CHAPTER 7

Categorical Data: Contingency Tables

7.1 INTRODUCTION

In Chapter 6, *discrete variables* came up by counting the number of times that specific outcomes occurred. In looking at the presence or absence of a risk factor and a disease, *odds ratio* and *relative risk* were introduced. In doing this, we looked at the relationship between two discrete variables; each variable took on one of two possible states (i.e., risk factor present or absent and disease present or absent). In this chapter we show how to analyze more general discrete data. Two types of generality are presented.

The first generalization considers two jointly distributed discrete variables. Each variable may take on more than two possible values. Some examples of discrete variables with three or more possible values might be: smoking status (which might take on the values "never smoked," "former smoker," and "current smoker"); employment status (which could be coded as "full-time," "part-time," "unemployed," "unable to work due to medical reason," "retired," "quit," and "other"); and clinical judgment of improvement (classified into categories of "considerable improvement," "slight improvement," "no change," "slight worsening," "considerable worsening," and "death").

The second generalization allows us to consider three or more discrete variables (rather than just two) at the same time. For example, method of treatment, gender, and employment status may be analyzed jointly. With three or more variables to investigate, it becomes difficult to obtain a "feeling" for the interrelationships among the variables. If the data fit a relatively simple mathematical model, our understanding of the data may be greatly increased.

In this chapter, our first *multivariate statistical model* is encountered. The model is the *log-linear model* for multivariate discrete data. The remainder of the book depends on a variety of models for analyzing data; this chapter is an exciting, important, and challenging introduction to such models!

7.2 TWO-WAY CONTINGENCY TABLES

Let two or more discrete variables be measured on each unit in an experiment or observational study. In this chapter, methods of examining the relationship among the variables are studied. In most of the chapter we study the relationship of two discrete variables. In this case we count the number of occurrences of each pair of possibilities and enter them in a table. Such tables are called *contingency tables*. Example 7.1 presents two contingency tables.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

Example 7.1. In 1962, Wangensteen et al., published a paper in the *Journal of the American Medical Association* advocating gastric freezing. A balloon was lowered into a subject's stomach, and coolant at a temperature of −17 to −20°C was introduced through tubing connected to the balloon. Freezing was continued for approximately 1 hour. The rationale was that gastric digestion could be interrupted and it was thought that a duodenal ulcer might heal if treatment could be continued over a period of time. The authors advanced three reasons for the interruption of gastric digestion: (1) interruption of vagal secretory responses; (2) "rendering of the central mucosa nonresponsive to food ingestion . . . "; and (3) "impairing the capacity of the parietal cells to secrete acid and the chief cells to secrete pepsin." Table 7.1 was presented as evidence for the effectiveness of gastric freezing. It shows a decrease in acid secretion.

On the basis of this table and other data, the authors state: "These data provide convincing objective evidence of significant decreases in gastric secretory responses attending effective gastric freezing" and conclude: "When profound gastric hypothermia is employed with resultant freezing of the gastric mucosa, the method becomes a useful agent in the control of many of the manifestations of peptic ulcer diathesis. Symptomatic relief is the rule, followed quite regularly by x-ray evidence of healing of duodenal ulcer craters and evidence of effective depression of gastric secretory responses." *Time* [1962] reported that "all [the patients'] ulcers healed within two to six weeks."

However, careful studies attempting to confirm the foregoing conclusion failed. Two studies in particular failed to confirm the evidence, one by Hitchcock et al. [1966], the other by Ruffin et al. [1969]. The latter study used an elaborate sham procedure (control) to simulate gastric freezing, to the extent that the tube entering the patient's mouth was cooled to the same temperature as in the actual procedure, but the coolant entering the stomach was at room temperature, so that no freezing took place. The authors defined an endpoint to have occurred if one of the following criteria was met: "perforation; ulcer pain requiring hospitalization for relief; obstruction, partial or complete, two or more weeks after hyperthermia; hemorrhage, surgery for ulcer; repeat hypothermia; or x-ray therapy to the stomach."

Several institutions cooperated in the study, and to ensure objectivity and equal numbers, random allocations to treatment and sham were balanced within groups of eight. At the termination of the study, patients were classified as in Table 7.2. The authors conclude: "The results of

Table 7.1 Gastric Response of 10 Patients with Duodenal Ulcer Whose Stomachs Were Frozen at −17 to −20°C for 1 Hour

		Average Percent Decrease in HCl after Gastric Freezing		
Patients	Patients with Decrease in Free HCl	Overnight Secretion	Peptone Stimulation	Insulin
10	10[a]	87	51	71

Source: Data from Wangensteen et al. [1962].

[a] All patients, except one, had at least a 50% decrease in free HCl in overnight secretion.

Table 7.2 Causes of Endpoints

Group	Patients	With Hemorrhage	With Operation	With Hospitalization	Not Reaching Endpoint
F (freeze)	69	9	17	9	34
S (sham)	68	9	14	7	38

Source: Data from Ruffin et al. [1969].

Table 7.3 Contingency Table for Gastric Freezing Data

	j			
i	1	2	$\cdots$	c
1	n_{11}	n_{12}	$\cdots$	n_{1c}
2	n_{21}	n_{22}	$\cdots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

this study demonstrate conclusively that the 'freezing' procedure was not better than the sham in the treatment of duodenal ulcer, confirming the work of others. ... It is reasonable to assume that the relief of pain and subjective improvement reported by early investigators was probably due to the psychological effect of the procedure."

Contingency tables set up from two variables are called *two-way tables*. Let the variable corresponding to rows have r (for "row") possible outcomes, which we index by i $(i = 1, 2, \dots, r)$. Let the variables corresponding to the column headings have c (for "column") possible states indexed by j $(j = 1, 2, \dots, c)$. One speaks of an *$r \times c$ contingency table*. Let n_{ij} be the number of observations corresponding to the ith state of the row variable and the jth state of the column variable. In the example above, $n_{11} = 9, n_{12} = 17, n_{13} = 9, n_{14} = 34, n_{21} = 9, n_{22} = 14, n_{23} = 7$, and $n_{24} = 38$. In general, the data are presented as shown in Table 7.3. Such tables usually arise in one of two ways:

1. A sample of observations is taken. On each unit we observe the values of two traits. Let π_{ij} be the probability that the row variable takes on level i and the column variable takes on level j. Since one of the combinations must occur,

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij} = 1 \qquad (1)$$

2. Each row corresponds to a sample from a different population. In this case, let π_{ij} be the probability the column variable takes on state j when sampling from the ith population. Thus, for each i,

$$\sum_{j=1}^{c}\pi_{ij} = 1 \qquad (2)$$

If the samples correspond to the column variable, the π_{ij} are the probabilities that the row variable takes on state i when sampling from population j. In this circumstance, for each j,

$$\sum_{i=1}^{r}\pi_{ij} = 1 \qquad (3)$$

Table 7.2 comes from the second model since the treatment is assigned by the experimenter; it is not a trait of the experimental unit. Examples for the first model are given below.

The usual null hypothesis in a model 1 situation is that of independence of row and column variables. That is (assuming row variable $= i$ and column variable $= j$), $P[i \text{ and } j] = P[i]P[j]$,

$$H_0\colon \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

In the model 2 situation, suppose that the row variable identifies the population. The usual null hypothesis is that all r populations have the same probabilities of taking on each value of the column variable. That is, for any two rows, denoted by i and i', say, and all j,

$$H_0\colon \pi_{ij} = \pi_{i'j}$$

If one of these hypotheses holds, we say that there is *no association*; otherwise, the table is said to have *association* between the categorical variables.

We will use the following notation for the sum over the elements of a row and/or column: $n_{i\cdot}$ is the sum of the elements of the ith row; $n_{\cdot j}$ is the sum of the elements of the jth column:

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}, \qquad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, \qquad n_{\cdot\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$$

It is shown in Note 7.1 that under either model 1 or model 2, the null hypothesis is reasonably tested by comparing n_{ij} with

$$\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

The latter is the value expected in the ijth cell given the observed marginal configuration and assuming either of the null hypotheses under model 1 or model 2. This is shown as

$n_{11} = 9$	$n_{12} = 17$	$n_{13} = 9$	$n_{14} = 34$	$n_{1\cdot} = 69$
$n_{21} = 9$	$n_{22} = 14$	$n_{23} = 7$	$n_{24} = 38$	$n_{2\cdot} = 68$
$n_{\cdot 1} = 18$	$n_{\cdot 2} = 31$	$n_{\cdot 3} = 16$	$n_{\cdot 4} = 72$	$n_{\cdot\cdot} = 137$

Under the null hypothesis, the table of expected values $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ is

$69 \times 18/137$	$69 \times 31/137$	$69 \times 16/137$	$69 \times 72/137$
$68 \times 18/137$	$68 \times 31/137$	$68 \times 16/137$	$68 \times 72/137$

or

9.07	15.61	8.06	36.26
8.93	15.39	7.94	35.74

It is a remarkable fact that both null hypotheses above may be tested by the χ^2 statistic,

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}}$$

Note that n_{ij} is the observed cell entry; $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ is the expected cell entry, so this statistic may be remembered as

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

For example, the array above gives

$$X^2 = \frac{(9-9.07)^2}{9.07} + \frac{(17-15.61)^2}{15.61} + \frac{(9-8.06)^2}{8.06}$$
$$+\frac{(34-36.26)^2}{36.26} + \frac{(9-8.93)^2}{8.93} + \frac{(14-15.39)^2}{15.39}$$
$$+\frac{(7-7.94)^2}{7.94} + \frac{(38-35.76)^2}{35.76} = 0.752$$

Under the null hypothesis, the X^2 statistic has approximately a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. This approximation is for large samples and is appropriate when all of the *expected* values, $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$, are 5 or greater. There is some evidence to indicate that the approximation is valid if all the expected values, except possibly one, are 5 or greater.

For our example, the degrees of freedom for the example are $(2-1)(4-1) = 3$. The rejection region is for X^2 too large. The 0.05 critical value is 7.81. As $0.752 < 7.81$, we do *not* reject the null hypothesis at the 0.05 significance level.

Example 7.2. Robertson [1975] examined seat belt use in automobiles with starter interlock and buzzer/light systems. The use or nonuse of safety belts by drivers in their vehicles was observed at 138 sites in Baltimore, Maryland; Houston, Texas; Los Angeles, California; the New Jersey suburbs; New York City; Richmond, Virginia; and Washington, DC during late 1973 and early 1974. The sites were such that observers could see whether or not seat belts were being used. The sites were freeway entrances and exits, traffic-jam areas, and other points where vehicles usually slowed to less than 15 miles per hour. The observers dictated onto tape the gender, estimated age, and racial appearance of the driver of the approaching car; as the vehicles slowed alongside, the observer recorded whether or not the lap belt and/or shoulder belt was in use, not in use, or could not be seen. The license plate numbers were subsequently sent to the appropriate motor vehicle administration, where they were matched to records from which the manufacturer and year were determined. In the 1973 models, a buzzer/light system came on when the seat belt was not being used. The buzzer was activated for at least 1 minute when the driver's seat was occupied, the ignition switch was on, the transmission gear selector was in a forward position, and the driver's lap belt was not extended at least 4 inches from its normal resting position. Effective on August 15, 1973, a federal standard required that the automobile could be started only under certain conditions. In this case, when the driver was seated, the belts had to be extended more than 4 inches from their normally stored position and/or latched. Robertson states that as a result of the strong negative public reaction to the interlock system, federal law has banned the interlock system. Data on the buzzer/light-equipped models and interlock-equipped models are given in Table 7.4. As can be seen from the table, column percentages were presented to aid assimilation of the information in the table.

Table 7.4 Robertson [1975] Seat Belt Data

	1973 Models (Buzzer/Light)		1974 Models (Interlock)		
Belt Use	%	Number	%	Number	Total
Lap and shoulder	7	432	48	1007	1439
Lap only	21	1262	11	227	1489
None	72	4257	41	867	5124
Total	100	5951	100	2101	8052

Percentages in two-way contingency tables are useful in aiding visual comprehension of the contents. There are three types of percent tables:

1. *Column percent tables* give the percentages for each column (the columns add to 100%, except possibly for rounding errors). This is best for comparing the distributions of different columns.
2. *Row percent tables* give the percentages for each row (the rows add to 100%). This is best for comparing the distributions of different rows.
3. The *total percent table* gives percentages, so that all the entries in a table add to 100%. This aids investigation of the proportions in each combination.

The column percentages in Table 7.4 facilitate comparison of seat belt use in the 1973 buzzer/light models and the 1974 interlock models. They illustrate that there are strategies for getting around the interlock system, such as disabling it, connecting the seat belt and leaving it connected on the seat, as well as possible other strategies, so that even with an interlock system, not everyone uses it. The computed value of the chi-square statistic for this table is 1751.6 with two degrees of freedom. The p-value is effectively zero, as shown in Table A.3 in the Appendix.

Given that we have a statistically significant association, the next question that arises is: To what may we attribute this association? To determine why the association occurs, it is useful to have an idea of which entries in the table differ more than would be expected by chance from their value under the null hypothesis of no association. Under the null hypothesis, for each entry in the table, the following *adjusted residual value* is approximately distributed as a standard normal distribution. The term *residual* is used since it looks at the difference between the observed value and the value expected under the null hypothesis. This difference is then standardized by its standard error,

$$z_{ij} = \frac{n_{ij} - (n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot})}{\sqrt{n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}\,(1 - n_{i\cdot}/n_{\cdot\cdot})\left(1 - n_{\cdot j}/n_{\cdot\cdot}\right)}} \tag{4}$$

For example, for the (1, 1) entry in the table, a standardized residual, is given by

$$\frac{(432 - 1439 \times 5951/8052)}{\sqrt{\dfrac{1439(5951)}{8052}\left(1 - \dfrac{1439}{8052}\right)\left(1 - \dfrac{5951}{8052}\right)}} = 41.83$$

The matrix of the residual values observed with the corresponding normal probability p-values is given in Table 7.5. Note that the values add to zero for the residuals across each row. This occurs because there are only two columns. The adjusted residual values observed are so far from zero that the normal p-values are miniscule.

In general, there is a problem in looking at a contingency table with many cells. Because there are a large number of residual values in the table, it may be that one or more of them differs by chance from zero at the 5% significance level. Even *under the null hypothesis*, because of the many possibilities examined, *this would occur much more than 5% of the time.* One conservative way to deal with this problem is to multiply the p-values by the number of rows minus one and the number of columns minus one. If the corresponding p-value is less than 0.05, one can conclude that the entry is different from that expected by the null hypothesis at the 5% significance level *even after looking at all of the different entries.* (This problem of looking at many possibilities, called the *multiple comparison problem*, is dealt with in considerable detail in Chapter 12.) For this example, even after multiplying by the number of rows minus one and the number of columns minus one, all of the entries differ from those expected under the null hypothesis. Thus, one can conclude, using the sign of the residual to tell us whether the

Table 7.5 Adjusted Residual Values (Example 7.2)

i	j	Residual (Z_{ij})	p-value	p-value × $(r-1) \times (c-1)$
1	1	−41.83	0+	0+
1	2	41.83	0+	0+
2	1	10.56	3×10^{-22}	6×10^{-22}
2	2	−10.56	3×10^{-22}	6×10^{-22}
3	1	24.79	9×10^{-53}	2×10^{-52}
3	2	−24.79	9×10^{-53}	2×10^{-52}

percentage is too high or too low, that in the 1973 models there is less lap and shoulder belt use than in the 1974 models. Further, if we look at the "none" category, there are fewer people without any belt use in the 1974 interlock models than in the 1973 buzzer/light-equipped models. One would conclude that the interlock system, although a system disliked by the public, was successful as a public health measure in increasing the amount of seat belt use.

Suppose that we decide there is an association in a contingency table. We can interpret the table by using residuals (as we have done above) to help to find out whether particular entries differ more than expected by chance. Another approach to interpretation is to characterize numerically the amount of association between the variables, or proportions in different groups, in the contingency table. To date, no single measure of the amount of association in contingency tables has gained widespread acceptance. There have been many proposals, all of which have some merit. Note 7.2 presents some measures of the amount of association.

7.3 CHI-SQUARE TEST FOR TREND IN 2 × k TABLES

There are a variety of techniques for improving the statistical power of χ^2 tests. Recall that power is a function of the alternative hypothesis. One weakness of the chi-square test is that it is an "omnibus" test; it tests for independence vs. dependence without specifying the nature of the latter. In some cases, a small subset of alternative hypotheses may be specified to increase the power of the chi-square test by defining a special test. One such situation occurs in $2 \times k$ tables when the alternative hypothesis is that there is an ordering in the variable producing the k categories. For example, exposure categories can be ordered, and the alternative hypothesis may be that the probability of disease *increases* with increasing exposure.

In this case the row variable takes on one of two states (say + or − for definiteness). For each state of the column variable ($j = 1, 2, \ldots, k$), let π_j be the conditional probability of a positive response. The test for trend is designed to have statistical power against the alternatives:

$$H_1 : \pi_1 \le \pi_2 \le \cdots \le \pi_k, \qquad \text{with at least one strict inequality}$$

$$H_2 : \pi_1 \ge \pi_2 \ge \cdots \ge \pi_k, \qquad \text{with at least one strict inequality}$$

That is, the alternatives of interest are that the proportion of + responses increases or decreases with the column variable. For these alternatives to be of interest, the column variable will have a "natural" ordering. To compute the statistic, a score needs to be assigned to each state j of the column variable. The scores x_j are assigned so that they increase or decrease. Often, the x_j are consecutive integers. The data are laid out as shown in Table 7.6.

Table 7.6 Scores Assigned to State *j*

	j				
i	1	2	$\cdots$	k	Total
1+	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1\cdot}$
2−	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	$n_{\cdot\cdot}$
Score	x_1	x_2	$\cdots$	x_k	

Before stating the test, we define some notation. Let

$$[n_1x] = \sum_{j=1}^{k} n_{1j}x_j - \frac{n_{1\cdot}\sum n_{\cdot j}x_j}{n_{\cdot\cdot}}$$

and

$$[x^2] = \sum_{j=1}^{k} n_{\cdot j}x_j^2 - \frac{(\sum n_{\cdot j}x_j)^2}{n_{\cdot\cdot}}$$

and

$$p = \frac{n_{1\cdot}}{n_{\cdot\cdot}}$$

Then the chi-square test for trend is defined to be

$$X^2_{\text{trend}} = \frac{[n_1x]^2}{[x^2]p(1-p)}$$

and when there is no association, this quantity has approximately a chi-square distribution with one degree of freedom. [In the terminology of Chapter 9, this is a chi-square test for the slope of a weighted regression line with dependent variable $p_j = n_{1j}/n_{\cdot j}$, predictor variable x_j, and weights $n_{1j}/p(1-p)$, where $j = 1, 2, \ldots, k$.]

Example 7.3. For an example of this test, we use data of Maki et al. [1977], relating risk of catheter-related infection to the duration of catheterization. An infection was considered to be present if there were 15 or more colonies of microorganisms present in a culture associated with the withdrawn catheter. A part of the data dealing with the number of positive cultures as related to duration of catheterization is given in Table 7.7. A somewhat natural set of values of the scores x_i is the duration of catheterization in days. The designation ≥4 is, somewhat arbitrarily, scored 4.

Before carrying out the analysis, note that a graph of the proportion of positive cultures vs. duration such as in the one shown in Figure 7.1 clearly suggests a trend. The general chi-square test on the 2 × 4 table produces a value of $X^2 = 6.99$ with three degrees of freedom and a significance level of 0.072.

Table 7.7 Relations of Results of Semiquantitative Culture and Catheterization

	Duration of Catheterization (days)				
Culture	1	2	3	≥ 4	Total
Positive[a]	1[b]	5	5	14	25
Negative	46	64	39	76	225
Total	47	69	44	90	250

Source: Data from Maki et al. [1977].

[a] Culture is positive if 15 or more colonies on the primary plate.

[b] Numbers in the body of the table are the numbers of catheters.

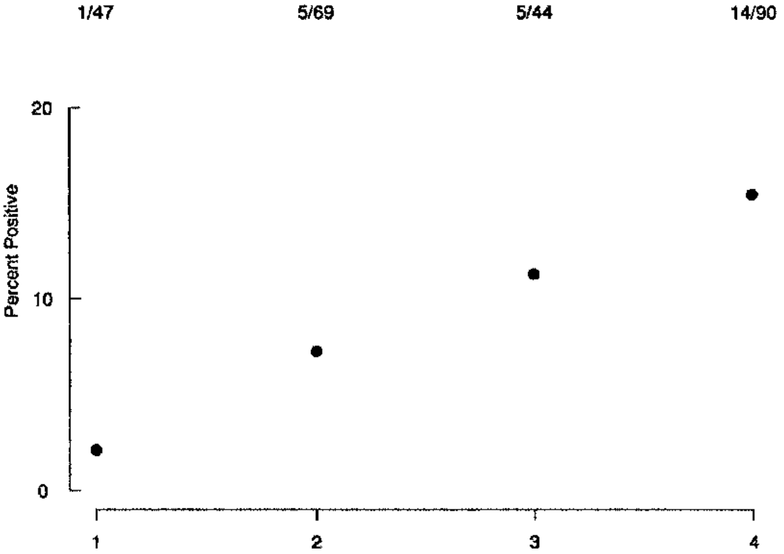

Figure 7.1 Graph of percentage of cultures positive vs. duration of catheterization. The fractions 1/47, etc., are the number of positive cultures to the total number of cultures for a particular day. (Data from Maki et al. [1977]; see Table 7.7.)

