were 20 patients undergoing elective cardiopulmonary bypass (CPB) grafting randomly allocated to be treated with either heparin-coated extracorporeal circuits (H) or uncoated circuits (U). Among the data collected were the following plasma terminal complement complex (SC5b-9) concentrations at baseline, 10 minutes after start of CPB, at cessation of CPB, and after the administration of protamine sulfate:

Patient	Treatment	Baseline	10 min CPB	End CPB	Protamine
1	U	0.37	0.81	1.88	2.12
2	U	0.48	0.73	3.28	3.31
3	U	0.48	0.42	2.94	1.46
4	H	0.37	0.44	1.28	3.82
5	H	0.38	0.31	0.50	0.68
6	U	0.38	0.43	1.39	5.04
7	H	0.46	0.57	1.03	1.29
8	H	0.32	0.35	0.75	1.10
9	U	0.41	0.94	1.57	2.53
10	U	0.37	0.38	2.07	1.69
11	H	0.48	0.33	1.12	1.04
12	H	0.39	0.39	1.69	1.62
13	U	0.27	0.41	1.28	2.26
14	H	0.51	0.27	1.17	1.05
15	H	0.97	0.75	1.82	1.31
16	U	0.53	1.57	4.49	2.15
17	U	0.41	0.47	1.60	1.87
18	U	0.46	0.65	1.49	1.24
19	H	0.75	0.78	1.49	1.57
20	H	0.64	0.52	2.11	2.44

Source: Data provided courtesy of Dr. Henk te Velthuis.

25. Heijdra et al. (A-27) state that many patients with severe chronic obstructive pulmonary disease (COPD) have low arterial oxygen saturation during the night. These investigators conducted a study to determine whether there is a causal relationship between respiratory muscle dysfunction and nocturnal saturation. Subjects were 20 (five females, 15 males) patients with COPD randomly assigned to receive either target-flow inspiratory muscle training (TF-IMT) at 60 percent of their maximal inspiratory mouth pressure (PI_{max}) or sham TF-IMT at 10 percent of PI_{max}. Among the data collected were the following endurance times (Time, s) for each subject at the beginning of training and 10 weeks later:

Time (s) TF-IMT 60% PI_{max}		Time (s) TF-IMT 10% PI_{max}	
Week 0	**Week 10**	**Week 0**	**Week 10**
330	544	430	476
400	590	400	320
720	624	900	650
249	330	420	330
144	369	679	486
440	789	522	369
440	459	116	110
289	529	450	474
819	1099	570	700
540	930	199	259

Source: Data provided courtesy of Dr. Yvonne F. Heijdra.

26. The three objectives of a study by Wolkin et al. (A-28) were to determine (a) the effects of chronic haloperidol treatment on cerebral metabolism in schizophrenic patients, (b) the relation between negative symptoms and haloperidol-induced regional changes in cerebral glucose utilization, and (c) the relation between metabolic change and clinical antipsychotic effect. Subjects were 18 male veterans' hospital inpatients (10 black, five white, and three Hispanic) with either acute or chronic decompensation of schizophrenia. Subjects ranged in age from 26 to 44 years, and their duration of illness ranged from 7 to 27 years. Among the data collected were the following pretreatment scores on the digit-symbol substitution subtest of the WAIS-R (DSY1RW) and haloperidol-induced change in absolute left dorsolateral prefrontal cortex (DLLA3V1) and absolute right dorsolateral prefrontal cortex (DLRA3V1) measured in units of μmol glucose/100 g tissue/min:

DSY1RW	DLLA3V1	DLRA3V1	DSY1RW	DLLA3V1	DLRA3V1
47	−7.97	−17.17	18	−4.91	−9.58
16	−8.08	−9.59	0	−1.71	.40
31	−10.15	−11.58	29	−4.62	−4.57
34	−5.46	−2.16	17	9.48	11.31
22	−17.12	−12.95	38	−6.59	−6.47
70	−12.12	−13.01	64	−12.19	−13.61
59	−9.70	−12.61	52	−15.13	−11.81
41	−9.02	−7.48	50	−10.82	−9.45
0	4.67	7.26	62	−4.92	−1.87

Source: Data provided courtesy of Dr. Adam Wolkin.

27. The purpose of a study by Maltais et al. (A-29) was to compare and correlate the increase in arterial lactic acid (La) during exercise and the oxidative capacity of the skeletal muscle in patients with chronic obstructive pulmonary disease (COPD) and control subjects (C). There were nine subjects in each group. The mean age of the patients was 62 years with a standard deviation of 5. Control subjects had a mean age of 54 years with a standard deviation of 3. Among the data collected were the

following values for the activity of phosphofructokinase (PFK), hexokinase (HK), and lactate dehydrogenase (LDH) for the two groups:

PFK		HK		LDH	
C	COPD	C	COPD	C	COPD
106.8	49.3	2.0	2.3	241.5	124.3
19.6	107.1	3.2	1.4	216.8	269.6
27.3	62.9	2.5	1.0	105.6	247.8
51.6	53.2	2.6	3.6	133.9	200.7
73.2	105.7	2.4	1.3	336.4	540.5
89.6	61.3	2.4	2.9	131.1	431.1
47.7	28.2	3.5	2.2	241.4	65.3
113.5	68.5	2.2	1.5	297.1	204.7
46.4	40.8	2.4	1.6	156.6	137.6

Source: Data provided courtesy of Dr. François Maltais.

28. Torre et al. (A-30) conducted a study to determine serum levels of nitrite in pediatric patients with human immunodeficiency virus type 1 (HIV-1) infection. Subjects included 10 healthy control children (six boys and four girls) with a mean age of 9.7 years and a standard deviation of 3.3. The remainder of the subjects were 21 children born to HIV-1-infected mothers. Of these, seven (three boys and four girls) were affected by AIDS. They had a mean age of 6 years with a standard deviation of 2.8. The remaining 14 children (seven boys and seven girls) became seronegative for HIV-1 during the first year of life. Their mean age was 3.3 years with a standard deviation of 2.3 years. Among the data collected were the following serum levels of nitrite (μmol/L):

Controls $n = 10$	Seronegativized Children $n = 14$	HIV-1-Positive Patients $n = 7$
0.301	0.335	0.503
0.167	0.986	0.268
0.201	0.846	0.335
0.234	1.006	0.946
0.268	2.234	0.846
0.268	1.006	0.268
0.201	0.803	0.268
0.234	0.301	
0.268	0.936	
0.301	0.268	
	0.134	
	0.335	
	0.167	
	0.234	

Source: Data provided courtesy of Dr. Donato Torre.

29. Seghaye et al. (A-31) analyzed the influence of low-dose aprotinin on complement activation, leukocyte stimulation, cytokine production, and the acute-phase response in children undergoing

cardiac operations. Inclusion criterion for the study was a noncyanotic congenital cardiac defect requiring a relatively simple primary surgical procedure associated with a low postoperative risk. Among the data collected were the following measurements on interleukin-6 (IL-6) and C-reactive protein (CRP) obtained 4 and 24 hours postoperatively, respectively:

IL-6	CRP	IL-6	CRP	IL-6	CRP
122	32	467	53	215	50
203	39	421	29	415	41
458	63	421	44	66	12
78	7	227	24	58	14
239	62	265	31	213	9
165	22	97	12		

Source: Data provided courtesy of Dr. Marie-Christine Seghaye.

Exercises for Use with Large Data Sets Available on the Following Website: www.wiley.com/college/daniel

1. California State Assembly Bill 2071 (AB 2071) mandated that patients at methadone clinics be required to undergo a minimum of 50 minutes of counseling per month. Evan Kletter (A-32) collected data on 168 subjects who were continuously active in treatment through the Bay Area Addiction Research and Treatment (BAART) centers for 1 year prior to, and 2 years after AB 2071's implementation. Prior to AB 2071, BAART center counselors spent two sessions of at least 15 minutes per session per month with each client. The subjects in the study were also identified as cocaine abusers. The observations in KLETTER are the percentages of failing a cocaine drug test for each of the subjects pre- and post-AB 2071. For example, a pre-value of 60 implies that the patient failed a cocaine test 60 percent of the time prior to adoption of AB 2071. Dr. Kletter performed a Wilcoxon rank sum test to determine if the percentage of failed tests decreased significantly after the passage of AB 2071. Use the data to determine what conclusion he was able to reach. Report the test statistic and p value.

REFERENCES

Methodology References

1. Frank Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics*, *1* (1945), 80–83.
2. A. M. Mood, *Introduction to the Theory of Statistics*, McGraw-Hill, New York, 1950.
3. J. Westenberg, "Significance Test for Median and Interquartile Range in Samples from Continuous Populations of Any Form," *Proceedings Koninklijke Nederlandse Akademie Van Wetenschappen*, *51* (1948), 252–261.
4. G. W. Brown and A. M. Mood, "On Median Tests for Linear Hypotheses," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 1951, 159–166.
5. H. B. Mann and D. R. Whitney, "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other," *Annals of Mathematical Statistics*, *18* (1947), 50–60.
6. A. N. Kolmogorov, "Sulla Determinazione Empirial di una Legge di Distribuizione," *Giornale dell' Institute Italiano degli Altuari*, *4* (1933), 83–91.
7. N. V. Smirnov, "Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples" (in Russian), *Bulletin Moscow University*, *2* (1939), 3–16.

8. W. H. KRUSKAL and W. A. WALLIS, "Use of Ranks in One-Criterion Analysis of Variance," *Journal of the American Statistical Association*, *47* (1952), 583–621; errata, *ibid., 48* (1953), 907–911.
9. M. FRIEDMAN, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association*, *32* (1937), 675–701.
10. M. FRIEDMAN, "A Comparison of Alternative Tests of Significance for the Problem of *m* Rankings," *Annals of Mathematical Statistics, II* (1940), 86–92.
11. C. SPEARMAN, "The Proof and Measurement of Association Between Two Things," *American Journal of Psychology*, *15* (1904), 72–101.
12. H. THEIL, "A Rank-Invariant Method of Linear and Polynomial Regression Analysis. III," *Koninklijke Nederlandse Akademie Van Wetenschappen, Proceedings, Series A*, *53* (1950), 1397–1412.
13. E. JACQUELIN DIETZ, "Teaching Regression in a Nonparametric Statistic Course," *American Statistician*, *43* (1989), 35–40.
14. JEAN D. GIBBONS, *Nonparametric Methods for Quantitative Analysis*, Third Edition, American Sciences Press, Syracuse, NY, 1996.
15. MARJORIE A. PETT, *Nonparametric Statistics for Health Care Research*, Sage Publications, Thousand Oaks, CA, 1997.