To calculate the chi-square test for trend, we calculate the quantities $[n_1x]$, $[x^2]$, and p as defined above.

$$[n_1x] = 82 - \frac{(25)(677)}{250} = 14.3$$

$$[x^2] = 2159 - \frac{677^2}{250} \doteq 325.6840$$

$$p = \frac{25}{250} = 0.1, \qquad (1-p) = 0.9$$

$$X^2_{\text{trend}} = \frac{[n_1x]^2}{[x^2]p(1-p)} \doteq \frac{14.3^2}{325.6840(0.1)(0.9)} \doteq 6.98$$

This statistic has one degree of freedom associated with it, and from the chi-square Table A.3, it can be seen that $0.005 < p < 0.01$; hence there is a significant linear trend.

Note two things about the chi-square test for trend. First, the degrees of freedom are one, *regardless* of how large the value k. Second, the values of the scores chosen (x_j) are not too crucial, and evenly spaced scores will give more statistical power against a trend than will the usual χ^2 test. The example above indicates one type of contingency table in which ordering is clear: when the categories result from grouping a continuous variable.

7.4 KAPPA: MEASURING AGREEMENT

It often happens in measuring or categorizing objects that the variability of the measurement or categorization is investigated. For example, one might have two physicians independently judge a patient's status as "improved," "remained the same," or "worsened." A study of psychiatric patients might have two psychiatrists independently classifying patients into diagnostic categories. When we have two discrete classifications of the same object, we may put the entries into a two-way *square* ($r = c$) contingency table. The chi-square test of this chapter may then be used to test for association. Usually, when two measurements are taken of the same objects, there is not much trouble showing association; rather, the concern is to study the degree or amount of agreement in the association. This section deals with a statistic, *kappa* (κ), designed for such situations. We will see that the statistic has a nice interpretation; the value of the statistic can be taken as a measure of the degree of agreement. As we develop this statistic, we shall illustrate it with the following example.

Example 7.4. Fisher et al. [1982] studied the reproducibility of coronary arteriography. In the coronary artery surgery study (CASS), coronary arteriography is the key diagnostic procedure. In this procedure, a tube is inserted into the heart and fluid injected that is opaque to x-rays. By taking x-ray motion pictures, the coronary arteries may be examined for possible narrowing, or *stenosis*. The three major arterial systems of the heart were judged with respect to narrowing. Narrowing was significant if it was 70% or more of the diameter of the artery. Because the angiographic films are a key diagnostic tool and are important in the decision about the appropriateness of bypass surgery, the quality of the arteriography was monitored and the amount of agreement was ascertained.

Table 7.8 presents the results for randomly selected films with two readings. One reading was that of the patient's clinical site and was used for therapeutic decisions. The angiographic film was then sent to another clinical site designated as a quality control site. The quality control site read the films blindly, that is, without knowledge of the clinical site's reading. From these readings, the amount of disease was classified as "none" (entirely normal), "zero-vessel disease but some disease," and one-, two-, and three-vessel disease.

We wish to study the amount of agreement. One possible measure of this is the proportion of the pairs of readings that are the same. This quantity is estimated by adding up the numbers

Table 7.8 Agreement with Respect to Number of Diseased Vessels

Quality Control	Clinical Site Reading					
Site Reading	Normal	Some	One	Two	Three	Total
Normal	13	8	1	0	0	22
Some	6	43	19	4	5	77
One	1	9	155	54	24	243
Two	0	2	18	162	68	250
Three	0	0	11	27	240	278
Total	20	62	204	247	337	870

on the diagonal of the table; those are the numbers where both the clinical and quality control sites read the same quantity. In such a situation, the contingency table will be square. Let r be the number of categories (in the table of this example, $r = 5$). The proportion of cases with agreement is given by

$$P_A = \frac{n_{11} + n_{22} + \cdots + n_{rr}}{n_{\cdot\cdot}} = \sum_{i=1}^{r} \frac{n_{ii}}{n_{\cdot\cdot}}$$

For this table, the proportion with agreement is given by $P_A = (13+43+155+162+240)/870 = 613/870 \doteq 0.7046$.

The proportion of agreement is limited because it is determined heavily by the proportions of people in the various categories. Consider, for example, a situation where each of two judges places 90% of the measurements in one category and 10% in the second category, such as in the following array:

81	9	90
9	1	10
90	10	100

Here there is no association whatsoever between the two measurements. In fact, the chi-square value is precisely zero by design; there is no more agreement between the patients than that expected by chance. Nevertheless, because both judges have a large proportion of the cases in the first category, in 82% of the cases there is agreement; that is, $P_A = 0.82$. We have a paradox: On the one hand, the agreement seems good (there is an agreement 82% of the time); on the other hand, the agreement is no more than can be expected by chance. To have a more useful measure of the amount of agreement, the *kappa statistic* was developed to adjust for the amount of agreement that one expects purely by chance.

If one knows the totals of the different rows and columns, the proportion of observations expected to agree by chance is given by the following equation:

$$P_C = \frac{n_{1\cdot}n_{\cdot 1} + \cdots + n_{r\cdot}n_{\cdot r}}{n_{\cdot\cdot}^2} = \sum_{i=1}^{r} \frac{n_{i\cdot}n_{\cdot i}}{n_{\cdot\cdot}^2}$$

For the angiography example, the proportion of agreement expected by chance is given by

$$P_C = \frac{22 \times 20 + 77 \times 62 + 243 \times 204 + 250 \times 247 + 278 \times 337}{870^2} \doteq 0.2777$$

The kappa statistic uses the fact that the best possible agreement is 1 and that, by chance, one expects an agreement P_C. A reasonable measure of the amount of agreement is the proportion of difference between 1 and P_C that can be accounted for by actual observed agreement. That is, kappa is the ratio of the agreement actually observed minus the agreement expected by chance, divided by 1 (which corresponds to perfect agreement), minus the agreement expected by chance:

$$\kappa = \frac{P_A - P_C}{1 - P_C}$$

For our example, the computed value of kappa is

$$\kappa = \frac{0.7046 - 0.2777}{1 - 0.2777} \doteq 0.59$$

The kappa statistic runs from $-P_C/(1 - P_C)$ to 1. If the agreement is totally by chance, the expected value is zero. Kappa is equal to 1 if and only if there is complete agreement between the two categorizations [Cohen, 1968; Fleiss, 1981].

Since the kappa statistic is generally used where it is clear that there will be statistically significant agreement, the real issue is the amount of agreement. κ is a measure of the amount of agreement. In our example, one can state that 59% of the difference between perfect agreement and the agreement expected by chance is accounted for by the agreement between the clinical and quality control reading sites.

Now that we have a parameter to measure the amount of agreement, we need to consider the effect of the sample size. For small samples, the estimation of κ will be quite variable; for larger samples it should be quite good. For relatively large samples, when there is no association, the variance of the estimate is estimated as follows:

$$\mathrm{var}_0(\kappa) = \frac{P_C + P_C^2 - \sum_{i=1}^{r}(n_{i\cdot}^2 n_{\cdot i} + n_{i\cdot} n_{\cdot i}^2)/n_{\cdot\cdot}^3}{n_{\cdot\cdot}(1 - P_C)^2}$$

The subscript on $\mathrm{var}_0(\kappa)$ indicates that it is the variance under the null hypothesis. The standard error of the estimate is the square root of this quantity. κ divided by the standard error is approximately a standard normal variable when there is no association between the quantities. This may be used as a statistical test for association in lieu of the chi-square test [Fleiss et al., 1969].

A more useful function of the general standard error is construction of a confidence interval for the true κ. A $100(1-\alpha)\%$ confidence interval for the population value of κ for large samples is given by

$$\left(\kappa - z_{1-\alpha/2}\sqrt{\mathrm{var}(\kappa)},\ \kappa + z_{1-\alpha/2}\sqrt{\mathrm{var}(\kappa)}\right)$$

The estimated standard error, allowing for association, is the square root of

$$\mathrm{var}(\kappa) = \frac{\sum_{i=1} \frac{n_{ii}}{n_{\cdot\cdot}}\left[1 - \left(\frac{n_{i\cdot} + n_{\cdot i}}{n_{\cdot\cdot}}\right)(1-\kappa)\right]^2 + \sum_{i \neq j}\sum \frac{n_{ij}}{n_{\cdot\cdot}}\left[\left(\frac{n_{\cdot i} + n_{j\cdot}}{n_{\cdot\cdot}}\right)(1-\kappa)\right]^2 - [\kappa - P_C(1-\kappa)]^2}{n_{\cdot\cdot}(1 - P_C)^2}$$

For our particular example, the estimated variance of κ is

$$\mathrm{var}(\kappa) = 0.000449$$

The standard error of κ is approximately 0.0212. The 95% confidence interval is

$$(0.57 - 1.96 \times 0.0212, 0.57 + 1.96 \times 0.0212) \doteq (0.55, 0.63)$$

A very comprehensive discussion of the use of κ in medical research can be found in Kraemer et al. [2002], and a discussion in the context of other ways to measure agreement is given by Nelson and Pepe [2000].

The kappa statistic has drawbacks. First, as indicated, the small sample variance is quite complicated. Second, while the statistic is supposed to adjust for marginal agreement is does not really do so (see, e.g., Agresti [2002, p. 453]). Third, κ ignores the ordering of the categories (see Maclure and Willett [1987]). Finally, it is difficult to embed κ in a statistical model: as, for example, a function of the odds ratio or correlation coefficient. Be sure to consider alternatives to kappa when measuring agreement; for example, the odds ratio and logistic regression as in Chapter 6 or the log-linear models discussed in the next section.

*7.5 LOG-LINEAR MODELS

For the first time we will examine statistical methods that deal with more than two variables at one time. Such methods are important for the following reasons: In one dimension, we have been able to summarize data with the normal distribution and its two parameters, the mean and the variance, or equivalently, the mean and the standard deviation. Even when the data did not appear normally distributed, we could get a feeling for our data by histograms and other graphical methods in one dimension. When we observe two numbers at the same time, or are working with two-dimensional data, we can plot the points and examine the data visually. (This is discussed further in Chapter 9. Even in the case of two variables, we shall see that it is useful to have models summarizing the data.) When we move to three variables, however, it is much harder to get a "feeling" for the data. Possibly, in three dimensions, we could construct visual methods of examining the data, although this would be difficult. With more than three variables, such physical plots cannot be obtained; although mathematicians may think of space and time as being a four-dimensional space, we, living in a three-dimensional world, cannot readily grasp what the points mean. In this case it becomes very important to simplify our understanding of the data by fitting a model to the data. *If* the model fits, it may summarize the complex situation very succinctly. In addition, the model may point out relationships that may reasonably be understood in a simple way. The fitting of probability models or distributions to many variables at one time is an important topic.

The models are necessarily mathematically complex; thus, the reader needs discipline and perseverance to work through and understand the methods. It is a very worthwhile task. Such methods are especially useful in the analysis of observational biomedical data. We now proceed to our first model for multiple variables, the log-linear model.

Before beginning the details of the actual model, we define some terms that we will be using. The models we investigate are for *multivariate categorical data*. We already know the meaning of *categorical data*: values of a variable or variables that put subjects into one of a finite number of categories. The term *multivariate* comes from the prefix *multi*-, meaning "many," and *variate*, referring to variables; the term refers to multiple variables at one time.

Definition 7.1. *Multivariate data* are data for which each observation consists of values for more than one random variable on each experimental unit. *Multivariate statistical analysis* consists of data analysis of multivariate data.

The majority of data collected are, in fact, multivariate data. If one measures systolic and diastolic blood pressure on each subject, there are two variables—thus, multivariate data. If we administer a questionnaire on the specifics of brushing teeth, flossing, and so on, the response of a person to each question is a separate variable, and thus one has multivariate data. Strictly speaking, some of the two-way contingency table data we have looked at are multivariate data since they cross-classify by two variables. On the other hand, tables that arose from looking at one quantity in different subgroups are not multivariate when the group was not observed on experimental units picked from a population but was part of a data collection or experimental procedure.

Additional terminology is included in the term *log-linear models*. We already have an idea of the meaning of a model. Let us consider the two terms *log* and *linear*. The logarithm was discussed in connection with the likelihood ratio chi-square statistics. (In this section, and indeed throughout this book, the logarithm will be to the base e.) Recall, briefly some of the properties of the logarithm. Of most importance to us is that the log of the product of terms is the sum of the individual logs. For example, if we have three numbers, a, b, and c (all positive), then

$$\ln(abc) = \ln a + \ln b + \ln c$$

Here, "ln" represents the *natural logarithm*, the log to the base e. Recall that by the definition of natural log, if one exponentiates the logarithm—that is, takes the number e to the power represented by the logarithm—one gets the original number back:

$$e^{\ln a} = a$$

Inexpensive hand calculators compute both the logarithm and the exponential of a number. If you are rusty with such manipulations, Problem 7.24 will give you practice in the use of logarithms and exponentials.

The second term we have used is the term *linear*. It is associated with a straight line or a linear relationship. For two variables x and y, y is a linear function of x if $y = a + bx$, where a and b are constants. For three variables, x, y, and z, z is a linear function of x and y if $z = a + bx + cy$, where a, b, and c are constant. In general, in a linear relationship, one *adds* a constant multiple for each of the variables involved. The linear models we use will look like the following: Let

$$g_{ij}^{IJ}$$

be the logarithm of the probability that an observation falls into the ijth cell in the two-dimensional contingency table. Let there be I rows and J columns. One possible model would be

$$g_{ij}^{IJ} = u + u_i^I + u_j^J$$

(For more detail on why the term *linear* is used for such models, see Note 7.4.)

We first consider the case of two-way tables. Suppose that we want to fit a model for independence. We know that independence in terms of the cell probabilities π_{ij} is equivalent to the following equation:

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

If we take logarithms of this equation and use the notation g_{ij} for the natural log of the cell probability, the following results:

$$g_{ij} = \ln \pi_{ij} = \ln \pi_{i\cdot} + \ln \pi_{\cdot j}$$

When we denote the natural logs of $\pi_{i\cdot}$ and $\pi_{\cdot j}$ by the quantities h_i^I and h_j^J, we have

$$g_{ij} = h_i^I + h_j^J$$

The quantities h_i^I and h_j^J are not all independent. They come from the marginal probabilities for the I row variables and the J column variables. For example, the h_i^I's satisfy the equation

$$e^{h_1^I} + e^{h_2^I} + \cdots + e^{h_I^I} = 1$$

This equation is rather awkward and unwieldy to work with; in particular, given $I - 1$ of the h_i's, determination of the other coefficient takes a bit of work. It is possible to choose a different normalization of the parameters if we add a constant. Rewrite the equation above as follows:

$$g_{ij} = \left(\sum_{i'=1}^{I} \frac{h_{i'}^I}{I}\right) + \left(\sum_{j'=1}^{J} \frac{h_{j'}^J}{J}\right) + \left(h_i^I - \sum_{i'=1}^{I} \frac{h_{i'}^I}{I}\right) + \left(h_j^J - \sum_{j'=1}^{J} \frac{h_{j'}^J}{J}\right)$$

The two quantities in parentheses farthest to the right both add to zero when we sum over the indices i and j, respectively. In fact, that is why those terms were added and subtracted. Thus, we can rewrite the equation for g_{ij} as follows:

$$g_{ij} = u + u_i^I + u_j^J, \qquad i = 1, \dots, I, \quad j = 1, \dots, J$$

where

$$\sum_{i=1}^{I} u_i^I = 0, \ \sum_{j=1}^{J} u_j^J = 0$$

It is easier to work with this normalization. Note that this is a linear model for the log of the cell probability π_{ij}; that is, this is a log-linear model.

Recall that estimates for the $\pi_{i\cdot}$ and $\pi_{\cdot j}$ were $n_{i\cdot}/n_{\cdot\cdot}$ and $n_{\cdot j}/n_{\cdot\cdot}$, respectively. If one follows through all of the mathematics involved, estimates for the parameters in the log-linear model result. At this point, we shall slightly abuse our notation by using the same notation for both the population parameter values and the estimated parameter values from the sample at hand. The estimates are

$$u = \frac{1}{I}\sum_{i=1}^{I} \ln \frac{n_{i\cdot}}{n_{\cdot\cdot}} + \frac{1}{J}\sum_{j=1}^{J} \ln \frac{n_{\cdot j}}{n_{\cdot\cdot}}$$

$$u_i^I = \ln \frac{n_{i\cdot}}{n_{\cdot\cdot}} - \frac{1}{I}\sum_{i'=1}^{I} \ln \frac{n_{i'\cdot}}{n_{\cdot\cdot}}$$

$$u_j^J = \ln \frac{n_{\cdot j}}{n_{\cdot\cdot}} - \frac{1}{J}\sum_{j'=1}^{I} \ln \frac{n_{\cdot j'}}{n_{\cdot\cdot}}$$

From these estimates we get fitted values for the number of observations in each cell. This is done as follows: By inserting the estimated parameters from the log-linear model and then taking the exponential, we have an estimate of the probability that an observation falls into the ijth cell. Multiplying this by $n_{\cdot\cdot}$, we have an estimate of the number of observations we should see in the cell if the model is correct. In this particular case, the fitted value for the ijth cell turns out to be the expected value from the chi-square test presented earlier in this chapter, that is, $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$.

Let us illustrate these complex formulas by finding the estimates for one of the examples above.

Example 7.1. (*continued*) We know that for the 2 × 4 table, we have the following values:

$$n_{\cdot 1} = 18, \quad n_{\cdot 2} = 31, \quad n_{\cdot 3} = 16, \quad n_{\cdot 4} = 72, \quad n_{1\cdot} = 69, \quad n_{2\cdot} = 68, \quad n_{\cdot\cdot} = 137$$

$$\ln(n_{1\cdot}/n_{\cdot\cdot}) \doteq -0.6859, \qquad \ln(n_{2\cdot}/n_{\cdot\cdot}) \doteq -0.7005$$

$$\ln(n_{\cdot 1}/n_{\cdot\cdot}) \doteq -2.0296, \qquad \ln(n_{\cdot 2}/n_{\cdot\cdot}) \doteq -1.4860$$

$$\ln(n_{\cdot 3}/n_{\cdot\cdot}) \doteq -2.1474, \qquad \ln(n_{\cdot 4}/n_{\cdot\cdot}) \doteq -0.6433$$

With these numbers, we may compute the parameters for the log-linear model. They are

$$\begin{aligned} u &\doteq \frac{-0.6859 - 0.7005}{2} + \frac{-2.0296 - 1.4860 - 2.1474 - 0.6433}{4} \\ &\doteq -0.6932 - 1.5766 = -2.2698 \end{aligned}$$

$$\begin{array}{ll} & u_1^J \doteq -2.0296 - (-1.5766) \doteq -0.4530 \\ u_1^I \doteq -0.6859 - (-0.6932) \doteq 0.0073 & u_2^J \doteq -1.4860 - (-1.5766) \doteq 0.0906 \\ u_2^I \doteq -0.7004 - (-0.6932) \doteq -0.0073 & u_3^J \doteq -2.1474 - (-1.5766) \doteq -0.5708 \\ & u_4^J \doteq -0.6433 - (-1.5766) \doteq 0.9333 \end{array}$$

The larger the value of the coefficient, the larger will be the cell probability. For example, looking at the two values indexed by i, the second state having a minus sign will lead to a slightly smaller contribution to the cell probability than the term with the plus sign. (This is also clear from the marginal probabilities, which are 68/137 and 69/137.) The small magnitude of the term means that the difference between the two I state values has very little effect on the cell probability. We see that of all the contributions for the j variable values, $j = 4$ has the biggest effect, 1 and 3 have fairly large effects (tending to make the cell probability small), while 2 is intermediate.

The chi-square goodness of fit and the likelihood ratio chi-square statistics that may be applied to this setting are

$$X^2 = \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}}$$

$$\text{LRX}^2 = 2 \sum \left(\text{observed} \ln \frac{\text{observed}}{\text{fitted}} \right)$$

Finally, if the model for independence does not hold, we may add more parameters. We can find a log-linear model that will fit any possible pattern of cell probabilities. The equation for the log of the cell probabilities is given by the following:

$$g_{ij} = u + u_i^I + u_j^J + u_{ij}^{IJ}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J$$

where

$$\sum_{i=1}^{I} u_i^I = 0, \qquad \sum_{j=1}^{J} u_j^J = 0, \qquad \sum_{i=1}^{I} u_{ij}^{IJ} = 0, \qquad \sum_{j=1}^{J} u_{ij}^{IJ} = 0$$

It seems rather paradoxical, or at least needlessly confusing, to take a value indexed by i and j and to set it equal to the sum of four values, including some indexed by i and j; the right-hand side is much more complex than the left-hand side. The reason for doing this is that, usually, the full (or saturated) model, which can give any possible pattern of cell probabilities, is not desirable. It is hoped during the modeling effort that the data will allow a simpler model, which would allow a simpler interpretation of the data. In the case at hand, we examine the possibility of the simpler interpretation that the two variables are independent. If they are not, the particular model is not too useful.

Note two properties of the fitted values. First, in order to fit the independence model, where each term depends on at most one factor or one variable, we only needed to know the marginal values of the frequencies, the $n_{i\cdot}$ and $n_{\cdot j}$. We did not need to know the complete distribution of the frequencies to find our fitted values. Second, when we had fit values to the frequency table, the fitted values summed to the marginal value used in the estimation; that is, if we sum across i or j, the sum of the expected values is equal to the sum actually observed.

At this point it seems that we have needlessly confused a relatively easy matter: the analysis of two-way contingency tables. If only two-way contingency tables were involved, this would be a telling criticism; however, the strength of log-linear models appears when we have more than two cross-classified categorical variables. We shall now discuss the situation for three cross-classified categorical variables. The analyses may be extended to any number of variables, but such extensions are not done in this book.

Suppose that the three variables are labeled X, Y, and Z, where the index i is used for the X variable, j for the Y variable, and k for the Z variable. (This is to say that X will take values $1, \dots, I$, Y will take on $1, \dots, J$, and so on.) The methods of this section are illustrated by the following example.

Example 7.5. The study of Weiner et al. [1979] is used in this example. The study involves exercise treadmill tests for men and women. Among men with chest pain thought probably to be angina, a three-way classification of the data is as follows: One variable looks at the resting electrocardiogram and tells whether or not certain parts of the electrocardiogram (the ST- and T-waves) are normal or abnormal. Thus, $J = 2$. A second variable considers whether or not the exercise test was positive or negative ($I = 2$). A positive exercise test shows evidence of an ischemic response (i.e., lack of appropriate oxygen to the heart muscles for the effort being exerted). A positive test is thought to be an indicator of coronary artery disease. The third variable was an evaluation of the coronary artery disease as determined by coronary arteriography. The disease is classified as normal or minimal disease, called zero-vessel disease, one-vessel disease, and multiple-vessel disease ($K = 3$). The data are presented in Table 7.9.

The most general log-linear model for the three factors is given by the following extension of the two-factor work:

$$g_{ijk} = u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ} + u_{ik}^{IK} + u_{jk}^{JK} + u_{ijk}^{IJK}$$

where

$$\sum_{i=1}^{I} u_i^I = \sum_{j=1}^{J} u_j^J = \sum_{k=1}^{K} u_k^K = 0$$

$$\sum_{i=1}^{I} u_{ij}^{IJ} = \sum_{j=1}^{J} u_{ij}^{IJ} = \sum_{i=1}^{I} u_{ik}^{IK} = \sum_{k=1}^{K} u_{ik}^{IK} = \sum_{j=1}^{J} u_{jk}^{JK} = \sum_{k=1}^{K} u_{jk}^{JK} = 0$$

$$\sum_{i=1}^{I} u_{ijk}^{IJK} = \sum_{j=1}^{J} u_{ijk}^{IJK} = \sum_{k=1}^{K} u_{ijk}^{IJK} = 0$$

Table 7.9 Exercise Test Data

		Number of Vessels Diseased (K)		
Exercise Test Response (I)	Resting Electrocardiogram ST- and T-Waves (J)	0 ($k = 1$)	1 ($k = 2$)	2 or 3 ($k = 3$)
+	Normal ($j = 1$)	30	64	147
($i = 1$)	Abnormal ($j = 2$)	17	22	80
−	Normal ($j = 1$)	118	46	38
($i = 2$)	Abnormal ($j = 2$)	14	7	11

Source: Weiner et al. [1979].

In other words, there is a u term for every possible combination of the variables, including no variables at all. For each term involving one or more variables, if we sum over any one variable, the sum is equal to zero. The term involving I, J, and K is called a *three-factor term*, or a *second-order interaction term*; in general, if a coefficient involves M variables, it is called an *M-factor term* or an $(M-1)$th-*order interaction term*.

With this notation we may now formulate a variety of simpler models for our three-way contingency table. For example, the model might be any one of the following simpler models:

$$H_1 : g_{ijk} = u + u_i^I + u_j^J + u_k^K$$

$$H_2 : g_{ijk} = u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ}$$

$$H_3 : g_{ijk} = u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ} + u_{ik}^{IK} + u_{jk}^{JK}$$

The notation has become so formidable that it is useful to introduce a shorthand notation for the hypotheses. One or more capitalized indices contained in brackets will indicate a hypothesis where the terms involving that particular set of indices as well as any terms involving subsets of the indices are to be included in the model. Any terms not specified in this form are assumed not to be in the model. For example,

$$[IJ] \longrightarrow u + u_i^I + u_j^J + u_{ij}^{IJ}$$

$$[K] \longrightarrow u + u_k^K$$

$$[IJK] \longrightarrow u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ} + u_{ik}^{IK} + u_{jk}^{JK} + u_{ijk}^{IJK}$$

The formulation of the three hypotheses given above in this notation would be simplified as follows:

$$H_1 : [I][J][K]$$

$$H_2 : [IJ][K]$$

$$H_3 : [IJ][IK][JK]$$

This notation describes a *hierarchical hypothesis*; that is, if we have two factor terms containing, say, variables I and J, we also have the one-factor terms for the same variables. The hypothesis would not be written $[IJ][I][J]$, for example, because the last two parts would be redundant, as already implied by the first. Using this bracket notation for the three-factor model, there are eight possible hypotheses of interest. All except the most complex one have a simple interpretation in terms of the probability relationships among the factors X, Y, and Z. This is given in Table 7.10.

Hypotheses 5, 6, and 7 are of particular interest. Take, for example, hypothesis 5. This hypothesis states that if you take into account the X variable, there is no association between Y and Z. In particular, if one only looks at the two-way table of Y and Z, an association may be seen, because in fact they are associated. However, if hypothesis 5 holds, one could then conclude that the association is due to interaction with the variable X and could be "explained away" by taking into account the values of X.

There is a relationship between hypotheses involving the bracket notation and the corresponding tables that one gets from the higher-dimensional contingency table. For example, consider the term $[IJ]$. This is related to the contingency table one gets by summing over K (i.e., over the Z variable). In general, a contingency table that results from summing over the cells for one or more variables in a higher-dimensional contingency table is called a *marginal table*. Very simple examples of marginal tables are the marginal total column and the marginal total row along the bottom of the two-way table.

Table 7.10 Three-Factor Hypotheses and their Interpretation

Hypothesis	Meaning in Words	Hypothesis Restated in Terms of the π_{ijk}'s
1. $[I][J][K]$	X, Y, and Z are independent	$\pi_{ijk} = \pi_{i\cdot\cdot}\pi_{\cdot j\cdot}\pi_{\cdot\cdot k}$
2. $[IJ][K]$	Z is independent of X and Y	$\pi_{ijk} = \pi_{ij\cdot}\pi_{\cdot\cdot k}$
3. $[IK][J]$	Y is independent of X and Z	$\pi_{ijk} = \pi_{i\cdot k}\pi_{\cdot j\cdot}$
4. $[I][JK]$	X is independent of Y and Z	$\pi_{ijk} = \pi_{i\cdot\cdot}\pi_{\cdot jk}$
5. $[IJ][IK]$	For X known, Y and Z are independent; that is, Y and Z are conditionally independent given X	$\pi_{ijk} = \pi_{ij.}\pi_{i.k}/\pi_{i\cdot\cdot}$
6. $[IJ][JK]$	X and Z are conditionally independent given Y	$\pi_{ijk} = \pi_{ij.}\pi_{.jk}/\pi_{.j.}$
7. $[IK][JK]$	X and Y are conditionally independent given Z	$\pi_{ijk} = \pi_{i.k}\pi_{.jk}/\pi_{\cdot\cdot k}$
8. $[IJ][IK][JK]$	No three-factor interaction	No simple form

Using the idea of marginal tables, we can discuss some properties of fits of the various hierarchical hypotheses for log-linear models. Three facts are important:

1. The fit is estimated using only the marginal tables associated with the bracket terms that state the hypothesis. For example, consider hypothesis 1, the independence of the X, Y, and Z variables. To compute the estimated fit, one only needs the one-dimensional frequency counts for the X, Y, and Z variables individually and does not need to know the joint relationship between them.
2. Suppose that one looks at the fitted estimates for the frequencies and sums the *fitted* values to give marginal tables. The marginal sum for the fit is equal to the marginal table for the actual data set when the marginal table is involved in the fitting.
3. The chi-square and likelihood ratio chi-square tests discussed above using the observed and fitted values still hold.

We consider fitting hypothesis 5 to the data of Example 7.5. The hypothesis stated that if one knows the response to the maximal treadmill test, the resting electrocardiogram ST- and T-wave abnormalities are independent of the number of vessels diseased. The observed frequencies and the fitted frequencies, as well as the values of the u-parameters for this model, are given in Table 7.11.

The relationship between the fitted parameter values and the expected, or fitted, number of observations in a cell is given by the following equations:

$$\widehat{\pi}_{ijk} = e^{u+u_i^I+u_j^J+u_k^K+u_{ij}^{IJ}+u_{ik}^{IK}}$$

The fitted value $= n\ldots\widehat{\pi}_{ijk}$, where $n\ldots$ is the total number of observations.
For these data, we compute the right-hand side of the first equation for the (1,1,1) cell. In this case,

$$\begin{aligned}\widehat{\pi}_{111} &= \exp(-2.885 + 0.321 + 0.637 - 0.046 - 0.284 - 0.680)\\ &= e^{-2.937} \doteq 0.053\end{aligned}$$

$$\text{fitted value} \doteq 594 \times 0.053 \doteq 31.48$$

Table 7.11 Fitted Model for the Hypothesis That the Resting Electrocardiogram ST- and T-Wave (Normal or Abnormal) Is Independent of the Number of Vessels Diseased (0, 1, and 2–3) Conditionally upon Knowing the Exercise Response (+ or −)

Cell (i, j, k)	Observed	Fitted	u-Parameters
(1,1,1)	30	31.46	$u = -2.885$
(1,1,2)	64	57.57	$u_1^I = -u_2^I = 0.321$
(1,1,3)	147	151.97	$u_1^J = -u_2^J = 0.637$
(1,2,1)	17	15.54	$u_1^K = -0.046, u_2^K = -0.200$
(1,2,2)	22	28.43	$u_3^K = 0.246$
(1,2,3)	80	75.04	$u_{1,1}^{IJ} = -0.284, u_{1,2}^{IJ} = 0.284$
(2,1,1)	118	113.95	$u_{2,1}^{IJ} = 0.284, u_{2,2}^{IJ} = -0.284$
(2,1,2)	46	45.75	$u_{1,1}^{IK} = -0.680, u_{1,2}^{IK} = 0.078$
(2,1,3)	38	42.30	$u_{1,3}^{IK} = 0.602$
(2,2,1)	14	18.05	$u_{2,1}^{IK} = 0.680, u_{2,2}^{IK} = -0.078$
(2,2,2)	7	7.25	$u_{2,3}^{IK} = -0.602$
(2,2,3)	11	6.70	

where exp(argument) is equal to the number e raised to a power equal to the argument. The computed value of 31.48 differs slightly from the tabulated value, because the tabulated value came from computer output that carried more accuracy than the accuracy used in this computation.

We may test whether the hypothesis is a reasonable fit by computing the chi-square value under this hypothesis. The likelihood ratio chi-square value is computed as follows:

$$\mathrm{LRX}^2 = 2(30\ln\frac{30}{31.46} + \cdots + 11\ln\frac{11}{6.70}) \doteq 6.86$$

To assess the statistical significance we need the degrees of freedom to examine the chi-square value. For the log-linear model the degrees of freedom is given by the following rule:

Rule 1. The chi-square statistic for model fit of a log-linear model has degrees of freedom equal to the total number of cells in the table $(I \times J \times K)$ minus the number of independent parameters fitted. By *independent parameters* we mean the following: The number of parameters fitted for the X variable is $I - 1$ since the u_i^I terms sum to zero. For each of the possible terms in the model, the number of independent parameters is given in Table 7.12.

For the particular model at hand, the number of independent parameters fitted is the sum of the last column in Table 7.13. There are 12 cells in the table, so that the number of degrees of freedom is 12 − 8, or 4. The p-value for a chi-square of 6.86 for four degrees of freedom is 0.14, so that we cannot reject the hypothesis that this particular model fits the data.

We are now faced with a new consideration. Just because this model fits the data, there may be other models that fit the data as well, including some simpler model. In general, one would like as simple a model as possible (Occam's razor); however, models with more parameters generally give a better fit. In particular, a simpler model may have a p-value much closer to the significance level that one is using. For example, if one model has a p of 0.06 and is simple, and a slightly more complicated model has a p of 0.78, which is to be preferred? If the sample size is small, the p of 0.06 may correspond to estimated cell values that differ considerably from the actual values. For a very large sample, the fit may be excellent. There is no hard-and-fast rule in the trade-off between the simplicity of the model and the goodness of the fit. To understand the data, we are happy with the simple model that fits fairly well, although presumably, it is not precisely the probability model that would fit the entirety of the population values. Here we would hope for considerable scientific understanding from the simple model.

Table 7.12 Degrees of Freedom for Log-Linear Model Chi-Square

Term	Number of Parameters
u	1
u_i^I	$I - 1$
u_j^J	$J - 1$
u_k^K	$K - 1$
u_{ij}^{IJ}	$(I - 1)(J - 1)$
u_{ik}^{IK}	$(I - 1)(K - 1)$
u_{jk}^{JK}	$(J - 1)(K - 1)$
u_{ijk}^{IJK}	$(I - 1)(J - 1)(K - 1)$

Table 7.13 Parameters for Example 7.5

	Number of Parameters	
Model Terms	General	Example 7.5
u	1	1
u_i^I	$I - 1$	1
u_j^J	$J - 1$	1
u_k^K	$K - 1$	2
u_{ij}^{IJ}	$(I - 1)(J - 1)$	1
u_{ik}^{IK}	$(I - 1)(K - 1)$	2

Table 7.14 Chi-Square Goodness-of-Fit Statistics for Example 7.5 Data

Model	d.f.	LRX^2	p-Value	X^2
$[I][J][K]$	7	184.21	< 0.0001	192.35
[IJ]$[K]$	6	154.35	< 0.0001	149.08
[IK]$[J]$	5	36.71	< 0.0001	34.09
[I]$[JK]$	5	168.05	< 0.0001	160.35
[IJ]$[IK]$	4	6.86	0.14	7.13
[IJ]$[JK]$	4	138.19	< 0.0001	132.30
[IK]$[JK]$	3	20.56	0.0001	21.84
[IJ]$[IK][JK]$	2	2.96	0.23	3.03

For this example, Table 7.14 shows for each of the eight possible models the degrees of freedom (d.f.), the LRX^2 value (with its corresponding p-value for reference), and the "usual" goodness-of-fit chi-square value. We see that there are only two possible models if we are to simplify at all rather than using the entire data set as representative. They are the model fit above and the model that contains each of the three two-factor interactions. The model fit above is simpler, while the other model below has a larger p-value, possibly indicating a better fit. One way of approaching this is through what are called *nested hypotheses*.

Definition 7.2. One hypothesis is *nested* within another if it is the special case of the other hypothesis. That is, whenever the nested hypothesis holds it necessarily implies that the hypothesis it is nested in also holds.

If nested hypotheses are considered, one takes the difference between the likelihood ratio chi-square statistic for the more restrictive hypothesis, minus the likelihood ratio chi-square statistic for the more general hypothesis. This difference will itself be a chi-square statistic if the special case holds. The degrees of freedom of the difference is equal to the difference of freedom for the two hypotheses. In this case, the chi-square statistic for the difference is $6.86 - 2.96 = 3.90$. The degrees of freedom are $4 - 2 = 2$. This corresponds to a p-value of more than 0.10. At the 5% significance level, there is marginal evidence that the more general hypothesis does fit the data better than the restrictive hypothesis. In this case, however, because of the greater simplicity of the restrictive hypothesis, one might choose it to fit the data. Once again, there is no hard and fast answer to the payoff between fit of the data and simplicity of interpretation of a hypothesis.

This material is an introduction to log-linear models. There are many extensions, some of which are mentioned briefly in the Notes at the end of the chapter. An excellent introduction to log-linear models is given in Fienberg [1977]. Other elementary books on log-linear models are those by Everitt [1992] and Reynolds [1977]. A more advanced and thorough treatment is given by Haberman [1978, 1979]. A text touching on this subject and many others is Bishop et al. [1975].

NOTES

7.1 Testing Independence in Model 1 and Model 2 Tables

This note refers to Section 7.2.

1. *Model 1.* The usual null hypothesis is that the results are statistically independent. That is (assuming row variable $= i$ and column variable $= j$):

$$P[i \text{ and } j] = P[i]P[j]$$

The probability on the left-hand side of the equation is π_{ij}. From Section 7.2, the marginal probabilities are found to be

$$\pi_{i\cdot} = \sum_{j=1}^{c} \pi_{ij} \quad \text{and} \quad \pi_{\cdot j} = \sum_{i=1}^{r} \pi_{ij}$$

The null hypothesis of statistical independence of the variables is

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

Consider how one might estimate these probabilities under two circumstances:

a. Without assuming the variables are independent.

b. Assuming the variables are independent.

In the first instance we are in a binomial situation. Let a success be the occurrence of the ijth pair. Let

$$n_{\cdot\cdot} = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

The binomial estimate for π_{ij} is the number of successes divided by the number of trials:

$$p_{ij} = \frac{n_{ij}}{n_{\cdot\cdot}}$$

If we assume independence, the natural approach is to estimate $\pi_{i\cdot}$ and $\pi_{\cdot j}$. But the occurrence of state i for the row variable is also a binomial event. The estimate of $\pi_{i\cdot}$ is the number of occurrences of state i for the row variable ($n_{i\cdot}$) divided by the sample size ($n_{\cdot\cdot}$). Thus,

$$p_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}}$$

Similarly, $\pi_{\cdot j}$ is estimated by

$$p_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}}$$

Under the hypothesis of statistical independence, the estimate of $\pi_{i\cdot}\pi_{\cdot j} = \pi_{ij}$ is

$$\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}^2}$$

The chi-square test will involve comparing estimates of the expected number of observations with and without assuming independence. With independence, we expect to observe $n_{\cdot\cdot}\pi_{ij}$ entries in the ijth cell. This is estimated by

$$n_{\cdot\cdot}p_{i\cdot}p_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

2. *Model 2*. Suppose that the row variable identifies the population. The null hypothesis is that all r populations have the same probabilities of taking on each value of the column variable. That is, for any two rows, denoted by i and i', say, and all j,

$$H_0 : \pi_{ij} = \pi_{i'j}$$

As in the first part above, we want to estimate these probabilities in two cases:

a. Without assuming anything about the probabilities.

b. Under H_0, that is, assuming that each population has the same distribution of the column variable.

Under (a), if no assumptions are made, π_{ij} is the probability of obtaining state j for the column variable in the $n_{i\cdot}$ trials from the ith population. Again the binomial estimate holds:

$$p_{ij} = \frac{n_{ij}}{n_{i\cdot}}$$

If the null hypothesis holds, we may "pool" all our $n_{\cdot\cdot}$ trials to get a more accurate estimate of the probabilities. Then the proportion of times the column variable takes on state j is

$$p_j = \frac{n_{\cdot j}}{n_{\cdot\cdot\cdot}}$$

As in the first part, let us calculate the numbers we expect in the cells under (a) and (b). If (a) holds, the expected number of successes in the $n_{i\cdot}$ trials of the ith population is $n_{i\cdot}\pi_{ij}$. We estimate this by

$$n_{i\cdot}(\frac{n_{ij}}{n_{i\cdot}}) = n_{ij}$$

Under the null hypothesis, the expected number $n_{i\cdot}\pi_{ij}$ is estimated by

$$n_{i\cdot}p_j = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

In summary, under either model 1 or model 2, the null hypothesis is reasonably tested by comparing n_{ij} with $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$.

7.2 Measures of Association in Contingency Tables

Suppose that we reject the null hypothesis of no association between the row and column categories in a contingency table. It is useful then to have a measure of the degree of association. In a series of papers, Goodman and Kruskal [1979] argue that no single measure of association for contingency tables is best for all purposes. Measures must be chosen to help with the problem at hand. Among the measures they discuss are the following:

1. *Measure* λ_C. Call the row variable or row categorization R and the column variable or column categorization C. Suppose that we wish to use the value of R to predict the value of C. The measure λ_C is an estimate of the proportion of the errors made in classification if we do not know R that can be eliminated by knowing R before making a prediction. From the data, λ_C is given by

$$\lambda_C = \frac{\left(\sum_{i=1}^{r} \max_j n_{ij}\right) - \max_j n_{\cdot j}}{n_{\cdot\cdot} - \max_j n_{\cdot j}}$$

λ_R is defined analogously.