Applications References

A-1. ROBERT B. PARKER, RYAN YATES, JUDITH E. SOBERMAN, and CASEY LAIZURE, "Effects of Grapefruit Juice on Intestial P-glycoprotein: Evaluation Using Digoxin in Humans," *Pharmacotherapy*, *23* (2003), 979–987.
A-2. CHRISTINA GRAPE, MARIA SANDGREN, LARS-OLAF HANSSON, MATS ERICSON, and TÖRES THEORELL, "Does Singing Promote Well-Being?: An Empirical Study of Professional and Amateur Singers During a Singing Lesson," *Integrative Physiological and Behavioral Science*, *38* (2003), 65–74.
A-3. R. S. ZUCKERMAN and S. HENEGHAN, "The Duration of Hemodynamic Depression During Laparoscopic Cholecystectomy," *Surgical Endoscopy*, *16* (2002), 1233–1236.
A-4. CAROLE W. CRANOR and DALE B. CHRISTENSEN, "The Asheville Project: Short-Term Outcomes of a Community Pharmacy Diabetes Care Program," *Journal of the American Pharmaceutical Association*, *43* (2003), 149–159.
A-5. YITAO LIU, LUDMILA BELAYEV, WEIZHAO ZHAO, RAUL BUSTO, and MYRON D. GINSBURG, "MRZ 2(579, a Novel Uncompetitive *N*-Methyl-D-Aspartate Antagonist, Reduces Infarct Volume and Brain Swelling and Improves Neurological Deficit After Focal Cerebral Ischemia in Rats," *Brain Research*, *862* (2000), 111–119.
A-6. MOHAMED A. ELSHAIKH, JAMES C. ULCHAKER, CHANDANA A. REDDY, KENNETH ANGERMEIER, ERIC A. KLEIN, and JAY P. CIEZKI, "Prophylactic Tamsulosin (Flomax®) in Patients Undergoing Prostate ^{125}I Brachytherapy for Prostate Carcinoma. Final Report of a Double-Blind Placebo-Controlled Randomized Study," *International Journal of Radiation Oncology, Biology, Physics*, (forthcoming).
A-7. G. LACROIX, S. TISSOT, F. ROGERIEUX, R. BEAULIEU, L. CORNU, C. GILLET, F. ROBIDEL, J. P. LEFÈVRE, and F. Y. BOIS, "Decrease in Ovalbumin-Induced Pulmonary Allergic Response by Benzaldehyde but Not Acetaldehyde Exposure in a Guinea Pig Model," *Journal of Toxicology and Environmental Health, Part A*, *65* (2002), 995–1012.
A-8. W. HERRMANN, H. SCHORR, J. P. KNAPP, A. MÜLLER, G. STEIN, and J. GEISEL, "Role of Homocysteine, Cystathionine and Methylmalonic Acid Measurement for Diagnosis of Vitamin Deficiency in High-Aged Subjects," *European Journal of Clinical Investigation*, *30* (2000), 1083–1089.
A-9. CAROL FLEXER, KATE KEMP BILEY, ALYSSA HINKLEY, CHERYL HARKEMA, and JOHN. HOLCOMB, "Using Sound-Field to Teach Phonemic Awareness to Pre-schoolers," *Hearing Journal*, *55* (3) (2002), 38–44.
A-10. SATOSHI NOZAWA, KATSUJI SHIMIZU, KEI MIYAMOTO, and MIZUO TANAKA, "Repair of Pars Interarticularis Defect by Segmental Wire Fixation in Young Athletes with Spondylolysis," *American Journal of Sports Medicine*, *31* (2003), 359–364.
A-11. M. BUTZ, K. H. WOLLINSKY, U. WIDEMUTH-CATRINESCU, A. SPERFELD, S. WINTER, H. H. MEHRKENS, A. C. LUDOLPH, and H. SCHREIBER, "Longitudinal Effects of Noninvasive Positive-Pressure Ventilation in Patients with Amyotrophic Lateral Sclerosis," *American Journal of Medical Rehabilitation*, *82* (2003), 597–604.
A-12. TOMOHIRO OTANI and SHOJI KISHI, "A Controlled Study of Vitrectomy for Diabetic Macular Edema," *American Journal of Ophthalmology*, *134* (2002), 214–219.
A-13. J. JOSE and S. R. ELL, "The Association of Subjective Nasal Patency with Peak Inspiratory Nasal Flow in a Large Healthy Population," *Clinical Otolaryngology*, *28* (2003), 352–354.

A-14. Wendy Gantt and the Wright State University Statistical Consulting Center (2002).
A-15. Mary White and the Wright State University Statistical Consulting Center (2001).
A-16. Dominic Sprott and the Wright State University Statistical Consulting Center (2003).
A-17. Peter Damm, Henrik Vestergaard, Claus Kühl, and Oluf Pedersen, "Impaired Insulin-Stimulated Non-oxidative Glucose Metabolism in Glucose-Tolerant Women with Previous Gestational Diabetes," *American Journal of Obstetrics and Gynecology*, *174* (1996), 722–729.
A-18. Bernard Gutin, Mark Litaker, Syed Islam, Tina Manos, Clayton Smith, and Frank Treiber, "Body-Composition Measurement in 9-11-yr-Old Children by Dual-Energy X-Ray Absorptiometry, Skinfold-Thickness Measurements, and Bioimpedance Analysis," *American Journal of Clinical Nutrition*, *63* (1996), 287–292.
A-19. Courtney Crim, Cesar A. Keller, Cherie H. Dunphy, Horacio M. Maluf, and Jill A. Ohar, "Flow Cytometric Analysis of Lung Lymphocytes in Lung Transplant Recipients," *American Journal of Respiratory and Critical Care Medicine*, *153* (1996), 1041–1046.
A-20. Masakazu Ichinose, Motohiko Miura, Hideyuki Yamauchi, Natsuko Kageyama, Masafumi Tomaki, Tatsuya Oyake, Yuzuru Ohuchi, Wataru Hida, Hiroshi Miki, Gen Tamura, and Kunio Shirato, "A Neurokinin 1-Receptor Antagonist Improves Exercise-Induced Airway Narrowing in Asthmatic Patients," *American Journal of Respiratory and Critical Care Medicine*, *153* (1996), 936–941.
A-21. Tomoaki Tomiya and Kenji Fujiwara, "Serum Transforming Growth Factor *a* Level as a Marker of Hepatocellular Carcinoma Complicating Cirrhosis," *Cancer*, *77* (1996), 1056–1060.
A-22. Khashayar Sakhaee, Lisa Ruml, Paulette Padalino, Sharon Haynes, and Charles Y. C. Pak, "The Lack of Influence of Long-Term Potassium Citrate and Calcium Citrate Treatment in Total Body Aluminum Burden in Patients with Functioning Kidneys," *Journal of the American College of Nutrition*, *15* (1996), 102–106.
A-23. Jean-Michel Dubuis, Jacqueline Glorieux, Faisca Richer, Cheri L. Deal, Jean H. Dussault, and GuyVan Vliet, "Outcome of Severe Congenital Hypothyroidism: Closing the Developmental Gap with Early High Dose Levothyroxine Treatment," *Journal of Clinical Endocrinology and Metabolism*, *81* (1996), 222–227.
A-24. Piotr Kuna, Mark Lazarovich, and Allen P. Kaplan, "Chemokines in Seasonal Allergic Rhinitis," *Journal of Allergy and Clinical Immunology*, *97* (1996), 104–112.
A-25. Chee Jeong Kim, Wang Seong Ryu, Ju Won Kwak, Chong Taik Park, and Un Ho Ryoo, "Changes in Lp(a) Lipoprotein and Lipid Levels After Cessation of Female Sex Hormone Production and Estrogen Replacement Therapy," *Archives of Internal Medicine*, *156* (1996), 500–504.
A-26. Henk te Velthuis, Piet G. M. Jansen, C. Erik Hack, León Eijsman, and Charles R. H. Wildevuur, "Specific Complement Inhibition with Heparin-Coated Extracorporeal Circuits," *Annals of Thoracic Surgery*, *61* (1996), 1153–1157.
A-27. Yvonne F. Heijdra, P. N. Richard Dekhuijzen, Cees L. A.van Herwaarden, and Hans Th. M. Folgering, "Nocturnal Saturation Improves by Target-Flow Inspiratory Muscle Training in Patients with COPD," *American Journal of Respiratory and Critical Care Medicine*, *153* (1996), 260–265.
A-28. Adam Wolkin, Michael Sanfilipo, Erica Duncan, Burton Angrist, Alfred P. Wolf, Thomas B. Cooper, Jonathan D. Brodie, Eugene Laska, and John P. Rostrosen, "Blunted Change in Cerebral Glucose Utilization After Haloperidol Treatment in Schizophrenic Patients with Prominent Negative Symptoms," *American Journal of Psychiatry*, *153* (1996), 346–354.
A-29. François Maltais, Andrée-Anne Simard, Clermont Simard, Jean Jobin, Pierre Desgagnés, and Pierre LeBlanc, "Oxidative Capacity of the Skeletal Muscle and Lactic Acid Kinetics During Exercise in Normal Subjects and in Patients with COPD," *American Journal of Respiratory and Critical Care Medicine*, *153* (1996), 288–293.
A-30. Donato Torre, Giulio Ferrario, Filippo Speranza, Roberto Martegani, and Claudia Zeroli, "Increased Levels of Nitrite in the Sera of Children Infected with Human Immunodeficiency Virus Type 1," *Clinical Infectious Diseases*, *22* (1996), 650–653.
A-31. Marie-Christine Seghaye, Jean Duchateau, Ralph G. Grabitz, Karsten Jablonka, Tobias Wenzl, Christiane Marcus, Bruno J. Messmer, and Goetz von Bernuth, "Influence of Low-Dose Aprotinin on the Inflammatory Reaction Due to Cardiopulmonary Bypass in Children," *Annals of Thoracic Surgery*, *61* (1996), 1205–1211.
A-32. Evan Kletter, "Counseling as an Intervention for the Cocaine-Abusing Methadone Maintenance Patient," *Journal of Psychoactive Drugs*, *35* (2003), 271–277.