2. *Symmetric measure* λ. λ_C does not treat the row and column classifications symmetrically. A symmetric measure may be found by assuming that the chances are 1/2 and 1/2 of needing to predict the row and column variables, respectively. The proportion of the errors in classification that may be reduced by knowing the other (row or column variable) when predicting is estimated by λ:

$$\lambda = \frac{\left(\sum_{i=1}^{r} \max_j n_{ij}\right) + \left(\sum_{j=1}^{c} \max_i n_{ij}\right) - \max_i n_{i\cdot} - \max_j n_{\cdot j}}{2n_{\cdot\cdot} - (\max_i n_{i\cdot} + \max_j n_{\cdot j})}$$

3. *Measure* γ *for ordered categories*. In many applications of contingency tables the categories have a natural order: for example, last grade in school, age categories, number of weeks hospitalized. Suppose that the orderings of the variables correspond to the indices i and j for the rows and columns. The γ measure is the difference in the proportion of the time that the two measures have the same ordering minus the proportion of the time that they have the opposite ordering, when there are no ties. Suppose that the indices for the two observations are i, j and i, j. The indices have the same ordering if

$$(1) i < i \text{ and } j < j \quad \text{or} \quad (2) i > i \text{ and } j > j$$

They have the opposite ordering if

$$(1)\ i < \boldsymbol{i} \text{ and } j > \boldsymbol{j} \quad \text{or} \quad (2)\ i > \boldsymbol{i} \text{ and } j < \boldsymbol{j}$$

There are ties if $i = \boldsymbol{i}$ and/or $j = \boldsymbol{j}$. The index is

$$\gamma = \frac{2S - 1 + T}{1 - T}$$

where

$$S = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}\sum_{\boldsymbol{i}>i}\sum_{\boldsymbol{j}>j} n_{\boldsymbol{ij}}}{n_{..}^2}$$

and

$$T = \frac{\sum_{i=1}^{r}\left(\sum_{j=1}^{c} n_{ij}\right)^2 + \sum_{j=1}^{c}\left(\sum_{i=1}^{r} n_{ij}\right)^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2}{n_{..}^2}$$

4. *Karl Pearson's contingency coefficient*, C. Since the chi-square statistic (X^2) is based on the square of the difference between the values observed in the contingency table and the values estimated, if association does not hold, it is reasonable to base a measure of association on X^2. However, chi-square increases as the sample size increases. One would like a measure of association that estimated a property of the total population. For this reason, $X^2/n_{..}$ is used in the next three measures. Karl Pearson proposed the measure C.

$$C = \sqrt{\frac{X^2/n_{..}}{1 + X^2/n_{..}}}$$

5. *Cramer's* V. Harold Cramer proposed a statistic with values between 0 and 1. The coefficient can actually attain both values.

$$V = \sqrt{\frac{X^2/n_{..}}{\text{minimum}(r-1, c-1)}}$$

6. *Tshuprow's* T, *and the* Φ^2 *coefficient*. The two final coefficients based on X^2 are

$$T = \sqrt{\frac{X^2/n_{..}}{\sqrt{(r-1)(c-1)}}} \quad \text{and} \quad \Phi = \sqrt{X^2/n_{..}}$$

We compute these measures of association for two contingency tables. The first table comes from the Robertson [1975] seat belt paper discussed in the text. The data are taken for 1974 cars with the interlock system. They relate age to seat belt use. The data and the column percents are given in Table 7.15. Although the chi-square value is 14.06 with $p = 0.007$, we can see from the column percentages that the relationship is weak. The coefficients of association are

$$\begin{array}{llll}\lambda_C = 0, & \lambda = 0.003, & C = 0.08, & T = 0.04 \\ \lambda_R = 0.006, & \gamma = -0.03, & V = 0.06, & \Phi = 0.08\end{array}$$

Table 7.15 Seat Belt Data by Age

	Age (Years)			Column Percents (Age)		
Belt Use	<30	30–49	≥ 50	< 30	30–49	≥ 50
Lap and shoulder	206	580	213	45	50	45
Lap only	36	125	65	8	11	14
None	213	459	192	47	39	41
				100	100	100

In general, all these coefficients lie between -1 or 0, and $+1$. They are zero if the variables are not associated at all. These values are small, indicating little association.

Consider the following data from Weiner et al. [1979], relating clinical diagnosis of chest pain to the results of angiographic examination of the coronary arteries:

	Frequency (Vessels Diseased)			**Row Percents (Vessels Diseased)**			
Chest Pain	**0**	**1**	**2 or 3**	**0**	**1**	**2 or 3**	**Total**
Definite angina	66	135	419	11	22	68	101
Probable angina	179	139	276	30	23	46	99
Nonischemic	197	39	15	78	16	6	100

The chi-square statistic is 418.48 with a p-value of effectively zero. Note that those with definite angina were very likely (89%) to have disease, and even the probability of having multivessel disease was 68%. Chest pain thought to be nonischemic was associated with "no disease" 78% of the time. Thus, there is a strong relationship. The measures of association are

$$\begin{array}{llll} \lambda_C = 0.24, & \lambda = \ \ 0.20, & C = 0.47, & T = 0.38 \\ \lambda_R = 0.16, & \gamma = -0.64, & V = 0.38, & \Phi = 0.53 \end{array}$$

More information on these measures of association and other potentially useful measures is available in Reynolds [1977] and in Goodman and Kruskal [1979].

7.3 *Testing for Symmetry in a Contingency Table*

In a square table, one sometimes wants to test the table for symmetry. For example, when examining two alternative means of classification, one may be interested not only in the amount of agreement (κ), but also in seeing that the pattern of misclassification is the same. In this case, estimate the expected value in the ijth cell by $(n_{ij} + n_{ji})/2$. The usual chi-square value is appropriate with $r(r-1)/2$ degrees of freedom, where r is the number of rows (and columns). See van Belle and Cornell [1971].

7.4 *Use of the Term Linear in Log-Linear Models*

Linear equations are equations of the form $y = c + a_1X_1 + a_2X_2 + \cdots + a_nX_n$ for some variables $X_1, \ldots, X_n$ and constants c and $a_1, \ldots, a_n$. The log-linear model equations can be put into this form. For concreteness, consider the model $[IJ][K]$, where $i = 1, 2$, $j = 1, 2$, and $k = 1, 2$. Define new variables as follows:

$$X_1 = \begin{cases} 1 & \text{if } i = 1, \\ 0 & \text{if } i = 2; \end{cases} \quad X_2 = \begin{cases} 1 & \text{if } i = 2, \\ 0 & \text{if } i = 1; \end{cases} \quad X_3 = \begin{cases} 1 & \text{if } j = 1, \\ 0 & \text{if } j = 2; \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if } j = 2, \\ 0 & \text{if } j = 1 \end{cases} \quad X_5 = \begin{cases} 1 & \text{if } k = 1, \\ 0 & \text{if } k = 2 \end{cases} \quad X_6 = \begin{cases} 1 & \text{if } k = 2, \\ 0 & \text{if } k = 1, \end{cases}$$

$$X_7 = \begin{cases} 1 & \text{if } i = 1, j = 1, \\ 0 & \text{otherwise;} \end{cases} \quad X_8 = \begin{cases} 1 & \text{if } i = 1, j = 2, \\ 0 & \text{otherwise;} \end{cases}$$

$$X_9 = \begin{cases} 1 & \text{if } i = 2, j = 1, \\ 0 & \text{otherwise} \end{cases} \quad X_{10} = \begin{cases} 1 & \text{if } i = 2, j = 2, \\ 0 & \text{otherwise} \end{cases}$$

Then the model is

$$\log \pi_{ijk} = u + u_1^I X_1 + u_2^I X_2 + u_1^J X_3 + u_2^J X_4 + u_1^K X_5 + u_2^K X_6$$
$$+ u_{1,1}^{IJ} X_7 + u_{1,2}^{IJ} X_8 + u_{2,1}^{IJ} X_9 + u_{2,2}^{IJ} X_{10}$$

Thus the log-linear model is a linear equation of the same form as $y = c + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$. We discuss such equations in Chapter 11. Variables created to pick out a certain state (e.g., $i = 2$) by taking the value 1 when the state occurs, and taking the value 0 otherwise, are called *indicator* or *dummy variables*.

7.5 Variables of Constant Probability in Log-Linear Models

Consider the three-factor X, Y, and Z log-linear model. Suppose that Z terms are entirely "omitted" from the model, for example, $[IJ]$ or

$$\log \pi_{ijk} = u + u_i^I + u_j^J + u_{ij}^{IJ}$$

The model then fits the situation where Z is uniform on its state; that is,

$$P[Z = k] = \frac{1}{k}, \qquad k = 1, \ldots, K$$

7.6 Log-Linear Models with Zero Cell Entries

Zero values in the contingency tables used for log-linear models are of two types. Some arise as *sampling zeros* (values could have been observed, but were not in the sample). In this case, if zeros occur in marginal tables used in the estimation:

- Only certain u-parameters may be estimated.
- The chi-square goodness-of-fit statistic has reduced degrees of freedom.

Some zeros are necessarily *fixed*; for example, some genetic combinations are fatal to offspring and will not be observed in a population. Log-linear models can be used in the analysis (see Bishop et al., [1975]; Haberman [1979]; Fienberg [1977]).

7.7 GSK Approach to Higher-Dimensional Contingency Tables

The second major method of analyzing multivariate contingency tables is due to Grizzle et al. [1969]. They present an analysis method closely related to multiple regression (Chapter 11). References in which this method are considered are Reynolds [1977] and Kleinbaum et al. [1988].

PROBLEMS

In Problems 7.1–7.9, perform the following tasks as well as any other work requested. Problems 7.1–7.5 are taken from the seat belt paper of Robertson [1975].

- **(a)** If a table of expected values is given with one or more missing values, compute the missing values.
- **(b)** If the chi-square value is not given, compute the value of the chi-square statistic.
- **(c)** State the degrees of freedom.
- **(d)** State whether the chi-square p-value is less than or greater than 0.01, 0.05, and 0.10 .
- **(e)** When tables are given with missing values for the adjusted residual values, p-values and $(r - 1) \times (c - 1) \times p$-values, fill in the missing values.
- **(f)** When percent tables are given with missing values, fill in the missing percentages for the row percent table, column percent table, and total percent table, as applicable.
- **(g)** Using the 0.05 significance level, interpret the findings. (Exponential notation is used for some numbers, e.g., $34,000 = 3.4 \times 10^4 = 3.4\text{E}4$; $0.0021 = 2.1 \times 10^{-3} = 2.1\text{E} - 3$.)
- **(h)** Describe verbally what the row and column percents mean. That is, "of those with zero vessels diseased . . . ," and so on.

7.1 In 1974 vehicles, seat belt use was considered in association with the ownership of the vehicle. ("L/S" means "both lap and shoulder belt.")

	Ownership			
Belt Use	**Individuals**	**Rental**	**Lease**	**Other Corporate**
L/S	583	145	86	182
Lap Only	139	24	24	31
None	524	59	74	145

Expected			
615.6	112.6	90.9	176.9
134.7	24.7	19.9	38.7
495.7	90.7	?	?

Adjusted Residuals			
−2.99	?	−0.76	0.60
0.63	−0.15	1.02	−1.44
2.65	−4.55	?	0.31

p−Values			
0.0028	5E − 6	0.4481	0.5497
0.5291	0.8821	?	?
0.0080	5.3E − 6	0.8992	0.7586

$(r - 1) \times (c - 1) \times p-$ Values			
0.017	3E − 5	1+	1+
1+	1+	1+	0.8869
0.048	3E − 5	1+	1+

Column Percents			
47	?	47	51
11	11	13	9
42	?	?	41

$d.f. =?$

$X^2 = 26.72$

7.2 In 1974 cars, belt use and manufacturer were also examined. One hundred eighty-nine cars from "other" manufacturers are not entered into the table.

	Manufacturer					
Belt Use	**GM**	**Toyota**	**AMC**	**Chrysler**	**Ford**	**VW**
L/S	498	25	36	74	285	33
Lap only	102	5	12	29	43	11
None	334	18	30	67	259	51

Adjusted Residuals					
3.06	0.33	−0.65	−1.70	−0.69	−3.00
0.49	−0.03	?	2.89	−3.06	0.33
−3.43	−0.32	−0.23	−0.08	2.63	?

p−Values					
0.0022	0.7421	0.5180	0.0898	0.4898	0.0027
0.6208	0.9730	?	0.0039	0.0022	0.7415
0.0006	0.7527	?	0.9366	0.0085	0.0043

Column Percents					
53	52	46	44	49	?
11	10	15	17	7	?
36	38	38	39	?	?

d.f. =?
$X^2 = 34.30$

7.3 The relationship between belt use and racial appearance in the 1974 models is given here. Thirty-four cases whose racial appearance was "other" are excluded from this table.

	Racial Appearance	
Belt Use	**White**	**Black**
L/S	866	116
Lap only	206	20
None	757	102

Expected	
868.9	113.1
?	26.0
?	98.9

Adjusted Residuals	
−0.40	0.40
1.33	−1.33
?	?

p−Values	
0.69	0.69
?	?
0.67	0.67

d.f. =?
X^2 =?

7.4 The following data are given as the first example in Note 7.2. In the 1974 cars, belt use and age were cross-tabulated.

Expected		
217.59	556.64	?
49.22	125.93	?
?	481.42	194.39

Adjusted Residuals		
?	2.06	−1.23
−2.26	−0.13	?
2.67	−2.00	−0.25

p−Values		
0.219	?	0.217
0.024	0.895	0.017
0.007	?	0.799

$(r-1) \times (c-1) \times p$−Values		
0.88	0.16	0.87
?	?	?
0.03	0.18	1+

Column %s		
45	?	45
?	?	14
47	39	41

Row %s		
?	?	?
16	55	29
25	53	22

d.f. =?
$X^2 = 14.06$

7.5 In the 1974 cars, seat belt use and gender of the driver were related as follows:

	Gender	
Belt Use	Female	Male
L/S	267	739
Lap only	85	142
None	261	606

Expected	
?	?
66.3	160.7
253.1	613.9

Adjusted Residuals	
?	?
2.90	−2.90
0.77	−0.77

p−Values	
0.0104	0.0104
0.0038	0.0038
?	?

$(r-1) \times (c-1) \times p$−Values	
0.02	0.02
0.01	0.01
?	?

Column %s	
44	50
14	?
43	?

Total %s	
13	35
4	?
?	?

d.f. =?
X^2 =?

7.6 The data are given in the second example of Note 7.2. The association of chest pain classification and amount of coronary artery disease was examined.

Adjusted Residuals			**$(r-1) \times (c-1) \times p$−Values**		
−13.95	0.33	12.54	1.0E − 30	1+	4.3E − 27
?	1.57	−1.27	1+	0.47	0.82
18.32	−2.47	−14.80	1.4E − 40	0.05	8.8E − 33

Row %s			**Column %s**			
11	22	68	?	43	59	d.f. =?
30	23	46	?	44	39	$X^2 = 418.48$
?	?	?	?	12	2	

7.7 Peterson et al. [1979] studied the age at death of children who died from sudden infant death syndrome (SIDS). The deaths from a variety of causes, including SIDS, were cross-classified by the age at death, as in Table 7.16, taken from death records in King County, Washington, over the years 1969–1977.

Table 7.16 Death Data for Problem 7.7[a]

	Age at Death				
Cause	0 Days	1–6 Days	2–4 Weeks	5–26 Weeks	27–51 Weeks
Hyaline membrane disease	19	51	7	0	0
Respiratory distress syndrome	68	191	46	0	3
Asphyxia of the newborn	105	60	7	4	2
Immaturity	104	34	3	0	0
Birth injury	115	105	17	2	0
Congenital malformation	79	101	72	75	32
Infection	7	38	36	43	18
SIDS	0	0	24	274	24
All other	60	51	28	58	35

[a] d.f. =?; $X^2 = 1504.18$.

(a) The values of $(r-1) \times (c-1) \times p$-value for the adjusted residual are given here multiplied by −1 if the adjusted residual is negative and multiplied by +1 if the adjusted residual is positive.

−1+	1.4E − 9	−1+	−3.8E − 5	−0.89
−0.43	4.6E − 26	1+	−3.3E − 20	−3.2E − 3
3.0E − 18	1+	−0.02	−4.5E − 10	−0.18
2.3E − 26	−1+	−5.8E − 3	−1.2E − 9	−0.08
1.1E − 11	3.9E − 4	−0.42	−3.8E − 15	−1.6E − 3
−0.20	−1+	7.7E − 6	−1+	0.12
−1.1E − 8	−1+	1.3E − 5	0.90	6.5E − 3
−31.2E − 25	−3.4E − 28	−0.19	1.7E − 57	1+
−1+	−0.03	1+	1+	2.9E − 9

What is the distribution of SIDS cases under the null hypothesis that all causes have the same distribution?

(b) What percent display (row, column, or total) would best emphasize the difference?

7.8 Morehead [1975] studied the relationship between the retention of intrauterine devices (IUDs) and other factors. The study participants were from New Orleans, Louisiana. Tables relating retention to the subjects' age and to parity (the number of pregnancies) are studied in this problem (one patient had a missing age).

(a) Was age related to IUD retention?

Age	**Continuers**	**Terminators**
19–24	41	48
25–29	50	40
30+	63	27

Expected	
50.95	?
51.52	38.5
51.52	38.5

Adjusted Residuals	
−2.61	2.61
−0.40	0.40
?	?

***p*−Values**	
0.0091	0.0091
?	?
0.0027	0.0027

Column %s	
26.6	41.7
?	34.8
?	23.5

Row %s	
46.1	53.9
?	?
70.0	30.0

d.f. =?
X^2 =?

(b) The relationship of parity and IUD retention gave these data:

Parity	**Continuers**	**Terminators**
1–2	59	53
3–4	39	34
5+	57	28

Adjusted Residuals	
−1.32	1.32
−0.81	0.81
?	?

Total %s	
?	19.6
14.4	?
21.1	?

d.f. =?
$X^2 = 4.74$

7.9 McKeown et al. [1952] investigate evidence that the environment is involved in infantile pyloric stenosis. The relationship between the age at onset of the symptoms in days, and the rank of the birth (first child, second child, etc.) was given as follows:

	Age at Onset of Symptoms (Days)						
Birth Rank	0–6	7–13	14–20	21–27	28–34	35–41	≥ 42
1	42	41	116	140	99	45	58
2	28	35	63	53	49	23	31
≥ 3	26	21	39	48	39	14	23

(a) Find the expected value (under independence) for cell ($i = 2, j = 3$). For this cell compute (observed - expected)^2/ expected.

(b) The chi-square statistic is 13.91. What are the degrees of freedom? What can you say about the p-value?

(c) In the paper, the authors present, the column percents, not the frequencies, as above. Fill in the missing values in both arrays below. The arrangement is the same as the first table.

44	42	53	58	53	55	52
29	36	29	?	26	28	28
?	?	18	20	21	17	21

The adjusted residual p-values are

0.076	0.036	0.780	0.042	0.863	0.636	0.041
0.667	0.041	0.551	0.035	0.710	0.874	0.734
0.084	0.734	?	0.856	0.843	0.445	0.954

What can you conclude?

(d) The authors note that the first two weeks appear to have different patterns. They also present the data as:

	Age at Onset (Days)	
Birth Rank	0–13	≥ 14
1	83	458
2	63	219
≥ 3	47	163

For this table, $X^2 = 8.35$. What are the degrees of freedom? What can you say about the p-value?

(e) Fill in the missing values in the adjusted residual table, p-value table, and column percent table. Interpret the data.

Adjusted Residuals		p–Values		Column %s	
−2.89	2.89	0.0039	0.0039	43	55
?	?	0.065	0.065	33	?
1.54	−1.54	?	?	24	?

(f) Why is it crucial to know whether prior to seeing these data the investigators had hypothesized a difference in the parity distribution between the first two weeks and the remainder of the time period?

Problems 7.10–7.16 deal with the chi-square test for trend. The data are from a paper by Kennedy et al. [1981] relating operative mortality during coronary bypass operations to various risk factors. For each of the tables, let the scores for the chi-square test for trend be consecutive integers. For each of the tables:

- **a.** Compute the chi-square statistic for trend. Using Table A.3, give the strongest possible statement about the p-value.
- **b.** Compute, where not given, the percentage of operative mortality, and plot the percentage for the different categories using equally spaced intervals.
- **c.** The usual chi-square statistic (with $k - 1$ degrees of freedom) is given with its p-value. When possible, from Table A.3 or the chi-square values, tell which statistic is more highly significant (has the smallest p-value). Does your figure in (b) suggest why?

7.10 The amount of anginal (coronary artery disease) chest pain is categorized by the Canadian Heart Classification from mild (class I) to severe (class IV).

	Anginal Pain Classification			
Surgical Mortality	**I**	**II**	**III**	**IV**
Yes	6	19	47	59
No	242	1371	2494	1314
% surgical mortality	2.4	1.4	1.8	?

Usual
$X^2 = 31.19$
$p = 7.7\text{E} - 7$

7.11 Congestive heart failure occurs when the heart is not pumping sufficient blood. A heart damaged by a myocardial infarction, heart attack, can incur congestive heart failure. A score from 0 (good) to 4 (bad) for congestive heart failure is related to operative mortality.

	Congestive Heart Failure Score				
Operative Mortality	**0**	**1**	**2**	**3**	**4**
Yes	73	50	13	12	4
No	4480	1394	404	164	36
% operative mortality	1.6	3.4	?	6.8	10.0

Usual
$X^2 = 46.45$
$p = 1.8\text{E} - 9$

7.12 A measure of left ventricular performance, or the pumping action of the heart, is the *ejection fraction*, which is the percentage of the blood in the left ventricle that is pumped during the beat. A high number indicates a more efficient performance.