CHAPTER 14

SURVIVAL ANALYSIS

CHAPTER OVERVIEW

This chapter provides an introduction to the analysis of data arising from studies where the time to the occurrence of an event is the outcome of interest. These types of studies have historically been used to monitor the survival time of patients who face the possibility of dying during the study, hence the use of the description of these techniques as "survival analysis." However, in this chapter we will learn techniques that can be used in the context of any outcome where the time to occurrence of an event is of interest. We will be employing techniques similar to those we have learned in previous chapters, including the methods for analyzing frequency data, the methods for developing linear models for making predictions, and topics in nonparametric statistics.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand time-to-event data and how censored observations can be handled statistically.
2. be able to develop and use survival curves to make conclusions.
3. be able to statistically compare survival curves.
4. understand how to develop models designed to handle time-to-event data.

14.1 INTRODUCTION

In many studies, the outcome of interest is related to the timing of the occurrence of an event. In a clinical setting, one may be interested in measuring how long a chronically ill

patient survives after receiving a certain treatment. In another scenario, one may be interested in determining which of three drugs, compared to a placebo, provides symptom relief most rapidly.

Imagine that a cardiac rehabilitation clinic is interested in determining if enrollment in a traditional health education program or enrollment in a program that provides diet and nutritional planning along with patient education is more effective at preventing the occurrence of a second myocardial infarction following a first heart attack. The study could begin when the first patient, following his or her first heart attack, is randomly assigned to a treatment program, with additional patients enrolled through time. Conversely, the study could begin with a cohort of subjects, each of whom has had their first heart attack, who are randomly assigned to a treatment program. In either case, there are potentially three outcomes that could occur with each patient, with the *event of interest* being a second heart attack. These are (*1*) the patient has a second heart attack; (*2*) the patient drops out of the study—thereby becoming a *loss to follow-up*—which could occur for any number of reasons, including death, or relocating geographically, for example; or (*3*) the event of interest does not occur to the patient during the period of study. These three mutually exclusive events are the foundation for survival analysis studies.

Though the vast majority of published research using the methods of survival analysis is clinical in nature, it should be mentioned that there are many nonclinical uses for survival analysis as well. With the advent of computer-based statistical programs to help with complex calculations, the use of survival analysis methodologies has increased demonstrably among many disciplines. For example, engineers may wish to know the time it takes for a battery to lose its charge, a quality-control scientist at a manufacturing plant may wish to understand at what point machines need to be recalibrated, or an ecologist may want to estimate how long the average carcass remains in a study area before it is scavenged.

14.2 TIME-TO-EVENT DATA AND CENSORING

Measurement data for survival analysis studies utilizes the time that it takes for a well-defined event of interest to occur. For each subject enrolled in a study, the researcher records the amount of time (this could be months, days, years, or any measure of time) elapsing between the point at which each subject entered into the study until he or she experiences one of the three possible events just presented—the event occurs, the event does not occur, or the subject is lost to follow-up. The total amount of time between the initial enrollment in the study and the occurrence of one of the three outcomes is known as the research subject's *survival time*, or *time-to-event*. Hence, the information gathered on each subject is often referred to as *survival data* or *time-to-event data*. In addition to the survival data, covariates, such as age, gender, medication type, and diet, for example can also be gathered for the development of complex models.

DEFINITION

***Survival data,* or *time-to-event data*, are measurements of elapsed time between the initial enrollment in a study and the final disposition of the study subject. This elapsed time could be represented by the time of initial diagnosis or it could be represented by the point in time when one**

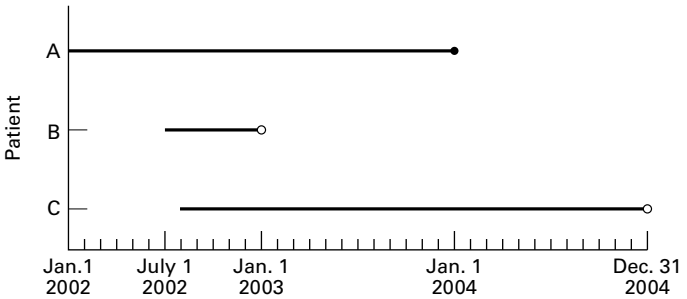


FIGURE 14.2.1 Patients entering a study at different times with known (●) and censored (○) survival times.

enters the study. *Survival* in this context simply means that an event has not occurred, not, necessarily, that the endpoint of interest involved an examination of "life" and "death."

Suppose we consider patients who entered into the heart-attack study described in the Introduction. For illustrative purposes, suppose we examine the fate of three patients who were in the study (Figure 14.2.1).

Patient A entered the study on January 1, 2002 and had a myocardial infarction on December 31, 2003. Patient A's survival time is therefore 24 months. Patient B entered the study on July 1, 2002 and moved out of state 6 months later on December 31, 2002. Patient B's survival time in the study is 6 months. Finally, Patient C entered the study on August 1, 2002 and remained in the study until it ended on December 31, 2004. Patient C's survival time is 29 months. We, therefore, have survivorship information on these three patients that might be useful for analysis; however, we notice that the survival times for Patients B and C are not known exactly. That is, Patient B provides an example of a patient lost to follow-up, and patient C provides an example of a patient that completed the study without experiencing the event of interest. Patients B and C have survival times that are called *censored survival times* and hence these survival times are referred to as *censored data*.

DEFINITION

***Censored data* are represented by measurements for which we have some information about survival time, but the exact survival time is not known.**

Censored data can occur in a number of ways. In *singly censored data*, a fixed number of subjects enter into a study at the same time. Once in the study, some of the subjects will not experience the event. Their survival time is known to be some length of time greater than the length of the study. This is known as *type I censoring*. It could also be that for research or ethical reasons the study is ended after a certain proportion of the subjects experience the condition of interest, with the remaining proportion having not experienced the event when the study is ended. This is called *type II censoring*. It should be noted that these concepts are not related to the concepts of Type I error and Type II error introduced in Chapter 7. Another type of censoring

that may occur is known as *progressively censored data* in which the period of study is fixed, but subjects may enter the experiment at different times. Patients may then either experience or not experience the event of interest, with those not experiencing the event having unknown survival times. This is called *type III censoring*. Data for which exact endpoints are not known, either because the subject dropped out of the study, was withdrawn from the study, or survived beyond the termination of the study are called *right-censored data* because the survival times extend beyond the right tail of the distribution of survival times. Conversely, we could have data for which exact beginning points are not known. This could arise, for example, if a subject with the condition enters the study, but it is not known exactly when the condition developed in the patient. These data are known as *left-censored data* because their survival times are truncated on the left side of the distribution of the survival time distribution, causing the difference in time between diagnosis and entering into the study to be unknown. Clearly, details surrounding censored data are complex and require much more detailed analysis than is covered in this introductory text. For those interested in further reading, we suggest the books by Kleinbaum and Klein (1), Lee (2), and Hosmer and Lemeshow (3).

Generally, for purposes of analysis, a dichotomous, or indicator, variable is used to distinguish survival times of those subjects who experience the event of interest and those that do not because of one of the censoring mechanisms described above. Typically this variable is called a *status variable*, with a zero indicating that an event did not occur and hence the survival time is censored, and a 1 indicating that the event of interest did occur.

In studies where different treatments are being investigated, we are interested in three items of information for each subject: (*1*) Which treatment was given to the patient? (*2*) For what length of time was the patient observed? (*3*) Did the patient experience the event of interest during the study or was the survival time censored for some reason? In studies that are not concerned with comparing different treatment conditions, only the last two items of data are relevant. Additionally, we may be interested in different covariates associated with patients (e.g., age, gender, income level) in order to develop more complex models, and therefore we may develop questions based on these covariates of interest.

With these three items of information in hand, along with any covariates of interest, we are able, in studies such as the myocardial infarction example mentioned in Section 14.1, to estimate the median survival time of the group of patients who received one treatment compared to another. Comparison of different treatment medians allows us to answer the following question: Based on the information from our study, which treatment do we conclude delays for a longer period of time, on the average, the occurrence of a second heart attack? The data collected in follow-up studies such as we have described may also be used to answer another question of considerable interest to the clinician: What is the estimated probability that a patient will survive for a specified length of time? Or, Is there a difference in survivorship of males and females who have experienced heart attacks? For the myocardial infarction study, the clinician may ask: "What is the probability that a patient who received treatment A will survive more than 2 years?" The methods employed to answer these types of questions are known as *survival analysis* methods.

Statistical Distribution Functions Before presenting survival analysis methods, it is important to consider data distributions commonly encountered in such analyses. Time-to-event data are distributed temporally, such that events occur either at some point, or within some interval, of time. These events are considered to represent a

random variable having some probability of occurrence at each time period for each subject in the study.

We have already encountered two useful representations of probability distributions in Chapter 4. These were the cumulative distribution function and the probability distribution function. If we let the event time be represented by T, then the cumulative distribution function of T is represented by $F(t)$, such that

$$F(t) = P(T \leq t) \tag{14.2.1}$$

That is, the cumulative distribution function represents the probability that an event time is less than or equal to some specified measurement time, t. As you recall from Chapter 4, $F(t)$ is an increasing function that runs from a value of zero (it is assumed theoretically that no events have occurred at the initiation of the study), to a value of 1 (it is assumed theoretically that all events have occurred at the conclusion of the study). In the context of survival analysis, a closely related function that is more commonly used than $F(t)$ is a function that runs from a value of 1 (it is assumed that all subjects at the initiation of the study have "survived" to that point) to a value of zero (it is assumed theoretically that none of the subjects have "survived" when the study ends, though some subjects may be censored). Conveniently, this is known as the survival distribution, $S(t)$, and is mathematically related to the cumulative distribution function by

$$S(t) = 1 - F(t) \tag{14.2.2}$$

Both of these distributions are illustrated in Figure 14.2.2. It is the survival curve we generally are most interested in, and comparisons of various survival curves provide a statistical means to compare such things as individual survival and differences in survival among different treatments.

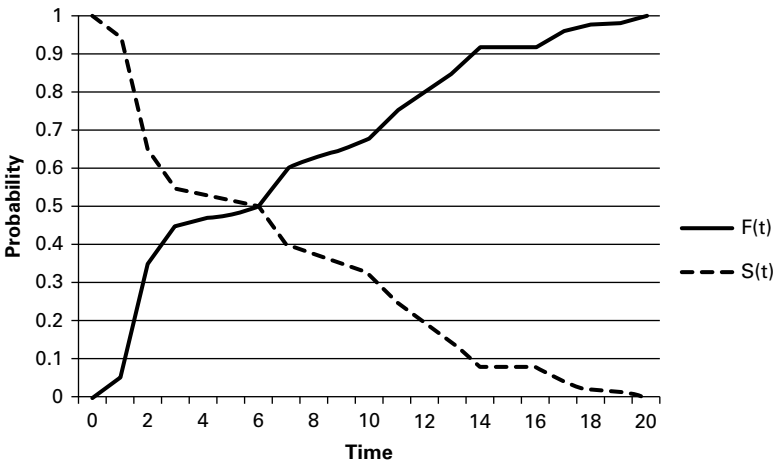


FIGURE 14.2.2 Illustration of the cumulative distribution function, F(t), and the survival distribution, S(t).

The probability distribution function, just as defined in Chapter 4, is represented by the set of probabilities that specify the possible values of a random variable. In the context of survival analysis, this density function represents the probability of an event occurring in a defined interval of time. We might ask, for example, what is the probability of surviving 2 months? Although fully appreciating the intricacies of this probability distribution requires knowledge of calculus, we can illustrate its meaning conceptually by remembering a concept from our discussion of the normal distribution in Chapter 4. When we calculated probabilities for the normal distribution, we were interested in calculating the area under a curve that was bounded by two values. Similarly, in survival analysis we are interested in calculating the probability of an event bounded by an interval of time, say Δt, and then finding our probability as the interval becomes very small, that is as $\Delta t \to 0$. Hence, the probability distribution function, $f(t)$, is defined by

$$f(t) = \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \quad \text{as } \Delta t \to 0 \tag{14.2.3}$$

That is, the set of probabilities of events that occur in an infinitesimally small interval of time defines the probability function. It is also possible to find this function by examining what happens during a change in $F(t)$, say $\Delta F(t)$, or a change in $S(t)$, say $\Delta S(t)$, in a given interval of time. That is

$$f(t) = \frac{\Delta F(t)}{\Delta t} = -\frac{\Delta S(t)}{\Delta t} \tag{14.2.4}$$

Finally, a function that is often encountered in survival analysis is the hazard function, $h(t)$. This function is used to define the instantaneous probability of an event occurring given that the subject has survived up to a given time, t. This function is defined as

$$h(t) = \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \quad \text{as } \Delta t \to 0 \tag{14.2.5}$$

Note that this function is based on a conditional probability, wherein we are interested in calculating the probability of an event occurring given that the subject has already survived to a defined time. The condition of having already survived to a given time means that the probability of surviving into the future is influenced by having already survived previous time periods. This idea can be very important in some instances, where surviving the early stages of a disease may dramatically decrease the potential of an event occurring in the near future. As an example, consider cancer where nonrecurrence, or remission, for a period of 5 years generally increases survivorship. This function can also be expressed in terms of two functions previously defined. This expression is

$$h(t) = \frac{f(t)}{S(t)} \tag{14.2.6}$$

Because the hazard function can exceed 1, it is not truly a probability, though it is based on the conditional probability of an event occurring. The hazard function is often defined in survival analysis by a known distribution such as the lognormal, exponential, or Weibull

distribution. Excellent descriptions of the various models used to represent hazard functions are provided by Allison (4) and Kleinbaum and Klein (1).

14.3 THE KAPLAN–MEIER PROCEDURE

Now let us show how we may use the data usually collected in follow-up studies of the type we have been discussing to estimate the probability of surviving for a specified length of time. The method we use was introduced by Kaplan and Meier (5) and for that reason is called the *Kaplan–Meier procedure*. Since the procedure involves the successive multiplication of individual estimated probabilities, it is sometimes referred to as the *product-limit method* of estimating survival probabilities.

As we shall see, the calculations include the computations of proportions of subjects in a sample who survive for various lengths of time. We use these sample proportions as estimates of the probabilities of survival that we would expect to observe in the population represented by our sample. In mathematical terms we refer to the process as the estimation of a survivorship function. Frequency distributions and probability distributions may be constructed from observed survival times, and these observed distributions may show evidence of following some theoretical distribution of known functional form. When the form of the sampled distribution is unknown, it is recommended that the estimation of a survivorship function be accomplished by means of a *nonparametric technique,* of which the Kaplan–Meier procedure is one. Nonparametric techniques are defined and discussed in detail in Chapter 13.

Calculations for the Kaplan–Meier Procedure

We let

n = the number of subjects whose survival times are available

p_1 = the proportion of subjects surviving at least the first time period (day, month, year, etc.)

p_2 = the proportion of subjects surviving the second time period after having survived the first time period

p_3 = the proportion of subjects surviving the third time period after having survived the second time period

$\vdots$

p_k = the proportion of subjects surviving the kth time period after having survived the $(k-1)$th time period

We use these proportions, which we may relabel $\hat{p}_1, \hat{p}_2, \hat{p}_3, \ldots, \hat{p}_k$ as estimates of the probability that a subject from the population represented by the sample will survive time periods 1, 2, 3, . . . , k, respectively.

For any time period, t, where $1 \leq t \leq k$, we estimate the probability of surviving the tth time period, p_t, as follows:

$$\hat{p}_t = \frac{\text{number of subjects surviving at least } (t-1) \text{ time periods who also survive the } t\text{th period}}{\text{number of subjects alive at end of time period } (t-1)} \tag{14.3.1}$$

The probability of surviving to time t, $S(t)$, is estimated by

$$\hat{S}(t) = \hat{p}_1 \times \hat{p}_2 \times \cdots \times \hat{p}_t \tag{14.3.2}$$

We illustrate the use of the Kaplan–Meier procedure with the following example.

EXAMPLE 14.3.1

To assess results and identify predictors of survival, Martini et al. (A-1) reviewed their total experience with primary malignant tumors of the sternum. They classified patients as having either low-grade (25 patients) or high-grade (14 patients) tumors. The event (status), time to event (months), and tumor grade for each patient are shown in Table 14.3.1. We wish to compare the 5-year survival experience of these two groups by means of the Kaplan–Meier procedure.

Solution: The data arrangement and necessary calculations are shown in Table 14.3.2. The entries for the table are obtained as follows.

TABLE 14.3.1 Survival Data, Subjects with Malignant Tumors of the Sternum

Subject	Time (Months)	Vital Status[a]	Tumor Grade[b]	Subject	Time (Months)	Vital Status[a]	Tumor Grade[b]
1	29	dod	L	21	155	ned	L
2	129	ned	L	22	102	dod	L
3	79	dod	L	23	34	ned	L
4	138	ned	L	24	109	ned	L
5	21	dod	L	25	15	dod	L
6	95	ned	L	26	122	ned	H
7	137	ned	L	27	27	dod	H
8	6	ned	L	28	6	dod	H
9	212	dod	L	29	7	dod	H
10	11	dod	L	30	2	dod	H
11	15	dod	L	31	9	dod	H
12	337	ned	L	32	17	dod	H
13	82	ned	L	33	16	dod	H
14	33	dod	L	34	23	dod	H
15	75	ned	L	35	9	dod	H
16	109	ned	L	36	12	dod	H
17	26	ned	L	37	4	dod	H
18	117	ned	L	38	0	dpo	H
19	8	ned	L	39	3	dod	H
20	127	ned	L				

[a]dod = dead of disease; ned = no evidence of disease; dpo = dead postoperation.
[b]L = low-grade; H = high-grade.
Source: Data provided courtesy of Dr. Nael Martini.

TABLE 14.3.2 Data Arrangement and Calculations for Kaplan–Meier Procedure, Example 14.3.1

1	2	3	4	5	6
Time (Months)	Vital Status 0 = Censored 1 = Dead	Patients at Risk	Patients Remaining Alive	Survival Proportion	Cumulative Survival Proportion
Patients with Low-Grade Tumors					
6	0				
8	0				
11	1	23	22	$22/23 = .956522$	.956522
15	1				
15	1	22	20	$20/22 = .909090$	.869564
21	1	20	19	$19/20 = .950000$	.826086
26	0				
29	1	18	17	$17/18 = .944444$	.780192
33	1	17	16	$16/17 = .941176$	.734298
34	0				
75	0				
79	1	14	13	$13/14 = .928571$	.681847
82	0				
95	0				
102	1	11	10	$10/11 = .909090$	.619860
109	0				
109	0				
117	0				
127	0				
129	0				
137	0				
138	0				
155	0				
212	1	2	1	$1/2 = .500000$	.309930
337	0				

(*Continued*)

TABLE 14.3.2 (*Continued*)

1	2	3	4	5	6
Time (Months)	Vital Status 0 = Censored 1 = Dead	Patients at Risk	Patients Remaining Alive	Survival Proportion	Cumulative Survival Proportion
			Patients with High-Grade Tumors		
0	1	14	13	$13/14 = .928571$	.928571
2	1	13	12	$12/13 = .923077$	.857142
3	1	12	11	$11/12 = .916667$	.785714
4	1	11	10	$10/11 = .909090$	.714285
6	1	10	9	$9/10 = .900000$	.642856
7	1	9	8	$8/9 = .888889$	.571428
9	1				
9	1	8	6	$6/8 = .750000$	.428572
12	1	6	5	$5/6 = .833333$	.357143
16	1	5	4	$4/5 = .800000$	.285714
17	1	4	3	$3/4 = .750000$	.214286
23	1	3	2	$2/3 = .666667$	.142857
27	1	2	1	$1/2 = .500000$	.071428
122	0	1	0		

1. We begin by listing the observed times in order from smallest to largest in Column 1.
2. Column 2 contains an indicator variable that shows vital status $(1 = \text{died}, 0 = \text{alive or censored})$.
3. In Column 3 we list the number of patients at risk for each time associated with the death of a patient. We need only be concerned about the times at which deaths occur because the survival rate does not change at censored times.
4. Column 4 contains the number of patients remaining alive just after one or more deaths.
5. Column 5 contains the estimated conditional probability of surviving, which is obtained by dividing Column 4 by Column 3. Note that although there were two deaths at 15 months in the low-grade group and two deaths at 9 months in the high-grade group, we calculate only one survival proportion at these points. The calculations take the two deaths into account.
6. Column 6 contains the estimated cumulative probability of survival. We obtain the entries in this column by successive multiplication. Each entry after the first in Column 5 is multiplied by the cumulative product of all previous entries.