	Ejection Fraction (%)				
Operative Mortality	**< 19**	**20–29**	**30–39**	**40–49**	**≥ 50**
Yes	1	4	5	22	74
No	14	88	292	685	3839
% operative mortality	6.7	?	?	3.1	1.9

Usual
$X^2 = 8.34$
$p = 0.080$

7.13 A score was derived from looking at how the wall of the left ventricle moved while the heart was beating (details in CASS [1981]). A score of 5 was normal; the larger the score, the worse the motion of the left ventricular wall looked. The relationship to operative mortality is given here.

	Wall Motion Score				
Operative Mortality	**5–7**	**8–11**	**12–15**	**16–19**	**≥ 20**
Yes	65	36	32	10	2
No	3664	1605	746	185	20
% operative mortality	1.7	2.2	?	5.1	9.1

Usual
$X^2 = 28.32$
$p = 1.1\text{E} - 5$

What do you conclude about the relationship? That is, if you were writing a paragraph to describe this finding in a medical journal, what would you say?

7.14 After the blood has been pumped from the heart, and the pressure is at its lowest point, a low blood pressure in the left ventricle is desirable. This left ventricular end diastolic pressure [LVEDP] is measured in millimeters of mercury (mmHg).

	LVEDP			
Operative Mortality	**0–12**	**13–18**	**19–24**	**≥24**
Yes	56	43	22	26
No	3452	1692	762	416
% operative mortality	?	2.5	2.8	5.9

Usual
$X^2 = 34.49$
$p = 1.6\text{E} - 7$

7.15 The number of diseased vessels and operative mortality are given by:

	Diseased Vessels		
Operative Mortality	**1**	**2**	**3**
Yes	17	43	91
No	1196	2018	3199
% operative mortality	1.4	2.1	?

Usual
$X^2 = 7.95$
$p = 0.019$

7.16 The left main coronary artery, if occluded (i.e., totally blocked), blocks two of the three major arterial vessels to the heart. Such an event almost always leads to death. Thus, people with much narrowing of the left main coronary artery usually receive surgical therapy. Is this narrowing also associated with higher surgical mortality?

	Percentage Narrowing			
Operative Mortality	**0–49**	**50–74**	**75–89**	**≥ 90**
Yes	116	8	10	19
No	5497	486	268	222
% operative mortality	2.1	1.6	?	7.9

Usual
$X^2 = 37.75$
$p = 3.2\text{E} - 8$

7.17 In Robertson's [1975] seat belt study, the observers (unknown to them) were checked by sending cars through with a known seat belt status. The agreement numbers between the observers and the known status were:

	Belt Use in Vehicles Sent		
Belt Use Reported	**S/L**	**Lap Only**	**No Belt**
Shoulder and lap	28	2	0
Lap only	3	33	6
No belt	0	15	103

(a) Compute P_A, P_C, and κ.

(b) Construct a 95% confidence interval for κ.

(c) Find the two-sided p-value for testing $\kappa = 0$ (for the entire population) by using $Z = \kappa/\text{SE}_0(\kappa)$.

7.18 The following table is from [Fisher et al., 1982]. The coronary artery tree has considerable biological variability. If the right coronary artery is normal-sized and supplies its usual share of blood to the heart, the circulation of blood is called *right dominant*. As the right coronary artery becomes less important, the blood supply is characterized as balanced and then *left dominant*. The data for the clinical site and quality control site joint readings of angiographic films are given here.

	Dominance (Clinical Site)		
Dominance (QC Site)	**Left**	**Balanced**	**Right**
Left	64	7	4
Balanced	4	35	32
Right	8	21	607

(a) Compute P_A, P_C, and κ (Section 7.4).

(b) Find var(κ) and construct a 90% confidence interval for the population value of κ.

7.19 Example 7.4 discusses the quality control data for the CASS arteriography (films of the arteries). A separate paper by Wexler et al. [1982] examines the study of the left ventricle. Problem 7.12 describes the ejection fraction. Clinical site and quality control site readings of ejection gave the following table:

	Ejection Fraction (QC Site)		
Ejection Fraction (Clinical Site)	**≥ 50%**	**30–49%**	**< 30%**
≥ 50%	302	27	5
30–49%	40	55	9
< 30%	1	9	18

(a) Compute P_A, P_C, and κ.

(b) Find SE(κ) and construct a 99% confidence interval for the population value of κ.

7.20 The value of κ depends on how we construct our categories. Suppose that in Example 7.4 we combine normal and other zero-vessel disease to create a zero-vessel disease category. Suppose also that we combine two- and three-vessel disease into a multivessel-disease category. Then the table becomes:

Vessels Diseased	**Vessels Diseased (Clinical Site)**		
(QC Site)	**0**	**1**	**Multi-**
0	70	20	9
1	10	155	78
Multi-	2	29	497

(a) Compute P_A, P_C, and κ.

(b) Is this kappa value greater than or less than the value in Example 7.4? Will this always occur? Why?

(c) Construct a 95% confidence interval for the population value of κ.

7.21 Zeiner-Henriksen [1972a] compared personal interview and postal inquiry methods of assessing infarction. His introduction follows:

> The questionnaire developed at the London School of Hygiene and Tropical Medicine and later recommended by the World Health Organization for use in field studies of cardiovascular disease has been extensively used in various populations. While originally developed for personal interviews, this questionnaire has also been employed for postal inquiries. The postal inquiry method is of course much cheaper than personal interviewing and is without interviewer error.
>
> A Finnish–Norwegian lung cancer study offered an opportunity to evaluate the repeatability at interview of the cardiac pain questionnaire, and to compare the interview symptom results with those of a similar postal inquiry. The last project, confined to a postal inquiry of the chest pain questions in a sub-sample of the 4092 men interviewed, was launched in April 1965, $2\frac{1}{2}$ to 3 years after the original interviews.
>
> The objective was to compare the postal inquiry method with the personal interview method as a means of assessing the prevalence of angina and possible infarction ....

The data are given in Table 7.17.

(a) Compute P_A, P_C, and κ.

(b) Construct a 90% confidence interval for the population value of $\kappa(\sqrt{\text{var}(\kappa)} = 0.0231)$.

(c) Group the data in three categories by:

(**i**) combining PI + AP, PI only, and AP only; (**ii**) combining the two PI/AP negatives categories; (**iii**) leaving "incomplete" as a third category. Recompute P_A, P_C, and κ. (This new grouping has the categories "cardiovascular symptoms," "no symptoms," and "incomplete.")

Table 7.17 Interview Data for Problem 7.21

	Interview						
		PI	AP	PI/AP Negative			
Postal Inquiry	PI[a] + AP[a]	Only	Only	Nonspecific	Other	Incomplete	Total
PI + AP	23	15	9	6	—	1	54
PI only	14	18	14	24	8	—	78
AP only	3	5	20	12	17	3	60
PI/AP negative							
Nonspecific	2	8	8	54	24	5	101
Other	2	3	5	62	279	1	352
Incomplete	—	2	—	22	37	—	61
Total	44	51	56	180	365	10	706

[a] PI, possible infarction; AP, angina pectoris.

Table 7.18 Interview Results for Problem 7.22

	Interview					
				I− A−		
Postal Inquiry[a]	I+ A+	I+ A−	I− A+	Nonspecific	Other	Total
I+ *A*+	11	3	1	1	—	16
I+ *A*−	2	14	—	4	—	20
I− *A*+	5	2	7	1	1	16
I− *A*−						
Nonspecific	1	4	5	39	9	58
Other	1	8	6	40	72	127
Total	20	31	19	85	82	237

[a] I+, positive infarction; I−, negative infarction; A+ and A−, positive or negative indication of angina.

7.22 In a follow-up study, Zeiner-Henriksen [1972b] evaluated the reproducibility of their method using reinterviews. Table 7.18 shows the results.

(a) Compute P_A, P_C, and κ for these data.

(b) Construct a 95% confidence interval for the population value of kappa. SE(κ) = 0.043.

(c) What is the value of the Z-statistic for testing no association that is computed from kappa and its estimated standard error $\sqrt{\text{var}_0(\kappa)} = 0.037$?

7.23 Weiner et al. [1979] studied men and women with suspected coronary disease. They were studied by a maximal exercise treadmill test. A positive test ($\geq$ 1 mm of ST-wave depression on the exercise electrocardiogram) is thought to be indicative of coronary artery disease. Disease was classified into zero-, one- (or single-), and multivessel disease. Among people with chest pain thought probably anginal (i.e., due to coronary artery disease), the following data are found.

	Vessels Diseased		
Category	**0**	**1**	**Multi-**
Males, + test	47	86	227
Males, − test	132	53	49
Females, + test	62	28	44
Females, − test	83	14	9

The disease prevalence is expected to be significantly different in men and women. We want to see whether the exercise test is related to disease separately for men and women.

(a) For males, the relationship of + or − test and disease give the data below. Fill in the missing values, interpret these data, and answer the questions.

	Vessels Diseased		
Exercise Test	**0**	**1**	**Multi-**
+	47	86	?
−	132	?	49

Expected		
108.5	84.2	167.3
70.5	54.8	?

Adjusted Residuals		
0+	0.73	0+
0+	0.73	0+

Row Percents		
?	?	63.1
56.4	22.6	20.9

Column Percents		
26.3	61.9	?
73.7	38.1	?

Formulate a question for which the row percents would be a good method of presenting the data. Formulate a question where the column percents would be more appropriate.

***7.24** **(a)** Find the natural logarithms, ln x, of the following x: 1.24, 0.63, 0.78, 2.41, 2.7182818, 1.00, 0.10. For what values do you think ln x is positive? For what values do you think ln x is negative? (A plot of the values may help.)

(b) Find the exponential, e^x, of the following x: −2.73, 5.62, 0.00, −0.11, 17.3, 2.45. When is e^x less than 1? When is e^x greater than 1?

(c) $\ln(a \times b) = \ln a + \ln b$. Verify this for the following pairs of a and b:

a :	2.00	0.36	0.11	0.62
b :	0.50	1.42	0.89	0.77

(d) $e^{a+b} = e^a \cdot e^b$. Verify this for the following pairs of numbers:

a:	−2.11	0.36	0.88	−1.31
b:	2.11	1.59	−2.67	−0.45

Table 7.19 Angina Data for Problem 7.25

Model[a]	d.f.	LRX^2	$p-$Value
$[I][J][K]$	7	114.41	0+
$[I][K]$	6	103.17	0+
$[IK][J]$	5	26.32	0+
$[I][JK]$	5	94.89	0+
$[IJ][IK]$	4	15.08	0.0045
$[IJ][JK]$	4	83.65	0+
$[IK][JK]$	3	6.80	0.079
$[IJ][IK][JK]$	2	2.50	0.286

[a] I, J, and K refer to variables as in Example 7.5.

Table 7.20 Hypothesis Data for Problem 7.25

Cell (i, j, k)	Observed	r Fitted	u-Parameters
(1,1,1)	17	18.74	$u = -3.37$
(1,1,2)	86	85.01	$u_1^I = -u_2^I = 0.503$
(1,1,3)	244	243.25	$u_1^J = -u_2^J = 0.886$
(1,2,1)	5	3.26	$u_1^K = -0.775, u_2^K = -0.128, u_3^K = 0.903$
(1,2,2)	14	14.99	$u_{1,1}^{IJ} = -u_{1,2}^{IJ} = -u_{2,1}^{IJ} = u_{2,2}^{IJ} = -0.157$
(1,2,3)	99	99.75	$u_{1,1}^{IK} = -u_{2,1}^{IK} = -0.728$
(2,1,1)	42	40.26	$u_{1,2}^{IK} = -u_{2,2}^{IK} = 0.143$
(2,1,2)	31	31.99	$u_{1,3}^{IK} = -u_{2,3}^{IK} = 0.586$
(2,1,3)	37	37.75	$u_{1,1}^{JK} = -u_{2,1}^{JK} = 0.145$
(2,2,1)	2	3.74	$u_{1,2}^{JK} = -u_{2,2}^{JK} = 0.138$
(2,2,2)	4	3.01	$u_{1,3}^{JK} = -u_{2,3}^{JK} = -0.283$
(2,2,3)	9	8.25	

***7.25** Example 7.5 uses Weiner et al. [1979] data for cases with probable angina. The results for the cases with definite angina are given in Table 7.19.

(a) Which models are at all plausible?

(b) The data for the fit of the $[IJ][IK][JK]$ hypothesis are given in Table 7.20.

Using the u-parameters, compute the fitted value for the (1,2,3) cell, showing that it is (approximately) equal to 99.75 as given.

(c) Using the fact that hypothesis 7 is nested within hypothesis 8, compute the chi-square statistic for the additional gain in fit between the models. What is the p-value (as best as you can tell from the tables)?

***7.26** As in Problem 7.25, the cases of Example 7.5, but with chest pain thought not to be due to heart disease (nonischemic), gave the goodness-of-fit likelihood ratio chi-square statistics shown in Table 7.21.

(a) Which model would you prefer? Why?

(b) For model $[IJ]$ $[IK]$, the information on the fit is given in Table 7.22.

Using the u-parameter values, verify the fitted value for the (2,1,1) cell.

(c) Interpret the probabilistic meaning of the model in words for the variables of this problem.

Table 7.21 Goodness-of-Fit Data for Problem 7.23

Model	d.f.	LRX^2	$p-$Value
$[I][J][K]$	7	35.26	0+
$[IJ][K]$	6	28.45	0+
$[IK][J]$	5	11.68	0.039
$[I][JK]$	5	32.46	0+
$[IJ][IK]$	4	4.87	0.30
$[IJ][JK]$	4	25.65	0+
$[IK][JK]$	3	8.89	0.031
$[IJ][IK][JK]$	2	2.47	0.29

Table 7.22 Fit Data for Problem 7.23

Cell (i, j, k)	Observed	r Fitted	u-Parameters
(1,1,1)	33	32.51	$u = -3.378$
(1,1,2)	13	12.01	$u_1^I = -u_2^I = 0.115$
(1,1,3)	7	8.48	$u_1^J = -u_2^J = 0.658$
(1,2,1)	13	13.49	$u_1^K = 1.364, u_2^K = -0.097, u_3^K = -1.267$
(1,2,2)	4	4.99	$u_{1,1}^{IJ} = -u_{1,2}^{IJ} = -u_{2,1}^{IJ} = u_{2,2}^{IJ} = -0.218$
(1,2,3)	5	3.52	$u_{1,1}^{IK} = -u_{2,1}^{IK} = -0.584$
(2,1,1)	126	128.69	$u_{1,2}^{IK} = -u_{2,2}^{IK} = -0.119$
(2,1,2)	21	18.75	$u_{1,3}^{IK} = -u_{2,3}^{IK} = 0.703$
(2,1,3)	3	2.56	
(2,2,1)	25	22.31	
(2,2,2)	1	3.25	
(2,2,3)	0	0.44	

***7.27** Willkens et al. [1981] study possible diagnostic criteria for Reiter's syndrome. This rheumatic disease was considered in the context of other rheumatic diseases. Eighty-three Reiter's syndrome cases were compared with 136 cases with one of the following four diagnoses: ankylosing spondylitis, seronegative definite rheumatoid arthritis, psoriatic arthritis, and gonococcal arthritis. A large number of potential diagnostic criteria were considered. Here we consider two factors: the presence or absence of urethritis and/or cervicitis (for females); and the duration of the initial attack evaluated as greater than or equal to one month or less than one month. The data are given in Table 7.23, and the goodness-of-fit statistics are given in Table 7.24.

(a) Fill in the question marks in Table 7.24.

(b) Which model(s) seem plausible (at the 0.05 significance level)?

(c) Since we are looking for criteria to differentiate between Reiter's syndrome and the other diseases, one strategy that makes sense is to assume independence of the disease category ($[K]$) and then look for the largest departures from the observed and fitted cells. The model we want is then $[IJ][K]$. The fit is given in Table 7.25.

Which cell of Reiter's syndrome cases has the largest excess of observed minus fitted?

(d) If you use the cell found in part (c) as your criteria for Reiter's syndrome, what are the specificity and sensitivity of this diagnostic criteria for these cases?

Table 7.23 Reiter's Syndrome Data for Problem 7.27

Urethritis and/or Cervicitis [I]	1 Disease [K]	Initial Attack [J] <1 Month	 ≥1 Month
Yes	Reiter's	2	70
	Other	11	3
No	Reiter's	1	10
	Other	20	132

Table 7.24 Goodness-of-Fit Data for Problem 7.27

Model	d.f.	LRX^2	$p-$Value
$[I][J][K]$	?	200.65	?
$[IJ][K]$	?	200.41	?
$[IK][J]$	?	40.63	?
$[I][JK]$	?	187.78	?
$[IJ][IK]$	?	40.39	?
$[IJ][JK]$	?	187.55	?
$[IK][JK]$	?	27.76	?
$[IJ][IK][JK]$	?	5.94	?

Table 7.25 Goodness-of-Fit Data for Problem 7.27

Cell (i, j, k)	Observed	Fitted
(1,1,1)	70	24.33
(1,1,2)	3	48.67
(1,2,1)	2	4.33
(1,2,2)	11	8.67
(2,1,1)	10	47.33
(2,1,2)	132	94.67
(2,2,1)	1	7.00
(2,2,2)	20	14.00

***7.28** We claim in the text that the three-factor log-linear model $[IJ][IK]$ means that the J and K variables are independent conditionally upon the I variable. Prove this by showing the following steps:

(a) By definition, Y and Z are independent conditionally upon X if

$$P[Y = j \text{ and } Z = k | X = i] = P[Y = j | X = i] P[Z = k | X = i]$$

Using the probabilities π_{ijk}, show that this is equivalent to

$$\frac{\pi_{ijk}}{\pi_{i\cdot\cdot}} = \left(\frac{\pi_{ij\cdot}}{\pi_{i\cdot\cdot}}\right)\left(\frac{\pi_{i\cdot k}}{\pi_{i\cdot\cdot}}\right)$$

(b) If the equation above holds true, show that

$$\ln \pi_{ijk} = u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ} + u_{ik}^{IK}$$

where

$$u_{ij}^{IJ} = \ln(\pi_{ij\cdot}) - \frac{1}{I}\sum_{i=1}^{I}\ln(\pi_{ij\cdot}) - \frac{1}{J}\sum_{j=1}^{J}\ln(\pi_{ij\cdot}) + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\ln(\pi_{ij\cdot})$$

$$u_{ik}^{IK} = \ln(\pi_{i\cdot k}) - \frac{1}{I}\sum_{i=1}^{I}\ln(\pi_{i\cdot k}) - \frac{1}{K}\sum_{k=1}^{K}\ln(\pi_{i\cdot k}) + \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\ln(\pi_{i\cdot k})$$

$$u_i^I = \frac{1}{J}\sum_{j=1}^{J}\ln(\pi_{ij\cdot}) + \frac{1}{K}\sum_{k=1}^{K}\ln(\pi_{i\cdot k}) - \ln(\pi_{i\cdot\cdot}) + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\ln(\pi_{ij\cdot})$$
$$+\frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\ln(\pi_{i\cdot k}) - \frac{1}{I}\sum_{i=1}^{I}\ln(\pi_{i\cdot\cdot})$$

$$u_j^J = \frac{1}{I}\sum_{i=1}^{I}\ln(\pi_{ij\cdot}) - \frac{1}{IJ}\sum_{j=1}^{J}\sum_{i=1}^{I}\ln(\pi_{ij\cdot})$$

$$u_k^J = \frac{1}{I}\sum_{i=1}^{I}\ln(\pi_{i\cdot k}) - \frac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\ln(\pi_{i\cdot k})$$

$$u = -\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\ln(\pi_{ij\cdot}) - \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\ln(\pi_{i\cdot k}) - \frac{1}{I}\sum_{i=1}^{I}\ln(\pi_{i\cdot\cdot})$$

(c) If the equation above holds, use $\pi_{ijk} = e^{\ln \pi_{ijk}}$ to show that the first equation then holds.

***7.29** The notation and models for the three-factor log-linear model extend to larger numbers of factors. For example, for variables W, X, Y, and Z (denoted by the indices i, j, k, and l, respectively), the following notation and model correspond:

$$[IJK][L] = u + u_i^I + u_j^J + u_k^K + u_l^L + u_{ij}^{IJ} + u_{ik}^{IK} + u_{jk}^{JK} + u_{ijk}^{IJK}$$

(a) For the four-factor model, write the log-linear u-terms corresponding to the following model notations: **(i)** $[IJ][KL]$; **(ii)** $[IJK][IJL][JKL]$; **(iii)** $[IJ][IK][JK][L]$.

(b) Give the bracket notation for the models corresponding to the u-parameters: **(i)** $u + u_i^I + u_j^J + u_k^K + u_l^L$; **(ii)** $u + u_i^I + u_j^J + u_k^K + u_l^L + u_{ij}^{IJ} + u_{kl}^{KL}$; **(iii)** $u + u_i^I + u_j^J + u_k^K + u_l^L + u_{ij}^{IJ} + u_{ik}^{IK} + u_{il}^{IL} + u_{jk}^{JK} + u_{ijk}^{IJK}$.