After the calculations are completed we examine Table 14.3.2 to determine what useful information it provides. From the table we note the following facts, which allow us to compare the survival experience of the two groups of subjects: those with low-grade tumors and those with high-grade tumors:

1. **Median survival time.** We can determine the median survival time by locating the time, in months, at which the cumulative survival proportion is equal to .5. None of the cumulative survival proportions are exactly .5, but we see that in the low-grade tumor group, the probability changes from .619860 to .309930 at 212 months; therefore, the median survival for this group is 212 months. In the high-grade tumor group, the cumulative proportion changes from .571428 to .428572 at 9 months, which is the median survival for this group.
2. **Five-year survival rate.** We can determine the 5-year or 60-month survival rate for each group directly from the cumulative survival proportion at 60 months. For the low-grade tumor group, the 5-year survival rate is .734298 or 73 percent; for the high-grade tumor group, the 5-year survival rate is .071428 or 7 percent.
3. **Mean survival time.** We may compute for each group the mean of the survival times, which we will call $\overline{T}_L$ and $\overline{T}_H$ for the low-grade and high-grade groups, respectively. For the low-grade tumor group we compute $\overline{T}_L = 2201/25 = 88.04$, and for the high-grade tumor group we compute $\overline{T}_H = 257/14 = 18.35$. Since so many of the times in the low-grade group are censored, the true mean survival time for that group is, in reality, higher (perhaps, considerably so) than 88.04. The true mean survival time for the high-grade group is also likely higher than the computed 18.35, but with just one censored time we do not expect as great a difference between the calculated mean and the true mean. Thus, we see that we have still another indication that the survival experience of the low-grade tumor group is more favorable than the survival experience of the high-grade tumor group.
4. **Average hazard rate.** From the raw data of each group we may also calculate another descriptive statistic that can be used to compare the two survival experiences. This statistic is called the *average hazard rate*. It is a measure of nonsurvival potential rather than survival. A group with a higher average hazard rate will have a lower probability of surviving than a group with a lower average hazard rate. We compute the average hazard rate, designated $\overline{h}$ by dividing the number of subjects who do not survive by the sum of the observed survival times. For the low-grade tumor group, we compute $\overline{h}_L = 9/2201 = .004089$. For the high-grade tumor group we compute $\overline{h}_H = 13/257 = .05084$, We see that the average hazard rate for the high-grade group is higher than for the low-grade group, indicating a smaller chance of surviving for the high-grade group.

The cumulative survival proportion column of Table 14.3.2 may be portrayed visually in a survival curve graph in which the cumulative survival proportions are represented by the vertical axis and the time in months by the horizontal axis. We note that the graph resembles stairsteps with "steps" occurring at the times when deaths occurred. The graph also allows us

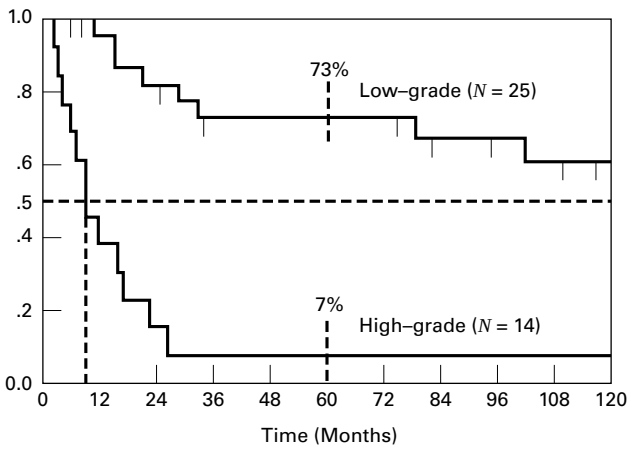


FIGURE 14.3.1 Kaplan–Meier survival curve, Example 14.3.1, showing median survival times and 5-year (60-month) survival rates.

to represent visually the median survival time and survival rates such as the 5-year survival rate. The graph for the cumulative survival data of Table 14.3.2 is shown in Figure 14.3.1.

These observations strongly suggest that the survival experience of patients with low-grade tumors is far more favorable than that of patients with high-grade tumors. ■

EXERCISES

14.3.1 Fifty-three patients with medullary thyroid cancer (MTC) were the subjects of a study by Dottorini et al. (A-2), who evaluated the impact of different clinical and pathological factors and the type of treatment on their survival. Thirty-two of the patients were females, and the mean age of all patients was 46.11 years with a standard deviation of 14.04 (range 18–35 years). The following table shows the status of each patient at various periods of time following surgery. Calculate the survival function using the Kaplan–meier procedure and plot the survival curve.

Subject	Time[a] (Years)	Status[b]	Subject	Time[a] (Years)	Status[b]
1	0	doc	28	6	alive
2	1	mtc	29	6	alive
3	1	mtc	30	6	alive
4	1	mtc	31	6	alive
5	1	mtc	32	7	mtc
6	1	mtc	33	8	alive
7	1	mtc	34	8	alive
8	1	mtc	35	8	alive
9	1	alive	36	8	alive
10	2	mtc	37	8	alive

(*Continued*)

Subject	Time[a] (Years)	Status[b]	Subject	Time[a] (Years)	Status[b]
11	2	mtc	38	9	alive
12	2	mtc	39	10	alive
13	2	alive	40	11	mtc
14	2	alive	41	11	doc
15	3	mtc	42	12	mtc
16	3	mtc	43	12	doc
17	3	alive	44	13	mtc
18	4	mtc	45	14	alive
19	4	alive	46	15	alive
20	4	alive	47	16	mtc
21	4	alive	48	16	alive
22	5	alive	49	16	alive
23	5	alive	50	16	alive
24	5	alive	51	17	doc
25	5	alive	52	18	mtc
26	6	alive	53	19	alive
27	6	alive			

[a] Time is number of years after surgery.
[b] doc = dead of other causes; mtc = dead of medullary thyroid cancer.
source: Data provided courtesy of Dr. Massimo E. Dottorini.

14.3.2 Banerji et al. (A-3) followed non–insulin-dependent diabetes mellitus (NIDDM) patients from onset of their original hyperglycemia and the inception of their near–normoglycemic remission following treatment. Subjects were black men and women with a mean age of 45.4 years and a standard deviation of 10.4. The following table shows the relapse/remission experience of 62 subjects. Calculate the survival function using the Kaplan–Meier procedure and plot the survival curve.

Total Duration of Remission (Months)	Remission Status[a]	Total Duration of Remission (Months)	Remission Status[a]	Total Duration of Remission (Months)	Remission Status[a]
3	1	8	2	26	1
3	2	9	2	27	1
3	1	10	1	28	2
3	1	10	1	29	1
3	1	11	2	31	2
4	1	13	1	31	1
4	1	16	1	33	2
4	1	16	2	39	2
5	1	17	2	41	1
5	1	18	2	44	1
5	1	20	1	46	1
5	1	22	1	46	2
5	1	22	2	48	1
5	1	22	2	48	2

(*Continued*)

Total Duration of Remission (Months)	Remission Status[a]	Total Duration of Remission (Months)	Remission Status[a]	Total Duration of Remission (Months)	Remission Status[a]
5	1	23	1	48	1
6	1	24	2	49	1
6	1	25	2	50	1
6	1	25	2	53	1
7	1	26	1	70	2
8	2	26	1	94	1
8	1				
8	2				

[a] 1 = yes (the patient is still in remission); 2 = no (the patient has relapsed).
Source: Data provided Courtesy of Dr. Mary Ann Banerji.

14.4 COMPARING SURVIVAL CURVES

Examination of a survival curve for a single group of individuals is valuable in that it allows one to see characteristics that are not as easily seen by examining a set of tabulated values. This includes visualizing the temporal trajectory to find time periods in which there were dramatic changes in survival, finding time periods in which relatively little change occurred, or in finding the approximate median of the data distribution. The construction of survival curves, however, finds its greatest use when comparisons among survival distributions are of interest. For example, one may wish to examine differences in treatment in which subjects were randomly assigned, or may wish to know which medication delays the onset of the event of interest for the longest period of time.

The results of comparing the survival experiences of different groups will not always be as dramatic as those of our previous example. For an objective comparison of the survival experiences of different groups, it is desirable that we have an objective technique for determining whether they are statistically significantly different. We know also that the observed results apply strictly to the samples on which the analyses are based. Of much greater interest is a method for determining if we may conclude that there is a difference between survival experiences in the populations from which the samples were drawn. In other words, at this point, we desire a method for testing the null hypothesis that there is no difference in survival experience between populations against the alternative that there is a difference. Such a test is provided by the *log-rank test*. The log-rank test is an application of the Mantel–Haenszel procedure discussed in Section 12.7. The extension of the procedure to survival data was proposed by Mantel (6). Though we may wish to compare survival curves of many populations, we will limit our discussion to the comparison of two groups: To accomplish this task, we calculate the log-rank statistic and proceed as follows:

1. Order the survival times until death for both groups combined, omitting censored times. Each time constitutes a stratum as defined in Section 12.7.
2. For each stratum or time, t_i, we construct a 2×2 table in which the first row contains the number of observed deaths, the second row contains the number of

TABLE 14.4.1 Contingency Table for Stratum (Time) t_i for Calculating the Log-Rank Test

	Group A	Group B	Total
Number of deaths observed	a_i	b_i	$a_i + b_i$
Number of patients alive	c_i	d_i	$c_i + d_i$
Number of patients "at risk"	$a_i + c_i$	$b_i + d_i$	$n_i = a_i + b_i + c_i + d_i$

patients alive, the first column contains data for one group, say, group A, and the second column contains data for the other group, say, group B. Table 14.4.1 shows the table for time t_i.

3. For each stratum compute the expected frequency for the upper left-hand cell of its table by Equation 12.7.5.
4. For each stratum compute v_i by Equation 12.7.6.
5. Finally, compute the Mantel–Haenszel statistic (now called the log-rank statistic) by Equation 12.7.7.

We illustrate the calculation of the log-rank statistic with the following example.

EXAMPLE 14.4.1

Let us refer again to the data on primary malignant tumors of the sternum presented in Example 14.3.1. Examination of the data reveals that there are 20 time periods (strata). For each of these a 2×2 table following the pattern of Table 14.4.1 must be constructed. The first of these tables is shown as Table 14.4.2. By Equations 12.7.5 and 12.7.6 we compute e_i and v_i as follows:

$$e_i = \frac{(0+1)(0+25)}{39} = .641$$

$$v_i = \frac{(0+1)(25+13)(0+25)(1+13)}{39^2(38)} = .230$$

The data for Table 14.4.2 and similar data for the other 19 time periods are shown in Table 14.4.3. Using data from Table 14.4.3, we compute the log-rank statistic by Equation 12.7.7 as follows:

$$\chi^2_{\text{MH}} = \frac{(9 - 17.811)^2}{3.140} = 24.724$$

TABLE 14.4.2 Contingency Table for First Stratum (Time Period) for Calculating the Log-Rank Test, Example 14.4.1

	Low-Grade	High-Grade	Total
Deaths	0	1	1
Patients alive	25	13	38
Patients at risk	25	13	39

TABLE 14.4.3 Intermediate Calculations for the Log-Rank Test, Example 14.4.1

Time, t_i	a_i	c_i	$a_i + c_i$	b_i	d_i	$b_i + d_i$	n_i	e_i	y_i
0	0	25	25	1	13	14	39	0.641	0.230
2	0	25	25	1	12	13	38	0.658	0.225
3	0	25	25	1	11	12	37	0.676	0.219
4	0	25	25	1	10	11	36	0.694	0.212
6	0	25	25	1	9	10	35	0.714	0.204
7	0	24	24	1	8	9	33	0.727	0.198
9	0	23	23	2	6	8	31	1.484	0.370
11	1	22	23	0	6	6	29	0.793	0.164
12	0	22	22	1	5	6	28	0.786	0.168
15	2	20	22	0	5	5	27	1.630	0.290
16	0	20	20	1	4	5	25	0.800	0.160
17	0	20	20	1	3	4	24	0.833	0.139
21	1	19	20	0	3	3	23	0.870	0.113
23	0	19	19	1	2	3	22	0.864	0.118
27	0	18	18	1	1	2	20	0.900	0.090
29	1	17	18	0	1	1	19	0.947	0.050
33	1	16	17	0	1	1	18	0.944	0.052
79	1	13	14	0	1	1	15	0.933	0.062
102	1	10	11	0	1	1	12	0.917	0.076
212	1	1	2	0	0	0	2	1.000	0.000
Totals	9							17.811	3.140

Reference to Appendix Table F reveals that since $24.724 > 7.879$, the p value for this test is $< .005$. We, therefore, reject the null hypothesis that the survival experience is the same for patients with low-grade tumors and high-grade tumors and conclude that they are different.

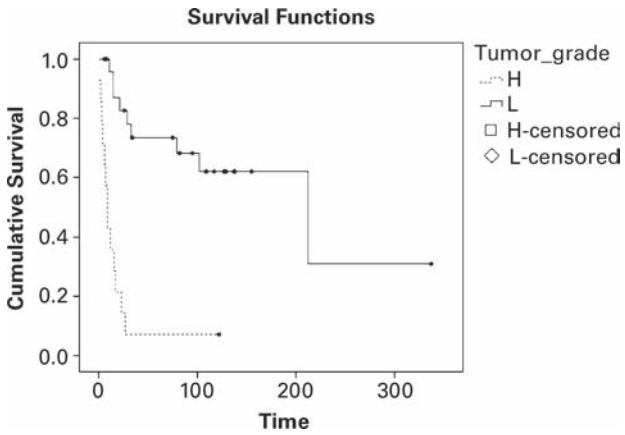
There are alternative procedures for testing the null hypothesis that two survival curves are identical. They include the Breslow test (also called the generalized Wilcoxon test) and the Tarone–Ware test. Both tests, as well as the log-rank test, are discussed in Parmar and Machin (7) and Allison (4). Like the log-rank test, the Breslow test and the Tarone–Ware test are based on the weighted differences between actual and expected numbers of deaths at the observed time points. Whereas the log-rank test ranks all deaths equally, the Breslow and Tarone–Ware tests give more weight to early deaths. For Example 12.8.1, SPSS computes a value of $24.93 (p < .001)$ for the Breslow test and a value of $25.22 (p < .001)$ for the Tarone–Ware test. Kleinbaum (27) discusses another test called the Peto test. Formulas for this test are found in Parmar and Machin (7). The Peto test also gives more weight to the early part of the survival curve, where we find the larger numbers of subjects at risk. When choosing a test, then, researchers who want to give more weight to the earlier part of the survival curve will select either the Breslow, the Tarone–Ware, or the Peto test. Otherwise, the log-rank test is appropriate.

We have covered only the basic concepts of survival analysis in this section. The reader wishing to pursue the subject in more detail may consult one or more of several books devoted to the topic, such as those by Kleinbaum (8), Lee (9), Marubini and Valsecchi (10), and Parmar and Machin (7).

Computer analysis

Several of the available statistical software packages, such as SPSS, are capable of performing survival analysis and constructing supporting graphs as described in this section.

A standard SPSS analysis of the data discussed in Examples 14.3.1 and 14.4.1 is shown in Figure 14.4.1. ■



Means and Medians for Survival Time

Tumor_ grade	Mean[a]				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
H	18.357	8.251	2.186	34.528	9.000	1.852	5.371	12.629
L	88.040	15.258	58.134	117.946	82.000	16.653	49.359	114.641
Overall	63.026	11.490	40.505	85.546	27.000	7.492	12.317	41.683

[a] Estimation is limited to the largest survival time if it is censored.

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	24.704	1	.000
Breslow (Generalized Wilcoxon)	24.927	1	.000
Tarone-Ware	25.217	1	.000

Test of equality of survival distributions for the different levels of tumor_grade.

FIGURE 14.4.1 SPSS output for Examples 14.3.1 and 14.4.1.

EXERCISES

14.4.1 If available in your library, read the article, "Impact of Obesity on Allogeneic Stem Cell Transplant Patients: A Matched Case-Controlled Study," by Donald R. Fleming et al. [*American Journal of Medicine, 102* (1997), 265–268] and answer the following questions:

(a) How was survival time determined?

(b) Why do you think the authors used the Wilcoxon test (Breslow test) for comparing the survival curves?

(c) Explain the meaning of the p values reported for Figures 1 through 4.

(d) What specific statistical results allow the authors to arrive at their stated conclusion?

14.4.2 If available in your library, read the article, "Improved Survival in Patients with Locally Advanced Prostate Cancer Treated with Radiotherapy and Goserelin," by Michel Bolla et al. [*New England Journal of Medicine, 337* (1997), 295–300], and answer the following questions:

(a) How was survival time determined?

(b) Why do you think the authors used the log-rank test for comparing the survival curves?

(c) Explain the meaning of the p values reported for Figures 1 and 2.

(d) What specific statistical results allow the authors to arrive at their stated conclusion?

14.4.3 Fifty subjects who completed a weight-reduction program at a fitness center were divided into two equal groups. Subjects in group 1 were immediately assigned to a support group that met weekly. Subjects in group 2 did not participate in support group activities. All subjects were followed for a period of 60 weeks. They reported weekly to the fitness center, where they were weighed and a determination was made as to whether they were within goal. Subjects were considered to be within goal if their weekly weight was within 5 pounds of their weight at time of completion of the weight-reduction program. Survival was measured from the date of completion of the weight-reduction program to the termination of follow-up or the point at which the subject exceeded goal. The following results were observed:

Subject	Time (Weeks)	Status (G = Within Goal G+ = Exceeded Goal L = Lost to Follow-Up)	Subject	Time (Weeks)	Status (G = Within Goal G+ = Exceeded Goal L = Lost to Follow-Up)
	Group 1			Group 2	
1	60	G	1	20	G+
2	32	L	2	26	G+
3	60	G	3	10	G+
4	22	L	4	2	G+
5	6	G+	5	36	G+
6	60	G	6	10	G+
7	60	G	7	20	G+
8	20	G+	8	18	L
9	32	G+	9	15	G+
10	60	G	10	22	G+
11	60	G	11	4	G+
12	8	G+	12	12	G+

(*Continued*)

Subject	Time (Weeks)	Status (G = Within Goal G+ = Exceeded Goal L = Lost to Follow-Up)	Subject	Time (Weeks)	Status (G = Within Goal G+ = Exceeded Goal L = Lost to Follow-Up)
Group 1			Group 2		
13	60	G	13	24	G+
14	60	G	14	6	G+
15	60	G	15	18	G+
16	14	L	16	3	G+
17	16	G+	17	27	G+
18	24	L	18	22	G+
19	34	L	19	8	G+
20	60	G	20	10	L
21	40	L	21	32	G+
22	26	L	22	7	G+
23	60	G	23	8	G+
24	60	G	24	28	G+
25	52	L	25	7	G+

Analyze these data using the methods discussed in this section.

14.5 COX REGRESSION: THE PROPORTIONAL HAZARDS MODEL

In previous chapters, we saw that regression models can be used for continuous outcome measures and for binary outcome measures (logistic regression). Additional regression techniques are available when the dependent measures may consist of a mixture of either time-to-event data or censored time observations. Returning to our example of a clinical trial of the effectiveness of two different medications to prevent a second myocardial infarction, we may wish to control for additional characteristics of the subjects enrolled in the study. For example, we would expect subjects to be different in their baseline systolic blood pressure measurements, family history of heart disease, weight, body mass, and other characteristics. Because all of these factors may influence the length of the time interval until a second myocardial infarction, we would like to account for these factors in determining the effectiveness of the medications. The regression method known as Cox regression (after D. R. Cox (11), who first proposed the method) or proportional hazard regression can be used to account for the effects of continuous and discrete covariate (independent variable) measurements when the dependent variable is possibly censored time-to-event data.

We describe this technique by first reviewing the *hazard function* from Section 14.2, which describes the conditional probability that an event will occur at a time just larger than t_i conditional on having survived event-free until time t_i. This function is often written as $h(t_i)$. The regression model requires that we assume the covariates have the effect of

either increasing or decreasing the hazard for a particular individual compared to some baseline value for the function. In our clinical trial example we might measure k covariates on each of the subjects where there are $i = 1, \ldots, n$ subjects and $h_0(t_i)$ is the baseline hazard function. We describe the regression model as

$$h(t_i) = h_0(t_i) \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_k z_{ik}) \tag{14.5.1}$$

The regression coefficients represent the change in the hazard that results from the risk factor, z_{ik}, that we have measured. Rearranging the above equation shows that the exponentiated coefficient represents the hazard ratio or the ratio of the conditional probabilities of an event. This is the basis for naming this method *proportional hazards regression*. You may recall that this is the same way we obtained the estimate of the odds ratio from the estimated coefficient when we discussed logistic regression in Chapter 11.

$$\frac{h(t_i)}{h_0(t_i)} = \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_k z_{ik}) \tag{14.5.2}$$

Estimating the covariate effects, $\hat{\beta}$ requires the use of a statistical software package because there is no straightforward single equation that will provide the estimates for this regression model. Computer output usually includes estimates of the regression coefficients, standard error estimates, hazard ratio estimates, and confidence intervals. In addition, computer output may also provide graphs of the hazard functions and survival functions for subjects with different covariate values that are useful to compare the effects of covariates on survival.