***7.30** Verify the values of the contingency coefficients, or measures of association, given in the first example of Note 7.2.

***7.31** Verify the values of the measures of association given in the second example of Note 7.2.

***7.32** Prove the following properties of some of the measures of association, or contingency coefficients, presented in Note 7.2.

(a) $0 \leq \lambda_C \leq 1$. Show by example that 0 and 1 are possible values.

(b) $0 \leq \lambda \leq 1$. Show by example that 0 and 1 are possible values. What happens if the two traits are independent in the sample $n_{ij} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$?

(c) $-1 \leq \gamma \leq 1$. Can γ be -1 or $+1$? If the traits are independent in the sample, show that $\gamma = 0$. Can $\gamma = 0$ otherwise? If yes, give an example.

(d) $0 < C < 1$.

(e) $0 \leq V \leq 1$.

(f) $0 \leq T \leq 1$ [use part (e) to show this].

(g) Show by example that ϕ^2 can be larger than 1.

***7.33** Compute the contingency coefficients of Note 7.2, omitting γ, for the data of:

(a) Problem 7.1.

(b) Problem 7.5.

REFERENCES

Agresti, A. [2002]. *Categorical Data Analysis*, 2nd ed. Wiley, New York.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. [1975]. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

CASS [1981]. (Principal investigators of CASS and their associates); Killip, T. (ed.); Fisher, L., and Mock, M. (assoc. eds.) National Heart, Lung and Blood Institute Coronary Artery Surgery Study. *Circulation*, **63**: part II, I-1 to I-81.

Cohen, J. [1968]. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**: 213–220.

Everitt, B. S. [1992]. *The Analysis of Contingency Tables*, 2nd ed. Halstead Press, New York.

Fienberg, S. E. [1977]. *The Analysis of Cross-Classified Categorical Data*. MIT Press, Cambridge, MA.

Fisher, L. D., Judkins, M. P., Lesperance, J., Cameron, A., Swaye, P., Ryan, T. J., Maynard, C., Bourassa, M., Kennedy, J. W., Gosselin, A., Kemp, H., Faxon, D., Wexler, L., and Davis, K. [1982]. Reproducibility of coronary arteriographic reading in the Coronary Artery Surgery Study (CASS). *Catheterization and Cardiovascular Diagnosis*, **8**: 565–575. Copyright © 1982 by Wiley-Liss.

Fleiss, J. L., Levin, B., and Park, M. C. [2003]. *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, New York.

Fleiss, J. L., Cohen, J., and Everitt, B. S. [1969]. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**: 323–327.

Goodman, L. A., and Kruskal, W. H. [1979]. *Measures of Association for Cross-Classifications*. Springer-Verlag, New York.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. [1969]. Analysis of categorical data by linear models. *Biometrics*, **25**: 489–504.

Haberman, S. J. [1978]. *Analysis of Qualitative Data, Vol. 1, Introductory Topics*. Elsevier, New York.

Haberman, S. J. [1979]. *Analysis of Qualitative Data, Vol. 2, New Developments*. Elsevier, New York.

Hitchcock, C. R., Ruiz, E., Sutherland, D., and Bitter, J. E. [1966]. Eighteen-month follow-up of gastric freezing in 173 patients with duodenal ulcer. *Journal of the American Medical Association*, **195**: 115–119.

Kennedy, J. W., Kaiser, G. C., Fisher, L. D., Fritz, J. K., Myers, W., Mudd, J. G., and Ryan, T. J. [1981]. Clinical and angiographic predictors of operative mortality from the collaborative study in coronary artery surgery (CASS). *Circulation*, **63**: 793–802.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam, A. [1997]. *Applied Regression Analysis and Multivariable Methods*, 3rd ed. Brooks/Cole, Pacific Grove, California.

Kraemer, H. C., Periyakoil, V. S., and Noda, A. [2002]. Tutorial in biostatistics: kappa coefficient in medical research. *Statistics in Medicine*, **21**: 2109–2119.

Maclure, M., and Willett, W. C. [1987]. Misinterpretation and misuses of the kappa statistic. *American Journal of Epidemiology*, **126**: 161–169.

Maki, D. G., Weise, C. E., and Sarafin, H. W. [1977]. A semi-quantitative culture method for identifying intravenous-catheter-related infection. *New England Journal of Medicine*, **296**: 1305–1309.

McKeown, T., MacMahon, B., and Record, R. G. [1952]. Evidence of post-natal environmental influence in the aetiology of infantile pyloric stenosis. *Archives of Diseases in Children*, **58**: 386–390.

Morehead, J. E. [1975]. Intrauterine device retention: a study of selected social-psychological aspects. *American Journal of Public Health*, **65**: 720–730.

Nelson, J. C., and Pepe, M. S. [2000]. Statistical description of interrater reliability in ordinal ratings. *Statistical Methods in Medical Research*, **9**: 475–496.

Peterson, D. R., van Belle, G., and Chinn, N. M. [1979]. Epidemiologic comparisons of the sudden infant death syndrome with other major components of infant mortality. *American Journal of Epidemiology*, **110**: 699–707.

Reynolds, H. T. [1977]. *The Analysis of Cross-Classifications*. Free Press, New York.

Robertson, L. S. [1975]. Safety belt use in automobiles with starter-interlock and buzzer-light reminder systems. *American Journal of Public Health*, **65**: 1319–1325. Copyright © 1975 by the American Health Association.

Ruffin, J. M., Grizzle, J. E., Hightower, N. C., McHarcy, G., Shull, H., and Kirsner, J. B. [1969]. A cooperative double-blind evaluation of gastric "freezing" in the treatment of duodenal ulcer. *New England Journal of Medicine*, **281**: 16–19.

Time [1962]. Frozen ulcers. *Time*, May 18, pp. 45–47.

van Belle, G., and Cornell, R. G. [1971]. Strengthening tests of symmetry in contingency tables. *Biometrics*, **27**: 1074–1078.

Wangensteen, C. H., Peter, E. T., Nicoloff, M., Walder, A. I., Sosin, H., and Bernstein, E. F. [1962]. Achieving "physiologic gastrectomy" by gastric freezing. *Journal of the American Medical Association*, **180**: 439–444. Copyright © 1962 by the American Medical Association.

Weiner, D. A., Ryan, T. J., McCabe, C. H., Kennedy, J. W., Schloss, M., Tristani, F., Chaitman, B. R., and Fisher, L. D. [1979]. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS). *New England Journal of Medicine*, **301**: 230–235.

Wexler, L., Lesperance, J., Ryan, T. J., Bourassa, M. G., Fisher, L. D., Maynard, C., Kemp, H. G., Cameron, A., Gosselin, A. J., and Judkins, M. P. [1982]. Interobserver variability in interpreting contrast left ventriculograms (CASS). *Catheterization and Cardiovascular Diagnosis*, **8**: 341–355.

Willkens, R. F., Arnett, F. C., Bitter, T., Calin, A., Fisher, L., Ford, D. K., Good, A. E., and Masi, A. T. [1981]. Reiter's syndrome: evaluation of preliminary criteria. *Arthritis and Rheumatism*, **24**: 844–849. Used with permission from J. B. Lippincott Company.

Zeiner-Henriksen, T. [1972a]. Comparison of personal interview and inquiry methods for assessing prevalences of angina and possible infarction. *Journal of Chronic Diseases*, **25**: 433–440. Used with permission of Pergamon Press, Inc.

Zeiner-Henriksen, T. [1972b]. The repeatability at interview of symptoms of angina and possible infarction. *Journal of Chronic Diseases*, **25**: 407–414. Used with permission of Pergamon Press, Inc.

CHAPTER 8

Nonparametric, Distribution-Free, and Permutation Models: Robust Procedures

8.1 INTRODUCTION

In Chapter 4 we worked with the normal distribution, noting the fact that many populations have distributions that are approximately normal. In Chapter 5 we presented elegant one- and two-sample methods for estimating the mean of a normal distribution, or the difference of the means, and constructing confidence intervals. We also examined the corresponding tests about the mean(s) from normally distributed populations. The techniques that we learned are very useful. Suppose, however, that the population under consideration is not normal. What should we do? If the population is not normal, is it appropriate to use the same t-statistic that applies when the sample comes from a normally distributed population? If not, is there some other approach that can be used to analyze such data?

In this chapter we consider such questions. In Section 8.2 we introduce terminology associated with statistical procedures needing few assumptions and in Section 8.3 we note that some of the statistical methods that we have already looked at require very few assumptions.

The majority of this chapter is devoted to specific statistical methods that require weaker assumptions than that of normality. Statistical methods are presented that apply to a wide range of situations. Methods of constructing statistical tests for specific situations, including computer simulation, are also discussed. We conclude with

1. An indication of newer research in the topics of this chapter
2. Suggestions for additional reading if you wish to learn more about the subject matter

8.2 ROBUSTNESS: NONPARAMETRIC AND DISTRIBUTION-FREE PROCEDURES

In this section we present terminology associated with statistical procedures that require few assumptions for their validity.

The first idea we consider is *robustness*:

Definition 8.1. A statistical procedure is *robust* if it performs well when the needed assumptions are not violated "too badly" or if the procedure performs well for a large family of probability distributions.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

By a *procedure* we mean an estimate, a statistical test, or a method of constructing a confidence interval. We elaborate on this definition to give the reader a better idea of the meaning of the term. The first thing to note is that the definition is *not* a mathematical definition. We have talked about a procedure performing "well" but have not given a precise mathematical definition of what "well" means. The term *robust* is analogous to beauty: Things may be considered more or less beautiful. Depending on the specific criteria for beauty, there may be greater or lesser agreement about the beauty of an object. Similarly, different statisticians may disagree about the robustness of a particular statistical procedure depending on the probability distributions of concern and use of the procedure. Nevertheless, as the concept of beauty is useful, the concept of robustness also proves to be useful conceptually and in discussing the range of applicability of statistical procedures.

We discuss some of the ways that a statistical test may be robust. Suppose that we have a test statistic whose distribution is derived for some family of distributions (e.g., normal distributions). Suppose also that the test is to be applied at a particular significance level, which we designate the *nominal* significance level. When other distributions are considered, the *actual* probability of rejecting the null hypothesis when it holds may differ from the *nominal* significance level if the distribution is not one of those used to derive the statistical test. For example, in testing for a specific value of the mean with a normally distributed sample, the t-test may be used. Suppose, however, that the distribution considered is not normal. Then, if testing at the 5% significance level, the actual significance level (the true probability of rejecting under the *null* hypothesis that the population mean has the hypothesized value) may not be 5%; it may vary. A statistical test would be robust over a larger family of distributions if the true significance level and nominal significance level were close to each other. Also, a statistical test is robust if under specific alternatives, the probability of rejecting the null hypothesis tends to be large even when the alternatives are in a more extensive family of probability distributions.

A statistical test may be robust in a particular way for large samples, but not for small samples. For example, for most distributions, if one uses the t-test for the mean when the sample size becomes quite large, the central limit theory shows that the nominal significance level is approximately the same as the true significance level when the null hypothesis holds. On the other hand, if the samples come from a skewed distribution and the sample size is small, the t-test can perform quite badly. Lumley et al., [2002] reviewed this issue and reported that in most cases the t-test performs acceptably even with 30 or so observations, and even in a very extreme example the performance was excellent with 250 observations.

A technique of constructing confidence intervals is robust to the extent that the nominal confidence level is maintained over a larger family of distributions. For example, return to the t-test. If we construct 95% confidence intervals for the mean, the method is robust to the extent that samples from a nonnormal distribution straddle the mean about 95% of the time. Alternatively, a method of constructing confidence intervals is nonrobust if the confidence with which the parameters are in the interval differs greatly from the nominal confidence level. An estimate of a parameter is robust to the extent that the estimate is close to the true parameter value over a large class of probability distributions.

Turning to a new topic, the normal distribution model is useful for summarizing data, because two parameters (in this case, the mean and variance, or equivalently, the mean and the standard deviation) describe the entire distribution. Such a set or family of distribution functions with each member described (or indexed) by a few parameters is called a *parametric family*. The distributions used for test statistics are also parametric families. For example, the t-distribution, the F-distribution, and the χ^2-distribution depend on one or two integer parameters: the degrees of freedom. Other examples of parametric families are the binomial distribution, with its two parameters n and π, and the Poisson distribution, with its parameter λ.

By contrast, *semiparametric families* and *nonparametric families* of distributions are families that cannot be conveniently characterized, or indexed, by a few parameters. For example, if one looked at all possible continuous distributions, it is not possible to find a few parameters that characterize all these distributions.

Definition 8.2. A family of probability distributions that can be characterized by a few parameters is a *parametric family*. A family is *nonparametric* if it can closely approximate any arbitrary probability distribution. A family of probability distributions that is neither parametric nor nonparametric is *semiparametric*.

In small samples the t-test holds for the family of normal distributions, that is, for a parametric family. It would be nice to have a test statistic whose distribution was valid for a larger family of distributions. In large samples the t-test qualifies, but in small samples it does not.

Definition 8.3. Statistical procedures that hold, or are valid for a nonparametric family of distributions, are called *nonparametric statistical procedures*.

The definition of nonparametric here can be made precise in a number of nonequivalent ways, and no single definition is in universal use. See also Note 8.1. The usefulness of the t-distribution in small samples results from the fact that samples from a normal distribution give the same t-distribution for all normal distributions under the null hypothesis. More generally, it is very useful to construct a test statistic whose distribution is the same for all members of some family of distributions. That is, assuming that the sample comes from some member of the family, and the null hypothesis holds, the statistic has a known distribution; in other words, the distribution does not depend upon, or is *free* of, which member of the underlying family of distributions is sampled. This leads to our next definition.

Definition 8.4. A statistical procedure is *distribution-free* over a specified family of distributions if the statistical properties of the procedure do not depend on (or are free of) the underlying distribution being sampled.

A test statistic is distribution-free if under the null hypothesis, it has the same distribution for all members of the family. A method of constructing confidence intervals is distribution-free if the nominal confidence level holds for all members of the underlying family of distributions.

The usefulness of the (unequal variances) t-test in large samples results from the fact that samples from any distribution give the same large-sample normal distribution under the null hypothesis that the means are equal. That is, the t-statistic becomes free of any information about the shape of the distribution as the sample size increases. This leads to a definition:

Definition 8.5. A statistical procedure is *asymptotically distribution-free* over a specified family of distributions if the statistical properties of the procedure do not depend on (or are free of) the underlying distribution being sampled for sufficiently large sample sizes.

In practice, one selects statistical procedures that hold over a wide class of distributions. Often, the wide class of distributions is nonparametric, and the resulting statistical procedure is distribution-free for the family. The procedure would then be both nonparametric and distribution-free. The terms *nonparametric* and *distribution-free* are used somewhat loosely and are often considered interchangeable. The term *nonparametric* is used much more often than the term *distribution-free*.

One would expect that a nonparametric procedure would not have as much statistical power as a parametric procedure *if* the sample observed comes from the parametric family. This is frequently, but not necessarily, true. One method of comparing procedures is to look at their relative efficiency. *Relative efficiency* is a complex term when defined precisely (see Note 8.2), but the essence is contained in the following definition:

Definition 8.6. The *relative efficiency* of statistical procedure A to statistical procedure B is the ratio of the sample size needed for B to the sample size needed for A in order that both procedures have the same statistical power.

For example, if the relative efficiency of A to B is 1.5, then B needs 50% more observations than A to get the same amount of statistical power.

8.3 SIGN TEST

Suppose that we are testing a drug to reduce blood pressure using a crossover design with a placebo. We might analyze the data by taking the blood pressure while not on the drug and subtracting it from the blood pressure while on the drug. These differences resulting from the matched or paired data will have an expected mean of zero if the drug under consideration had no more effect than the placebo effect. If we want to assume normality, a one-sample t-test with a hypothesized mean of zero is appropriate. Suppose, however, that we knew from past experience that there were occasional large fluctuations in blood pressure due to biological variability. If the sample size were small enough that only one or two such fluctuations were expected, we would be hesitant to use the t-test because of the known fact that one or two large observations, or outliers, destroyed the probability distribution of the test (see Problem 8.20). What should we do?

An alternative nonparametric way of analyzing the data is the following. Suppose that there is no treatment effect. All of the difference between the blood pressures measured on-drug and on-placebo will be due to biological variability. Thus, the difference between the two measurements will be due to symmetric random variability; the number is equally likely to be positive or negative. The *sign test* is appropriate for the null hypothesis that observed values have the same probability of being positive or negative: If we look at the number of positive numbers among the differences (and exclude values equal to zero), under the null hypothesis of no drug effect, this number has a binomial distribution, with $\pi = \frac{1}{2}$. *A test of the null hypothesis could be a test of the binomial parameter* $\pi = \frac{1}{2}$. This was discussed in Chapter 6 when we considered McNemar's test. Such tests are called *sign tests*, since we are looking at the sign of the difference.

Definition 8.7. Tests based on the sign of an observation (i.e., plus or minus), and which test the hypothesis that the observation is equally likely to be a plus or minus, are called *sign test procedures*.

Note that it is possible to use a sign test in situations where numbers are not observed, but there is only a rating. For example, one could have a blinded evaluation of patients as worse on-drug than on-placebo, the same on-drug as on-placebo, and better on-drug than on-placebo. By considering only those who were better or worse on the drug, the null hypothesis of no effect is equivalent to testing that each outcome is equally likely; that is, the binomial probability is 1/2, the sign test may be used. Ratings of this type are useful in evaluating drugs when numerical quantification is not available. As tests of $\pi = \frac{1}{2}$ for binomial random variables were discussed in Chapter 6, we will not elaborate here. Problems 8.1 to 8.3 use the sign test.

Suppose that the distribution of blood pressures *did* follow a normal distribution: How much would be lost in the way of efficiency by using the sign test? We can answer this question mathematically in large sample sizes. The relative efficiency of the sign test with respect to the t-test when the normal assumptions are satisfied is 0.64; that is, compared to analyzing data using the t-test, 36% of the samples are effectively thrown away. Alternatively, one needs 1/0.64, or 1.56 times as many observations for the sign test as one would need using the t-test to have the same statistical power in a normal distribution. On the other hand, if the data came from a different mathematical distribution, the Laplace or double exponential distribution, the sign test would be more efficient than the t-test.

In some cases a more serious price paid by switching to the sign test is that a different scientific question is being answered. With the t-test we are asking whether the average blood pressure is lower on drug than on placebo; with the sign test we are asking whether the majority of patients have lower blood pressure on drug than on placebo. The answers may be different and it is important to consider which is the more important question.

The sign test is useful in many situations. It is a "quick-and-dirty" test that one may compute mentally without the use of computational equipment; provided that statistical tables are available, you can get a quick estimate of the statistical significance of an appropriate null hypothesis.

8.4 RANKS

Many of the nonparametric, distribution-free tests are based on one simple and brilliant idea. The approach is motivated by an example.

Example 8.1. The following data are for people who are exercised on a treadmill to their maximum capacity. There were five people in a group that underwent heavy distance-running training and five control subjects who were sedentary and not trained. The maximum oxygen intake rate adjusted for body weight is measured in mL/kg per minute. The quantity is called $\mathrm{VO2_{MAX}}$. The values for the untrained subjects were 45, 38, 48, 49, and 51. The values for the trained subjects were 63, 55, 59, 65, and 77. Because of the larger spread among the trained subjects, especially one extremely large $\mathrm{VO2_{MAX}}$ (as can be seen from Figure 8.1), the values do not look like they are normally distributed. On the other hand, it certainly appears that the training has some benefits, since the five trained persons all exceed the treadmill times of the five sedentary persons. Although we do not want to assume that the values are normally distributed, we should somehow use the fact that the larger observations come from one group and the smaller observations come from the other group. We desire a statistical test whose distribution can be tabulated under the null hypothesis that the probability distributions are the same in the two groups.

The crucial idea is the rank of the observation, which is the position of the observation among the other observations when they are arranged in order.

Definition 8.8. The *rank* of an observation, among a set of observations, is its position when the observations are arranged from smallest to largest. The smallest observation has rank 1, the next smallest has rank 2, and so on. If observations are tied, the rank assigned is the average of the ranks appropriate to the equal numbers.

For example, the ranks of the 10 observations given above would be found as follows: first, order the observations from the smallest to largest; then number them from left to right, beginning at 1.

Observation	38	45	48	49	51	55	59	63	65	77
Rank	1	2	3	4	5	6	7	8	9	10

We now consider several of the benefits of using ranks. In the example above, suppose there was no difference in the $\mathrm{VO2_{\ MAX}}$ value between the two populations. Then we have 10 independent samples (five from each population). Since there would be nothing to distinguish between observations, the five observations from the set of people who experienced training would be equally likely to be any five of the given observations. That is, if we consider the

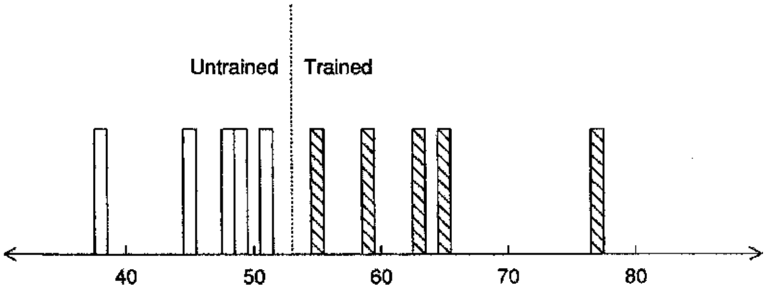

Figure 8.1 $\mathrm{VO2_{\ MAX}}$ in trained and untrained persons.

ranks from 1 to 10, all subsets of size 5 would be equally likely to represent the ranks of the five trained subjects. This is true regardless of the underlying distribution of the 10 observations.