EXAMPLE 14.5.1

To determine whether time to relapse among drug users is related to patient age and/or the drug of choice, Cross (unpublished clinical data) reviewed a random sample of case files for high-risk drug users in an outpatient treatment clinic. The data represent the self-reported time that relapse occurred (or the time at which the patient was lost to follow-up), patient status, drug of choice, and patient age. The data are summarized in Table 14.5.1.

TABLE 14.5.1 Survival Data for Patients in an Outpatient Treatment Clinic

Subject	Time (Weeks)	Status 0 = Censored 1 = Relapse	Drug 1 = Opiate 2 = Other	Age	Subject	Time (Weeks)	Status 0 = Censored 1 = Relapse	Drug 1 = Opiate 2 = Other	Age
1	12	1	1	21	21	21	1	2	28
2	8	1	1	18	22	41	1	2	31
3	5	1	1	17	23	23	0	2	22
4	17	1	1	17	24	15	1	2	31
5	19	1	1	25	25	15	0	2	25
6	12	0	1	30	26	21	1	2	19

(*Continued*)

TABLE 14.5.1 (*Continued*)

Subject	Time (Weeks)	Status 0 = Censored 1 = Relapse	Drug 1 = Opiate 2 = Other	Age	Subject	Time (Weeks)	Status 0 = Censored 1 = Relapse	Drug 1 = Opiate 2 = Other	Age
7	10	1	1	16	27	45	1	2	21
8	11	1	1	23	28	37	1	2	23
9	5	1	1	31	29	51	1	2	15
10	2	1	1	21	30	50	1	2	29
11	10	1	1	19	31	42	1	2	28
12	7	0	1	18	32	21	1	2	31
13	19	1	1	18	33	20	1	2	31
14	11	1	1	21	34	15	1	2	26
15	11	1	1	23	35	40	1	2	28
16	19	1	1	15	36	39	1	2	31
17	19	1	1	17	37	33	1	2	23
18	24	1	1	21	38	37	1	2	23
19	21	1	1	22	39	15	0	2	29
20	14	1	1	17	40	52	0	2	37

Source: Data provided courtesy of Dr. Chad L. Cross.

For this example, we will employ the Cox Regression method algorithms provided in SPSS software. All references to tables and figures in the explanations below refer to Figure 14.5.1, which shows selected SPSS output for this example.

1. **Overall test.** SPSS provides an overall test of significance much like that reported for logistic regression discussed in Chapter 11. In this test, the likelihood is used to compare a model with no parameters (the null model) and a model with the variables of interest included. If there is a significant difference in the likelihood function between the model with parameters and the null model, then the Cox regression model is significant, and at least one of the variables of interest is significantly related to the outcome variable. An examination of the output shows that the Omnibus Test for Model Coefficients with age and drug entered in the model is significantly different from the null model, with $p < .001$.
2. **Variables in the model.** Next SPSS provides a table for each of the variables entered into the model. Much like a standard regression model, the model parameter, its standard error, and a significance test are provided to test the null, H_o: $\beta = 0$. For these data, type of drug was significantly predictive of time to relapse ($p < .001$), but age was not ($p = .792$).
3. **Survival curves.** Since drug of choice was found to be significantly related to the time of relapse, it is instructive to examine the survival curves for these data. It is clear from examining these curves that there is a difference in time to relapse, with those reporting opiate use as their primary drug of choice relapsing at a much faster rate than those reporting use of drugs other than opiates.
4. **Hazard ratios.** The hazard ratios are provided for each variable in the model. As in logistic regression where we calculated odds ratios, hazard ratios are found by

Omnibus Tests of Model Coefficients

−2 Log Likelihood	Overall (score)		
	Chi-square	df	Sig.
167.407	25.558	2	.000

Variables in the Equation

							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Drug	2.139	.531	16.239	1	.000	8.492	3.000	24.036
age	−.009	.032	.070	1	.792	.991	.930	1.057

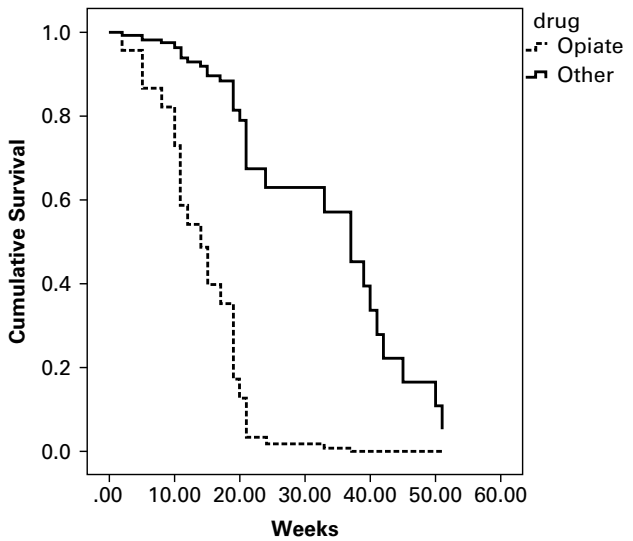


FIGURE 14.5.1 Cox Regression survival analysis output from SPSS software for Example 14.5.1.

calculating $\exp(\beta)$. Examining the variable drug, where opiates were used as the indicator variable in SPSS, the hazard of relapse is nearly 8.5 times more likely for opiates compared to other drugs, controlling for the covariate of age. Although we can calculate the hazard ratio for age in much the same way as for drug, it is often useful for quantitative covariates to consider calculating the function $100(\exp(\beta) - 1)$, which provides an estimate of the percent change in the hazard when the covariate increases by one unit. In the present example for age, this leads to

$100(.991 - 1) = -.9$. Therefore, for each 1 year increase in age, the hazard for relapse decreases by an average of about .9 percent.

5. **Conclusion.** Based on the results of this limited sample, we have learned that age of the patient, though not statistically significant, suggest that in general age may be somewhat protective in that risk of relapse decreases with age. We have also learned that those experiencing addiction to opiates are prone to relapse much earlier in their treatment. The results of this preliminary study may be used to develop further studies to determine if different, and perhaps more intensive, treatment programs are more successful for targeting those experience opiate addiction compared to other drugs. ■

Clearly Cox regression can become very complex as the number of variables increases. As with standard regression models discussed in early chapters, one may opt to use selection procedures (forward, backward, or stepwise) or examine interactions among variables in the models. Additionally, one may have time-dependent covariates in which the value of the covariate may change at each measurement time. Examples of this may be marriage or diagnosis with a health condition. These covariates are in contrast to time-constant covariates, which do not change (e.g., gender). In summary, Cox regression is a very useful technique for modeling survival data. For those interested in further reading, the texts by Kleinbaum and Klein (1), Lee (2), Hosmer and Lemeshow (3), and Allison (4) are highly recommended.

EXERCISES

14.5.1 In a study examining time-to-onset of cancer after exposure to UV light in rats, age (months) was used as a covariate in a Cox regression model. In the model, the parameter estimate for weight was .19 and had a p-value of .021. Provide an interpretation of this parameter estimate in terms of the hazard ratio.

14.5.2 In the study described in Exercise 14.5.1, the researchers were also interested to know if there was a difference between gender in the time it took to develop cancer. For gender, the parameter estimate was .77 and had a p-value of 0.014. Provide an interpretation of this parameter estimate in terms of the hazard ratio.

14.5.3 The intent of a study by Weaver et al. (A-4) was to assess whether occult lymph node metastases are important indicators of disease recurrence or survival in breast cancer patients. The data below provide some of the pertinent results of a Cox regression model for these data.

(a) Calculate the regression parameter coefficients for each variable.

(b) Provide an interpretation of these results using the concepts learned in this section.

Variable	Hazard Ratio (HR)	95% CI for HR	p-value
Age (50+ vs. <50)	1.69	(1.24, 2.31)	.001
Tumor size (>2 cm vs. ≤ 2 cm)	1.32	(.98, 1.76)	.060
Chemotherapy vs. no chemotherapy	.88	(.68, 1.13)	.31
Radiation vs. no radiation	0.54	(.40, .73)	.001

14.6 SUMMARY

In this chapter an introduction to time-to-event data was provided. In particular, the concept of data censoring, in which exact times are not known for subjects, was introduced. Distributions useful in survival analysis, including the cumulative distribution function, the survival function, and the hazard function were discussed. Calculating basic survival curves using the Kaplan–Meier procedure was discussed, as were methods for comparing survival curves using nonparametric methods. Regression concepts using Cox regression were provided, and detailed analysis of examples was given. The relationship of several methods covered in this chapter was tied to concepts learned earlier in the text, including linear regression, analysis of frequency data, and nonparametric statistics.

SUMMARY OF FORMULAS FOR CHAPTER 14

Formula Number	Name	Formula
14.2.1	Cumulative distribution function	$F(t) = P(T \leq t)$
14.2.2	Survival function	$S(t) = 1 - F(t)$
14.2.3	Probability distribution function	$f(t) = \dfrac{P(t \leq T < t + \Delta t)}{\Delta t}, \quad \text{as} \quad \Delta t \to 0$
14.2.4	Relationship of probability distribution function to the cumulative distribution function and the survival function	$f(t) = \dfrac{\Delta F(t)}{\Delta t} = -\dfrac{\Delta S(t)}{\Delta t}$
14.2.5	Hazard function	$h(t) = \dfrac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}, \quad \text{as} \quad \Delta t \to 0$
14.2.7	Relationship of the hazard function to the probability distribution function and the survival function	$h(t) = \dfrac{f(t)}{S(t)}$
14.3.1	Survival probability	$\hat{p}_i = \dfrac{\text{number of subjects surviving at least } (t-1) \text{ time period who also survive the } t\text{th period}}{\text{number of subjects alive at end of time period } (t-1)}$
14.3.2	Estimated survival function	$\hat{S}(t) = \hat{p}_1 \times \hat{p}_2 \times \cdots \times \hat{p}_t$

14.5.1	Hazard regression model	$h(t_i) = h_0(t_i)\exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_k z_{ik})$
14.5.2	Proportional hazard model	$\frac{h(t_i)}{h_o(t_i)} = \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_k z_{ik})$
Symbol Key	• β = regression coefficient • Δ = change • $F(t)$ = cumulative distribution function • $f(t)$ = probability density function • $h(t)$ = hazard function • p = probability • $S(t)$ = survival function • T = time of interest • t = time to event • z = risk factor in Cox regression	