We repeat for emphasis: *If we consider continuous probability distributions (so that there are no ties) under the null hypothesis that two groups of observations come from the same distribution, the ranks have the same distribution*! Thus, tests based on the ranks will be nonparametric tests over the family of continuous probability distributions. Another way of making the same point: Any test that results from using the ranks will be distribution-free, because the distribution of the ranks does not depend on the underlying probability distribution under the null hypothesis.

There is a price to be paid in using rank tests. If we have a small number of observations, say two in each group, even if the two observations in one group are larger than both observations in the other group, a rank test will not allow rejection of the null hypothesis that the distributions are the same. On the other hand, if one knows that the data are approximately normally distributed if the two large observations are considerably larger than the smaller observations, the t-test would allow one to reject the null hypothesis that the distributions are the same. However, this increased statistical power in tiny samples *critically* depends on the normality assumptions. With small sample sizes, one cannot check the adequacy of the assumptions. One may reject the null hypothesis incorrectly (when, in fact, the two distributions are the same) because a large outlying value is observed. This price is specific to small samples: In large samples a particular rank-based test may be more or less powerful than the t-test. Note 8.6 describes another disadvantage of rank tests.

Many nonparametric statistical tests can be devised using the simple idea of ranks. In the next three sections of this chapter we present specific rank tests of certain hypotheses.

8.5 WILCOXON SIGNED RANK TEST

In this section we consider our first rank test. The test is an alternative to the one-sample t-test. Whenever the one-sample t-test of Chapter 5 is appropriate, this test may also be used, as its assumptions will be satisfied. However, since the test is a nonparametric test, its assumptions will be satisfied much more generally than under the assumptions needed for the one-sample t-test. In this section we first discuss the needed assumptions and null hypothesis for this test. The test itself is then presented and illustrated by an example. For large sample sizes, the value of the test statistic may be approximated by a standard normal distribution; the appropriate procedure for this is also presented.

8.5.1 Assumptions and Null Hypotheses

The signed rank test is appropriate for statistically independent observations. The null hypothesis to be tested is that each observation comes from a distribution that is symmetric with a mean of zero. That is, for any particular observation, the value is equally likely to be positive or negative.

For the one-sample t-test, we have independent observations from a normal distribution; suppose that the null hypothesis to be tested has a mean of zero. When the mean is zero, the distribution is symmetric about zero, and positive or negative values are equally likely. Thus, the signed rank test may be used wherever the one-sample t-test of mean zero is appropriate. For large sample sizes, the signed rank test has an efficiency of 0.955 relative to the t-test; the price paid for using this nonparametric test is equivalent to losing only 4.5% of the observations. In addition, when the normal assumptions for the t-test hold and the mean is not zero, the signed rank test has equivalent statistical power.

An example where the signed rank test is appropriate is a crossover experiment with a drug and a placebo. Suppose that subjects have the sequence "placebo, then drug" or "drug, then placebo," each assigned at random, with a probability of 0.5. The null hypothesis of interest is that the drug has the same effect as the placebo. If one takes the difference between measurements taken on the drug and on the placebo, and if the treatment has no effect, the distribution of the difference will not depend on whether the drug was given first or second. The probability is one-half that the placebo was given first and that the observation being looked at is the second observation minus the first observation. The probability is also 1/2 that the observation being examined came from a person who took the drug first. In this case, the observation being used in the signed rank test would be the first observation minus the second observation. Since under the null hypothesis, these two differences have the same distribution except for a minus sign, the distribution of observations under the null hypothesis of "no treatment effect" is symmetric about zero.

8.5.2 Alternative Hypotheses Tested with Power

To use the test, we need to know what type of alternative hypotheses may be detected with some statistical power. For example, suppose that one is measuring blood pressure, and the drug supposedly lowers the blood pressure compared to a placebo. The difference between the measurements on the drug and the blood pressure will tend to be negative. If we look at the observations, two things will occur. First, there will tend to be more observations that have a negative value (i.e., a minus sign) than expected by chance. Second, if we look at the values of the data, the largest absolute values will tend to be negative values. The differences that are positive will usually have smaller absolute values. The signed rank test is designed to use both sorts of information. The signed rank statistic is designed to have power where the alternatives of interest correspond roughly to a shift of the distribution (e.g., the median, rather than being zero, is positive or negative).

8.5.3 Computation of the Test Statistic

We compute the signed rank statistic as follows:

1. Rank the absolute values of the observations from smallest to largest. Note that we do *not* rank the observations themselves, but rather, the absolute values; that is, we ignore minus signs. Drop observations equal to zero.
2. Add up the values of the ranks assigned to the positive observations. Do the same to the negative observations. The smaller of the two values is the value of the Wilcoxon signed rank statistic used in Table A.9 in the Appendix.

The procedure is illustrated in the following example.

Example 8.2. Brown and Hurlock [1975] investigated three methods of preparing the breasts for breastfeeding. The methods were:

1. Toughening the skin of the nipple by nipple friction or rolling
2. Creams to soften and lubricate the nipple
3. Prenatal expression of the first milk secreted before or after birth (colostrum)

Each subject had one randomly chosen treated breast and one untreated breast. Nineteen different subjects were randomized to each of three treatment groups; that is, each subject received the three treatments in random order. The purpose of the study was to evaluate methods of preventing postnatal nipple pain and trauma. The effects were evaluated by the mothers filling out a subjective questionnaire rating nipple sensitivity from "comfortable" (1) to "painful" (2) after each feeding. The data are presented in Table 8.1.

We use the signed rank test to examine the statistical significance of the nipple-rolling data. The first step is to rank the absolute values of the observations, omitting zero values. The observations ranked by absolute value and their ranks are given in Table 8.2.

Note the tied absolute values corresponding to ranks 4 and 5. The average rank 4.5 is used for both observations. Also note that two zero observations were dropped.

Table 8.1 Mean Subjective Difference between Treated and Untreated Breasts

Nipple Rolling	Masse Cream	Expression of Colostrum
−0.525	0.026	−0.006
0.172	0.739	0.000
−0.577	−0.095	−0.257
0.200	−0.040	−0.070
0.040	0.006	0.107
−0.143	−0.600	0.362
0.043	0.007	−0.263
0.010	0.008	0.010
0.000	0.000	−0.080
−0.522	−0.100	−0.010
0.007	0.000	0.048
−0.122	0.000	0.300
−0.040	0.060	0.182
0.000	−0.180	−0.378
−0.100	0.000	−0.075
0.050	0.040	−0.040
−0.575	0.080	−0.080
0.031	−0.450	−0.100
−0.060	0.000	−0.020

Source: Data from Brown and Hurlock [1975].

Table 8.2 Ranked Observation Data

Observation	Rank	Observation	Rank
0.007	1	−0.122	10
0.010	2	−0.143	11
0.031	3	0.172	12
0.040	4.5	0.200	13
−0.040	4.5	−0.522	14
0.043	6	−0.525	15
0.050	7	−0.575	16
−0.060	8	−0.577	17
−0.100	9		

The sum of the ranks of the positive numbers is $S = 1+2+3+4.5+6+7+12+13 = 48.5$. This is less than the sum of the negative ranks. For a sample size of 17, Table A.9 shows that the two-sided p-value is ≥ 0.10. If there are no ties, Owen [1962] shows that $P[S \geq 48.5] = 0.1$ and the two-sided p-value is 0.2. No treatment effect has been shown.

8.5.4 Large Samples

When the number of observations is moderate to large, we may compute a statistic that has approximately a standard normal distribution under the null hypothesis. We do this by subtracting the mean under the null hypothesis from the observed signed rank statistic, and dividing by the standard deviation under the null hypothesis. Here we do not take the minimum of the sums of positive and negative ranks; the usual one- and two-sided normal procedures can be used. The mean and variance under the null hypothesis are given in the following two equations:

$$E(S) = \frac{n(n+1)}{4} \tag{1}$$

$$\operatorname{var}(S) = \frac{n(n+1)(2n+1)}{24} \tag{2}$$

From this, one gets the following statistic, which is approximately normally distributed for large sample sizes:

$$Z = \frac{S - E(S)}{\sqrt{\operatorname{var}(S)}} \tag{3}$$

Sometimes, data are recorded on such a scale that ties can occur for the absolute values. In this case, tables for the signed rank test are conservative; that is, the probability of rejecting the null hypothesis when it is true is *less* than the nominal significance level. The asymptotic statistic may be adjusted for the presence of ties. The effect of ties is to reduce the variance in the statistic. The rank of a term involved in a tie is replaced by the average of the ranks of those tied observations. Consider, for example, the following data:

$$6, -6, -2, 0, 1, 2, 5, 6, 6, -3, -3, -2, 0$$

Note that there are not only some ties, but zeros. In the case of zeros, the zero observations are omitted from the computation as noted before. These data, ranked by absolute value, with average ranks replacing the given rank when the absolute values are tied, are shown below. The first row (A) represents the data ranked by absolute value, omitting zero values; the second row (B) gives the ranks; and the third row (C) gives the ranks, with ties averaged (in this row, ranks of positive numbers are shown in bold type):

A	1	−2	2	−2	−3	−3	5	6	−6	6	6
B	1	2	3	4	5	6	**7**	8	9	10	11
C	**1**	3	**3**	3	5.5	5.5	**7**	**9.5**	9.5	**9.5**	**9.5**

Note that the ties are with respect to the absolute value (without regard to sign). Thus the three ranks corresponding to observations of −2 and +2 are 2, 3, and 4, the average of which is 3. The S-statistic is computed by adding the ranks for the positive values. In this case,

$$S = 1 + 3 + 7 + 9.5 + 9.5 + 9.5 = 39.5$$

Before computing the asymptotic statistic, the variance of S must be adjusted because of the ties. To make this adjustment, we need to know the number of groups that have ties and the number of ties in each group. In looking at the data above, we see that there are three sets of ties, corresponding to absolute values 2, 3, and 6. The number of ties corresponding to observations of absolute value 2 (the "2 group") is 3; the number of ties in the "3 group" is 2; and the number of ties in the "6 group" is 4. In general, let q be the number of groups of ties, and let t_i, where i goes from 1 to q, be the number of observations involved in the particular group. In this case,

$$t_1 = 3, \qquad t_2 = 2, \qquad t_3 = 4, \qquad q = 3$$

In general, the variance of S is reduced according to the equation:

$$\text{var}(S) = \frac{n(n+1)(2n+1) - \frac{1}{2}\sum_{i=1}^{q} t_i(t_i - 1)(t_i + 1)}{24} \tag{4}$$

For the data that we are working with, we started with 13 observations, but the n used for the test statistic is 11, since two zeros were eliminated. In this case, the expected mean and variance are

$$E(S) = 11 \times \frac{12}{4} = 33$$

$$\text{var}(S) = \frac{11 \times 12 \times 23 - \frac{1}{2}(3 \times 2 \times 4 + 2 \times 1 \times 3 + 4 \times 3 \times 5)}{24} \doteq 135.6$$

Using test statistic S gives

$$Z = \frac{S - E(S)}{\sqrt{\text{var}(S)}} = \frac{39.5 - 33}{\sqrt{135.6}} \doteq 0.56$$

With a Z-value of only 0.56, one would not reject the null hypothesis for commonly used values of the significance level. For testing at a 0.05 significance level, if n is 15 or larger with few ties, the normal approximation may reasonably be used. Note 8.4 and Problem 8.22 have more information about the distribution of the signed-rank test.

Example 8.2. (*continued*) We compute the asymptotic Z-statistic for the signed rank test using the data given. In this case, $n = 17$ after eliminating zero values. We have one set of two tied values, so that $q = 1$ and $t_1 = 2$. The null hypothesis mean is $17 \times 18/4 = 76.5$. This variance is $[17 \times 18 \times 35 - (1/2) \times 2 \times 1 \times 3]/24 = 446.125$. Therefore, $Z = (48.5 - 76.5)/21.12 \doteq -1.326$. Table A.9 shows that a two-sided p is about 0.186. This agrees with $p = 0.2$ as given above from tables for the distribution of S.

8.6 WILCOXON (MANN–WHITNEY) TWO-SAMPLE TEST

Our second example of a rank test is designed for use in the two-sample problem. Given samples from two different populations, the statistic tests the hypothesis that the distributions of the two populations are the same. The test may be used whenever the two-sample t-test is appropriate. Since the test given depends upon the ranks, it is nonparametric and may be used more generally. In this section, we discuss the null hypothesis to be tested, and the efficiency of the test relative to the two-sample t-test. The test statistic is presented and illustrated by two examples. The large-sample approximation to the statistic is given. Finally, the relationship between two equivalent statistics, the Wilcoxon statistic and the Mann–Whitney statistic, is discussed.

8.6.1 Null Hypothesis, Alternatives, and Power

The null hypothesis tested is that each of two independent samples has the same probability distribution. Table A.10 for the Mann–Whitney two-sample statistic assumes that there are no ties. Whenever the two-sample t-test may be used, the *Wilcoxon statistic* may also be used. The statistic is designed to have statistical power in situations where the alternative of interest has one population with generally larger values than the other. This occurs, for example, when the two distributions are normally distributed, but the means differ. For normal distributions with a shift in the mean, the efficiency of the Wilcoxon test relative to the two-sample t-test is 0.955.

For other distributions with a shift in the mean, the Wilcoxon test will have relative efficiency near 1 if the distribution is *light-tailed* and greater than 1 if the distribution is *heavy-tailed*.

However, as the Wilcoxon test is designed to be less sensitive to extreme values, it will have less power against an alternative that adds a few extreme values to the data. For example, a pollutant that generally had a normally distributed concentration might have occasional very high values, indicating an illegal release by a factory. The Wilcoxon test would be a poor choice if this were the alternative hypothesis. Johnson et al. [1987] shows that a *quantile test* (see Note 8.5) is more powerful than the Wilcoxon test against the alternative of a shift in the extreme values, and the U.S. EPA [1994] has recommended using this test. In large samples a t-test might also be more powerful than the Wilcoxon test for this alternative.

8.6.2 Test Statistic

The test statistic itself is easy to compute. The combined sample of observations from both populations are ordered from the smallest observation to the largest. The sum of the ranks of the population with the smaller sample size (or in the case of equal sample sizes, an arbitrarily designated first population) gives the value of the Wilcoxon statistic.

To evaluate the statistic, we use some notation. Let m be the number of observations for the smaller sample, and n the number of observations in the larger sample. The Wilcoxon statistic W is the sum of the ranks of the m observations when both sets of observations are ranked together.

The computation is illustrated in the following example:

Example 8.3. This example deals with a small subset of data from the Coronary Artery Surgery Study [CASS, 1981]. Patients were studied for suspected or proven coronary artery disease. The disease was diagnosed by coronary angiography. In coronary angiography, a tube is placed into the aorta (where the blood leaves the heart) and a dye is injected into the arteries of the heart, allowing x-ray motion pictures (angiograms) of the arteries. If an artery is narrowed by 70% or more, the artery is considered significantly diseased. The heart has three major arterial systems, so the disease (or lack thereof) is classified as zero-, one-, two-, or three-vessel disease (abbreviated 0VD, 1VD, 2VD, and 3VD). Narrowed vessels do not allow as much blood to give oxygen and nutrients to the heart. This leads to chest pain (angina) and total blockage of arteries, killing a portion of the heart (called a *heart attack* or *myocardial infarction*). For those reasons, one does not expect people with disease to be able to exercise vigorously. Some subjects in CASS were evaluated by running on a treadmill to their maximal exercise performance. The treadmill increases in speed and slope according to a set schedule. The total time on the treadmill is a measure of exercise capacity. The data that follow present treadmill time in seconds for men with normal arteries (but suspected coronary artery disease) and men with three-vessel disease are as follows:

Normal	1014	684	810	990	840	978	1002	1111		
3VD	864	636	638	708	786	600	1320	750	594	750

Note that $m = 8$ (normal arteries) and $n = 10$ (three-vessel disease). The first step is to rank the combined sample and assign ranks, as in Table 8.3. The sum of the ranks of the smaller normal group is 101. Table A.10, for the closely related Mann–Whitney statistic of Section 8.6.4, shows that we reject the null hypothesis of equal population distributions at a 5% significance level.

Under the null hypothesis, the expected value of the Wilcoxon statistic is

$$E(W) = \frac{m(m+n+1)}{2} \tag{5}$$

Table 8.3 Ranking Data for Example 8.3

Value	Rank	Group	Value	Rank	Group	Value	Rank	Group
594	1	3VD	750	7.5	3VD	978	13	Normal
600	2	3VD	750	7.5	3VD	990	14	Normal
636	3	3VD	786	9	3VD	1002	15	Normal
638	4	3VD	810	10	Normal	1014	16	Normal
684	5	Normal	840	11	Normal	1111	17	Normal
708	6	3VD	864	12	3VD	1320	18	3VD

In this case, the expected value is 76. As we conjectured (*before* seeing the data) that the normal persons would exercise longer (i.e., W would be large), a one-sided test that rejects the null hypothesis if W is too large might have been used. Table A.10 shows that at the 5% significance level, we would have rejected the null hypothesis using the one-sided test. (This is also clear, since the more-stringent two-sided test rejected the null hypothesis.)

8.6.3 Large-Sample Approximation

There is a large-sample approximation to the Wilcoxon statistic (W) under the null hypothesis that the two samples come from the same distribution. The approximation may fail to hold if the distributions are different, even if neither has systematically larger or smaller values. The mean and variance of W, with or without ties, is given by equations (5) through (7). In these equations, m is the size of the smaller group (the number of ranks being added to give W), n the number of observations in the larger group, q the number of groups of tied observations (as discussed in Section 8.6.2), and t_i the number of ranks that are tied in the ith set of ties. First, without ties,

$$\operatorname{var}(W) = \frac{mn(m+n+1)}{12} \tag{6}$$

and with ties,

$$\operatorname{var}(W) = \frac{mn(m+n+1)}{12} - \left[\sum_{i=1}^{q} t_i(t_i-1)(t_i+1)\right] \frac{mn}{12(m+n)(m+n-1)} \tag{7}$$

Using these values, an asymptotic statistic with an approximately standard normal distribution is

$$Z = \frac{W - E(W)}{\sqrt{\operatorname{var}(W)}} \tag{8}$$

Example 8.3. (*continued*) The normal approximation is best used when $n \geq 15$ and $m \geq 15$. Here, however, we compute the asymptotic statistic for the data of Example 8.3.

$$E(W) = \frac{8(10+8+1)}{2} = 76$$

$$\operatorname{var}(W) = \frac{8 \cdot 10(8+10+1)}{12} - 2(2-1)(2+1)\left[\frac{8 \cdot 10}{12(8+10)(8+10+1)}\right]$$

$$= 126.67 - 0.12 = 126.55$$

$$Z = \frac{101-76}{\sqrt{126.55}} \doteq 2.22$$

The one-sided p-value is 0.013, and the two-sided p-value is $2(0.013) = 0.026$. In fact, the exact one-sided p-value is 0.013. Note that the correction for ties leaves the variance virtually unchanged.

Example 8.4. The Wilcoxon test may be used for data that are ordered and ordinal. Consider the angiographic findings from the CASS [1981] study for men and women in Table 8.4. Let us test whether the distribution of disease is the same in the men and women studied in the CASS registry.

You probably recognize that this is a contingency table, and the χ^2-test may be applied. If we want to examine the possibility of a trend in the proportions, the χ^2-test for trend could be used. That test assumes that the proportion of females changes in a linear fashion between categories. Another approach is to use the Wilcoxon test as described here.

The observations may be ranked by the six categories (none, mild, moderate, 1VD, 2VD, and 3VD). There are many ties: 4517 ties for the lowest rank, 1396 ties for the next rank, and so on. We need to compute the average rank for each of the six categories. If J observations have come before a category with K tied observations, the average rank for the k tied observations is

$$\text{average rank} = \frac{2J + K + 1}{2} \tag{9}$$

For these data, the average ranks are computed as follows:

K	*J*	Average	*K*	*J*	Average
4,517	0	2,259	4,907	6,860	9,314
1,396	4,517	5,215.5	5,339	11,767	14,437
947	5,913	6,387	6,997	17,106	20,605

Now our smaller sample of females has 2360 observations with rank 2259, 572 observations with rank 5215.5, and so on. Thus, the sum of the ranks is

$$\begin{aligned} W &= 2360(2259) + 572(5215.5) + 291(6387) + 1020(9314) + 835(14{,}437) + 882(20{,}605) \\ &= 49{,}901{,}908 \end{aligned}$$

The expected value from equation (5) is

$$E(W) = \frac{5960(5960 + 18{,}143 + 1)}{2} = 71{,}829{,}920$$

Table 8.4 Extent of Coronary Artery Disease by Gender

Extent of Disease	Male	Female	Total
None	2,157	2,360	4,517
Mild	824	572	1,396
Moderate	656	291	947
Significant			
1VD	3,887	1,020	4,907
2VD	4,504	835	5,339
3VD	6,115	882	6,997
Total	18,143	5,960	24,103

Source: Data from CASS [1981].