REVIEW QUESTIONS AND EXERCISES

1. Describe in words the concept of data censoring.
2. Define the following:
   (a) Hazard ratio
   (b) Hazard function
   (c) Probability distribution function
   (d) Survival function
   (e) Kaplan–Meier estimate
3. Explain the concepts underlying the Cox regression model.
4. What is the difference between right censoring and left censoring? Provide an example of each.
5. Discuss why it is often preferable to use a nonparametric test for comparisons of survival curves.
6. Why is Cox regression called a "proportional hazards" model?
7. If the probability distribution function at time 5 is equal to .25 and the survival function at time 5 is equal to .15, what is the hazard function at time 5?
8. If we find that a measurement in the time interval between time 2 and 10 results in a probability distribution function estimate of 0.03, what is the estimated change in the cumulative distribution function?
9. Using the data from question 8, what is the estimated change in the survival function?
10. Explain why the cumulative distribution function and the survival function are mirror images of one another.
11. The objective of a study by Lee et al. (A-5) was to improve understanding of the biologic behavior of gastric epithelioid stromal tumors. They studied the clinical features, histologic findings, and DNA ploidy of a series of the tumors to identify factors that might distinguish between benign and malignant variants of these tumors and have relevance for prognosis. Fifty-five patients with tumors

were classified on the basis of whether their tumors were high-grade malignant (grade 2), low-grade malignant (grade 1), or benign (grade 0). Among the data collected were the following:

Patient	Tumor Grade	Outcome (1 = Death from Disease)	Number of Days to Last Follow-Up or Death	Patient	Tumor Grade	Outcome (1 = Death from Disease)	Number of Days to Last Follow-Up or Death
1	0	0	87	8	0	0	1616
2	0	0	775	9	0	0	1982
3	0	0	881	10	0	0	2035
4	0	0	914	11	0	0	2191
5	0	0	1155	12	0	0	2472
6	0	0	1162	13	0	0	2527
7	0	0	1271	14	0	0	2782
15	0	0	3108	36	0	0	7318
16	0	0	3158	37	0	0	7447
17	0	0	3609	38	0	0	9525
18	0	0	3772	39	0	0	9938
19	0	0	3799	40	0	0	10429
20	0	0	3819	41	1	1	450
21	0	0	4586	42	1	1	556
22	0	0	4680	43	1	1	2102
23	0	0	4989	44	1	0	2756
24	0	0	5675	45	1	0	3496
25	0	0	5936	46	1	1	3990
26	0	0	5985	47	1	0	5686
27	0	0	6175	48	1	0	6290
28	0	0	6177	49	1	0	8490
29	0	0	6214	50	2	1	106
30	0	0	6225	51	2	1	169
31	0	0	6449	52	2	1	306
32	0	0	6669	53	2	1	348
33	0	0	6685	54	2	1	549
34	0	0	6873	55	2	1	973
35	0	0	6951				

Source: Data provided courtesy of Dr. Michael B. Farnell.

12. Girard et al. (A-6) conducted a study to identify prognostic factors of improved survival after resection of isolated pulmonary metastases (PM) from colorectal cancer. Among the data collected were the following regarding number of resected PM, survival, and outcome for 77 patients who underwent a complete resection at the first thoracic operation:

Patient	Number of Resected PM	Survival (Months)	Status	Patient	Number of Resected PM	Survival (Months)	Status
1	1	24	Alive	8	1	15	Dead
2	1	67	Alive	9	1	10	Dead
3	1	42	Alive	10	1	41	Dead
4	> 1	28	Dead	11	> 1	41	Dead
5	1	37	Dead	12	1	27	Dead
6	1	133	Alive	13	1	93	Alive
7	1	33	Dead	14	> 1	0	Dead
15	1	60	Dead	47	1	54	Dead

(*Continued*)

Patient	Number of Resected PM	Survival (Months)	Status	Patient	Number of Resected PM	Survival (Months)	Status
16	1	43	Dead	48	> 1	57	Alive
17	> 1	73	Alive	49	> 1	16	Dead
18	1	55	Alive	50	1	29	Dead
19	1	46	Dead	51	1	14	Dead
20	1	66	Alive	52	> 1	29	Dead
21	1	10	Dead	53	> 1	99	Dead
22	> 1	3	Dead	54	> 1	23	Dead
23	> 1	7	Dead	55	1	74	Alive
24	> 1	129	Alive	56	1	169	Alive
25	1	19	Alive	57	> 1	24	Dead
26	> 1	15	Dead	58	> 1	9	Dead
27	1	39	Alive	59	1	43	Dead
28	1	15	Dead	60	1	3	Alive
29	> 1	30	Dead	61	> 1	20	Dead
30	1	35	Alive	62	1	2	Dead
31	> 1	18	Dead	63	> 1	41	Dead
32	1	27	Dead	64	> 1	27	Dead
33	1	121	Alive	65	1	45	Alive
34	> 1	8	Dead	66	1	26	Dead
35	1	24	Alive	67	> 1	10	Dead
36	1	127	Alive	68	1	143	Alive
37	1	26	Dead	69	1	16	Dead
38	> 1	7	Dead	70	1	29	Alive
39	> 1	26	Dead	71	1	17	Dead
40	> 1	17	Dead	72	> 1	20	Dead
41	1	18	Dead	73	1	92	Alive
42	1	17	Dead	74	> 1	15	Dead
43	> 1	10	Dead	75	1	5	Dead
44	> 1	33	Dead	76	> 1	73	Alive
45	> 1	42	Alive	77	1	19	Dead
46	1	40	Alive				

Source: Data provided courtesy of Dr. Philippe Girard.

13. In a study by Alicikus et al. (A-7), long-term control of prostate cancer receiving radiotherapy was examined in patients after 10 years. The authors using Cox regression analysis to analyze these data, which resulted in the data summarized in the table below. For these data:

(a) Calculate the parameter estimates for the Cox regression model.

(b) Provide an explanation of the hazard ratios (HR) and their meaning.

(c) For age, provide an alternative measure for the HR and provide its meaning in terms of the percent change in years.

Variable	Hazard Ratio (HR)	95% CI for HR	*p*-value
Age	1.02	(.96, 1.08)	.51
Hormone therapy (yes vs. no)	.89	(.44, 1.81)	.75
Pre-PSA, >10 ng/mL vs. ≤10 ng/mL	2.41	(1.19, 4.88)	.015
Tumor classification	1.42	(1.17, 1.71)	<.001

Source: Zumre A. Alicikus, Yoshiya Yamada, Zhigang Zhang, Xin Pei, Margie Hung, Marisa Kollmeier, Brett Cox, and Michael J. Zelefsky, "Ten-year Outcomes of High-Dose, Intensity-Modulated Radiotherapy for Localized Prostate Cancer," *Cancer*, *117* (2010), 1429–1437.

REFERENCES

Methodology References

1. David G. Kleinbaum and Mitchel Klein, *Survival Analysis: A Self-Learning Text*, Second Edition, Springer, New York, 2005.
2. Elisa T. Lee, *Statistical Methods for Survival Data Analysis*, Third Edition, Wiley, New York, 2003.
3. David W. Hosmer, JR. and Stanley Lemeshow, *Applied Survival Analysis: Regression Modeling of Time to Event data*, Wiley, New York, 1999.
4. Paul D. Allison, *Survival Analysis using SAS®: A Practical Guide*, Second Edition, SAS Publishing, Cary, NC, 2010.
5. E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, *53* (1958), 457–481.
6. Nathan Mantel, "Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration," *Cancer Chemotherapy Reports*, *50* (March1966), 163–170.
7. Maheesh K. B. Parmar and David Machin, *Survival Analysis: A Practical Approach*, Wiley, New York, 1995.
8. David G. Kleinbaum, *Survival Analysis: A Self-Learning Text*, Springer, New York, 1996.
9. Elisa T. Lee, *Statistical Methods for Survival Data Analysis*, Lifetime Learning Publications, Belmont, CA, 1980.
10. Ettore Marubini and Maria Grazia Valsecchi, *Analysing Survival Data from Clinical Trials and Observational Studies*, Wiley, New York, 1995.
11. David R. Cox, "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, *B34* (1972), 187–220.

Application References

1. Nael Martini, Andrew G. Huvos, Michael E. Burt, Robert T. Heelan, Manjit S. Bains, Patricia M. McCormack, Valerie W. Rusch, Michael Weber, Robert J. Downey, and Robert J. Ginsberg, "Predictions of Survival in Malignant Tumors of the Sternum," *Journal of Thoracic and Cardiovascular Surgery*, *111* (1996), 96–106.
2. Massimo E. Dottorini, Agnese Assi, Maria Sironi, Gabriele Sangalli, Gianluigi Spreafico, and Luigia Colombo, "Multivariate Analysis of Patients with Medullary Thyroid Carcinoma," *Cancer*, *77* (1996), 1556–1565.
3. Mary Ann Banerji, Rochelle L. Chaiken, and Harold E. Lebovitz, "Long-Term Normoglycemic Remission in Black Newly Diagnosed NIDDM Subjects," *Diabetes*, *45* (1996), 337–341.
4. Donald L. Weaver, Takamaru Ashikaga, David N. Krag, Joan M. Skelly, Stewart J. Anderson, Seth P. Harlow, Thomas B. Julian, Eleftherios P. Mamounas, and Normal Wolmark, "Effect of Occult Metastases on Survival in Node-Negative Breast Cancer," *The New England Journal of Medicine*, *364* (2011), 412–421.
5. Joy S. Y. Lee, Antonio G. Nascimento, Michael B. Farnell, J. Aidan Carney, William S. Harmsen, and Duane M. Ilstrup, "Epithelioid Gastric Stromal Tumors (Leiomyoblastomas): A Study of Fifty-five Cases," *Surgery*, *118* (1995), 653–661.
6. Philippe Girard, Michel Ducreux, Pierre Baldeyrou, Philippe Lasser, Brice Gayet, Pierre Ruffié, and Dominique Grunenwald, "Surgery for Lung Metastases from Colorectal Cancer: Analysis of Prognostic Factors," *Journal of Clinical Oncology*, *14* (1996), 2047–2053.
7. Zumre A. Alicikus, Yoshiya Yamada, Zhigang Zhang, Xin Pei, Margie Hung, Marisa Kollmeier, Brett Cox, and Michael J. Zelefsky, "Ten-year Outcomes of High-Dose, Intensity-Modulated Radiotherapy for Localized Prostate Cancer," *Cancer*, *117* (2010), 1429–1437.