From equation (7), the variance, taking into account ties, is

$$\begin{aligned}\mathrm{var}(W) &= 5960 \times 18{,}143 \times \frac{5960 + 18{,}143 + 1}{12} \\ &\quad -(4517 \times 4516 \times 4518 + \cdots + 6997 \times 6996 \times 6998)\frac{5960 \times 18{,}143}{12 \times 20{,}103 \times 20{,}102} \\ &= 2.06 \times 10^{11}\end{aligned}$$

From this,

$$z = \frac{W - E(W)}{\sqrt{\mathrm{var}(W)}} \doteq -48.29$$

The p-value is extremely small and the population distributions clearly differ.

8.6.4 Mann–Whitney Statistic

Mann and Whitney developed a test statistic that is equivalent to the Wilcoxon test statistic. To obtain the value for the Mann–Whitney test, which we denote by U, one arranges the observations from the smallest to the largest. The statistic U is obtained by counting the number of times an observation from the group with the smallest number of observations precedes an observation from the second group. With no ties, the statistics U and W are related by the following equation:

$$U + W = \frac{m(m + 2n + 1)}{2} \tag{10}$$

Since the two statistics add to a constant, using one of them is equivalent to using the other. We have used the Wilcoxon statistic because it is easier to compute by hand. The values of the two statistics are so closely related that books of statistical tables contain tables for only one of the two statistics, since the transformation from one to the other is almost immediate. Table A.10 is for the Mann–Whitney statistic.

To use the table for Example 8.3, the Mann–Whitney statistic would be

$$U = \frac{8[8 + 2(10) + 1]}{2} - 101 = 116 - 101 = 15$$

From Table A.10, the two-sided 5% significance levels are given by the tabulated values and mn minus the tabulated value. The tabulated two-sided value is 63, and $8 \times 10 - 63 = 17$. We do reject for a two-sided 5% test. For a one-sided test, the upper critical value is 60; we want the lower critical value of $8 \times 10 - 60 = 20$. Clearly, again we reject at the 5% significance level.

8.7 KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST

Definition 3.9 showed one method of describing the distributions of values from a population: the *empirical cumulative distribution*. For each value on the real line, the empirical cumulative distribution gives the proportion of observations less than or equal to that value. One visual way of comparing two population samples would be a graph of the two empirical cumulative distributions. If the two empirical cumulative distributions differ greatly, one would suspect that the populations being sampled were not the same. If the two curves were quite close, it would be reasonable to assume that the underlying population distributions were essentially the same.

The *Kolmogorov–Smirnov statistic* is based on this observation. The value of the statistic is the maximum absolute difference between the two empirical cumulative distribution functions. Note 8.7 discusses the fact that the Kolmogorov–Smirnov statistic is a rank test. Consequently, the test is a nonparametric test of the null hypothesis that the two distributions are the same. When the two distributions have the same shape but different locations, the Kolmogorov–Smirnov statistic is far less powerful than the Wilcoxon rank-sum test (or the t-test if it applies), but the Kolmogorov–Smirnov test can pick up any differences between distributions, whatever their form.

The procedure is illustrated in the following example:

Example 8.4. (*continued*) The data of Example 8.3 are used to illustrate the statistic. Using the method of Chapter 3, Figure 8.2 was constructed with both distribution functions.

From Figure 8.2 we see that the maximum difference is 0.675 between 786 and 810. Tables of the statistic are usually tabulated not in terms of the maximum absolute difference D, but in terms of $(mn/d)D$ or mnD, where m and n are the two sample sizes and d is the lowest common denominator of m and n. The benefit of this is that $(mn/d)D$ or mnD is always an integer. In this case, $m = 8$, $n = 10$, and $d = 2$. Thus, $(mn/d)D = (8)(10/2)(0.675) = 27$ and $mnD = 54$. Table 44 of Odeh et al. [1977] gives the 0.05 critical value for mnD as 48. Since $54 > 48$, we reject the null hypothesis at the 5% significance level. Tables of critical values are not given in this book but are available in standard tables (e.g., Odeh et al. [1977]; Owen [1962]; Beyer [1990]) and most statistics packages. The tables are designed for the case with no ties. If there are ties, the test is conservative; that is, the probability of rejecting the null hypothesis when it is true is even less than the nominal significance level.

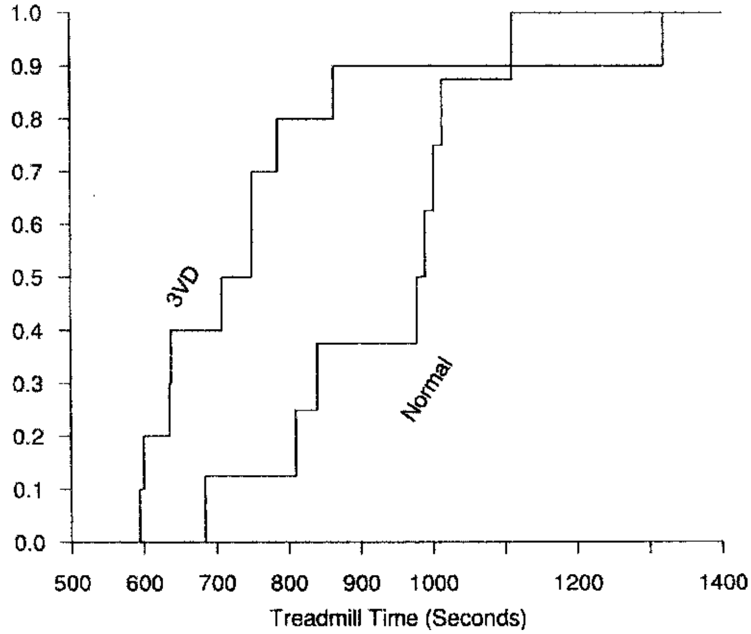

Figure 8.2 Empirical cumulative distributions for the data of Example 8.3.

The large-sample distribution of D is known. Let n and m both be large, say, both 40 or more. The large-sample test rejects the null hypothesis according to the following table:

Significance Level	**Reject the Null Hypothesis if:**
0.001	KS $\geq$ 1.95
0.01	KS $\geq$ 1.63
0.05	KS $\geq$ 1.36
0.10	KS $\geq$ 1.22

KS is defined as

$$\text{KS} = \max_{x} \sqrt{\frac{nm}{n+m}} |F_n(x) - G_m(x)| = \sqrt{\frac{nm}{n+m}} D \tag{11}$$

where F_n and G_m are the two empirical cumulative distributions.

8.8 NONPARAMETRIC ESTIMATION AND CONFIDENCE INTERVALS

Many nonparametric tests have associated estimates of parameters. Confidence intervals for these estimates are also often available. In this section we present two estimates associated with the Wilcoxon (or Mann–Whitney) two-sample test statistic. We also show how to construct a confidence interval for the median of a distribution.

In considering the Mann–Whitney test statistic described in Section 8.6, let us suppose that the sample from the first population was denoted by X's, and the sample from the second population by Y's. Suppose that we observe mX's and nY's. The Mann–Whitney test statistic U is the number of times an X was less than a Y among the nmX and Y pairs. As shown in equation (12), the Mann–Whitney test statistic U, when divided by mn, gives an unbiased estimate of the probability that X is less than Y.

$$E\left(\frac{U}{mn}\right) = P[X < Y] \tag{12}$$

Further, an approximate $100(1-\alpha)\%$ confidence interval for the probability that X is less than Y may be constructed using the asymptotic normality of the Mann–Whitney test statistic. The confidence interval is given by the following equation:

$$\frac{U}{mn} \pm Z_{1-\alpha/2} \sqrt{\frac{1}{\min(m,n)} \frac{U}{mn} \left(1 - \frac{U}{mn}\right)} \tag{13}$$

In large samples this interval tends to be too long, but in small samples it can be too short if U/mn is close to 0 or 1 [Church and Harris, 1970]. In Section 8.10.2 we show another way to estimate a confidence interval.

Example 8.5. This example illustrates use of the Mann–Whitney test statistic to estimate the probability that X is less than Y and to find a 95% confidence interval for $P[X < Y]$.

Examine the normal/3VD data in Example 8.3. We shall estimate the probability that the treadmill time of a randomly chosen person with normal arteries is less than that of a three-vessel disease patient.

Note that 1014 is less than one three-vessel treadmill time; 684 is less than 6 of the three-vessel treadmill times, and so on. Thus,

$$U = 1 + 6 + 2 + 1 + 2 + 1 + 1 + 1 = 15$$

We also could have found U by using equation (9) and $W = 101$ from Example 8.3. Our estimate of $P[X < Y]$ is $15/(8 \times 10) = 0.1875$. The confidence interval is

$$0.1875 \pm (1.96)\sqrt{\frac{1}{8}(0.1875)(1 - 0.1875)} = 0.1875 \pm 0.2704$$

We see that the lower limit of the confidence interval is below zero. As zero is the minimum possible value for $P[X < Y]$, the confidence interval could be rounded off to $[0, 0.458]$.

If it is known that the underlying population distributions of X and Y are the same shape and differ only by a shift in means, it is possible to use the Wilcoxon test (or any other rank test) to construct a confidence interval. This is an example of a *semiparametric* procedure: it does not require the underlying distributions to be known up to a few parameters, but it does impose strong assumptions on them and so is not *nonparametric*. The procedure is to perform Wilcoxon tests of $X + \delta$ vs. Y to find values of δ at which the p-value is exactly 0.05. These values of δ give a 95% confidence interval for the difference in locations.

Many statistical packages will compute this confidence interval and may not warn the user about the assumption that the distributions have the same shape but a different location. In the data from Example 8.5, the assumption does not look plausible: The treadmill times for patients with three-vessel disease are generally lower but with one outlier that is higher than the times for all the normal subjects.

In Chapter 3 we saw how to estimate the median of a distribution. We now show how to construct a confidence interval for the median that will hold for any distribution. To do this, we use *order statistics*.

Definition 8.9. Suppose that one observes a sample. Arrange the sample from the smallest to the largest number. The smallest number is the *first-order statistic*, the second smallest is the *second-order statistic*, and so on; in general, the ith-*order statistic* is the ith number in line.

The notation used for an order statistic is to put the subscript corresponding to the particular order statistic in parentheses. That is,

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

To find a $100(1 - \alpha)\%$ confidence interval for the median, we first find from tables of the binomial distribution with $\pi = 0.5$, the largest value of k such that the probability of k or fewer successes is less than or equal to $\alpha/2$. That is, we choose k to be the largest value of k such that

$$P[\text{number of heads in } n \text{ flips of a fair coin} = 0 \text{ or } 1 \text{ or} \ldots \text{or } k] \le \frac{\alpha}{2}$$

Given the value of k, the confidence interval for the median is the interval between the $(k + 1)$- and $(n - k)$-order statistics. That is, the interval is

$$(X_{(k+1)}, X_{(n-k)})$$

Example 8.6. The treadmill times of 20 females with normal or minimal coronary artery disease in the CASS study are

$$570, 618, 30, 780, 630, 738, 900, 750, 750, 540, 660,$$
$$780, 720, 750, 936, 900, 762, 840, 816, 690$$

We estimate the median time and construct a 90% confidence interval for the median of this population distribution. The order statistics (ordered observations) from 1 to 20 are

$$30, 540, 570, 618, 630, 660, 690, 720, 738, 750, 750,$$
$$750, 762, 780, 780, 816, 840, 900, 900, 936$$

Since we have an odd number of observations,

$$\text{median} = \frac{X_{(10)} + X_{(11)}}{2} = \frac{750 + 750}{2} = 750$$

If X is binomial, $n = 20$ and $\pi = 0.5$, $P[X \leq 5] = 0.0207$ and $P[X \leq 6] = 0.0577$. Thus, $k = 5$. Now, $X_{(6)} = 690$ and $X_{(15)} = 780$. Hence, the confidence interval is (690, 780). The actual confidence is $100(1 - 2 \times 0.0207)\% \doteq 95.9\%$. Because of the discrete nature of the data, the nominal 90% confidence interval is also a 95.9% confidence interval.

*8.9 PERMUTATION AND RANDOMIZATION TESTS

In this section we present a method that may be used to generate a wide variety of statistical procedures. The arguments involved are subtle; you need to pay careful attention to understand the logic. We illustrate the idea by working from an example.

Suppose that one had two samples, one of size n and one of size m. Consider the null hypothesis that the distributions of the two populations are the same. Let us suppose that, in fact, this null hypothesis is true; the combined $n + m$ observations are independent and sampled from the same population. Suppose now that you are told that one of the $n + m$ observations is equal to 10. Which of the $n + m$ observations is most likely to have taken the value 10? There is really nothing to distinguish the observations, since they are all taken from the same distribution or population. Thus, any of the $n + m$ observations is equally likely to be the one that was equal to 10. More generally, suppose that our samples are taken in a known order; for example, the first n observations come from the first population and the next m from the second. Let us suppose that the null hypothesis still holds. Suppose that you are now given the observed values in the sample, all $n + m$ of them, but not told which value was obtained from which ordered observation. Which arrangement is most likely? Since all the observations come from the same distribution, and the observations are independent, there is nothing that would tend to associate any one sequence or arrangement of the numbers with a higher probability than any other sequence. In other words, every assignment of the observed numbers to the $n + m$ observations is equally likely. This is the idea underlying a class of tests called *permutation tests*. To understand why they are called this, we need the definition of a permutation:

Definition 8.10. Given a set of $(n + m)$ objects arranged or numbered in a sequence, a *permutation* of the objects is a rearrangement of the objects into the same or a different order. The number of permutations is $(n + m)!$.

What we said above is that if the null hypothesis holds in the two-sample problem, all permutations of the numbers observed are equally likely. Let us illustrate this with a small example. Suppose that we have two observations from the first group and two observations from the second group. Suppose that we know that the four observations take on the values 3,

Table 8.5 Permutations of Four Observations

x		y		$\overline{x}-\overline{y}$	x		y		$\overline{x}-\overline{y}$
3	7	8	10	−4	7	8	3	10	1
3	7	10	8		7	8	10	3	
7	3	8	10		8	7	3	10	
7	3	10	8		8	7	10	3	
3	8	7	10	−3	7	10	3	8	3
3	8	10	7		7	10	8	3	
8	3	7	10		10	7	3	8	
8	3	10	7		10	7	8	3	
3	10	7	8	−1	8	10	3	7	4
3	10	8	7		8	10	7	3	
10	3	7	8		10	8	3	7	
10	3	8	7		10	8	7	3	

7, 8, and 10. Listed in Table 8.5 are the possible permutations where the first two observations would be considered to come from the first group and the second two from the second group. (Note that x represents the first group and y represents the second.)

If we only know the four values 3, 7, 8, and 10 but do not know in which order they came, any of the 24 possible arrangements listed above are equally likely. If we wanted to perform a two-sample test, we could generate a statistic and calculate its value for each of the 24 arrangements. We could then order the values of the statistic according to some alternative hypothesis so that the more extreme values were more likely under the alternative hypothesis. By looking at what sequence *actually occurred*, we can get a p-value for this set of data. The p-value is determined by the position of the statistic among the possible values. The p-value is the number of possibilities as extreme or more extreme than that observed divided by the number of possibilities.

Suppose, for example, that with the data above, we decided to use the difference in means between the two groups, $\overline{x}-\overline{y}$, as our test statistic. Suppose also that our alternative hypothesis is that group 1 has a larger mean than group 2. Then, if any of the last four rows of Table had occurred, the one-sided p-value would be 4/24, or 1/6. Note that this would be the most extreme finding possible. On the other hand, if the data had been 8, 7, 3, and 10, with an $\overline{x}-\overline{y}=1$, the p-value would be 12/24, or 1/2.

The tests we have been discussing are called *permutation tests*. They are possible when a permutation of all or some subset of the data is considered equally likely under the null hypothesis; the test is based on this fact. These tests are sometimes also called *conditional tests*, because the test takes some portion of the data as fixed or known. In the case above, we assume that we know the actual observed values, although we do not know in which order they occurred. We have seen an example of a conditional test before: Fisher's exact test in Chapter 6 treated the row and column totals as known; conditionally, upon that information, the test considered what happened to the entries in the table. The permutation test can be used to calculate appropriate p-values for tests such as the t-test when, in fact, normal assumptions do not hold. To do this, proceed as in the next example.

Example 8.7. Given two samples, a sample of size n of X observations and a sample of size m of Y observations, it can be shown (Problem 8.24) that the two-sample t-test is a monotone function of $\overline{x}-\overline{y}$; that is, as $\overline{x}-\overline{y}$ increases, t also increases. Thus, if we perform a permutation test on $\overline{x}-\overline{y}$, we are in fact basing our test on extreme values of the t-statistic. The illustration above is equivalent to a t-test on the four values given. Consider now the data

$$x_1 = 1.3, \qquad x_2 = 2.3, \qquad x_3 = 1.9, \qquad y_1 = 2.8, \qquad y_2 = 3.9$$

The 120 permutations $(3+2)!$ fall into 10 groups of 12 permutations with the same value of $\overline{x}-\overline{y}$ (a complete table is included in the Web appendix). The observed value of $\overline{x}-\overline{y}$ is -1.52, the lowest possible value. A one-sided test of $E(Y) < E(X)$ would have $p = 0.1 = 12/120$. The two-sided p-value is 0.2.

The Wilcoxon test may be considered a permutation test, where the values used are the ranks and not the observed values. For the Wilcoxon test we know what the values of the ranks will be; thus, one set of statistical tables may be generated that may be used for the entire sample. For the general permutation test, since the computation depends on the numbers actually observed, it cannot be calculated until we have the sample in hand. Further, the computations for large sample sizes are very time consuming. If n is equal to 20, there are over 2×10^{18} possible permutations. Thus, the computational work for permutation tests becomes large rapidly. This would appear to limit their use, but as we discuss in the next section, it is possible to sample permutations rather than evaluating every one.

We now turn to *randomization tests*. Randomization tests proceed in a similar manner to permutation tests. In general, one assumes that some aspects of the data are known. If certain aspects of the data are known (e.g., we might know the numbers that were observed, but not which group they are in), one can calculate a number of equally likely outcomes for the complete data. For example, in the permutation test, if we know the actual values, all possible permutations of the values are equally likely under the null hypothesis. In other words, it is as if a permutation were to be selected at random; the permutation tests are examples of randomization tests.

Here we consider another example. This idea is the same as that used in the signed rank test. Suppose that under the null hypothesis, the numbers observed are independent and symmetric about zero. Suppose also that we are given the absolute values of the numbers observed but not whether they are positive or negative. Take a particular number a. Is it more likely to be positive or negative? Because the distribution is symmetric about zero, it is not more likely to be either one. It is equally likely to be $+a$ or $-a$. Extending this to all the observations, every pattern of assigning pluses or minuses to our absolute values is equally likely to occur under the null hypothesis that all observations are symmetric about zero. We can then calculate the value of a test statistic for all the different patterns for pluses and minuses. A test basing the p-value on these values would be called a *randomization test*.

Example 8.8. One can perform a randomization one-sample t-test, taking advantage of the absolute values observed rather than introducing the ranks. For example, consider the first four paired observations of Example 8.2. The values are -0.0525, 0.172, 0.577, and 0.200. Assign all 16 patterns of pluses and minuses to the four absolute values (0.0525, 0.172, 0.577, and 0.200) and calculate the values of the paired or one-sample t-test. The 16 computed values, in increasing order, are -3.47, -1.63, -1.49, $\mathbf{-0.86}$, -0.46, -0.34, -0.08, -0.02, 0.02, 0.08, 0.34, 0.46, 0.86, 1.48, 1.63, and 3.47. The observed t-value (in bold type) is -0.86. It is the fourth of 16 values. The two-sided p-value is $2(4/16) = 0.5$.

*8.10 MONTE CARLO OR SIMULATION TECHNIQUES

*8.10.1 Evaluation of Statistical Significance

To compute statistical significance, we need to compare the observed values with something else. In the case of symmetry about the origin, we have seen it is possible to compare the observed value to the distribution where the plus and minus signs are independent with probability 1/2. In cases where we do not know a prior appropriate comparison distribution, as in a drug trial, the distribution without the drug is found by either using the same subjects in a crossover trial or forming a control group by a separate sample of people who are not treated with the drug. There are cases where one can conceptually write down the probability structure that would generate the distribution under the null hypothesis, but in practice could not calculate the distribution. One example of this would be the permutation test. As we mentioned previously, if there are 20 different values in the sample, there are more than 2×10^{18} different permutations. To generate them all would not be feasible, even with modern electronic computers. However, one could evaluate the particular value of the test statistic by generating a second sample from the null distribution with all permutations being equally likely. If there were some way to generate permutations randomly and compute the value of the statistic, one could take the observed statistic (thinking of this as a sample of size 1) and compare it to the randomly generated value under the null hypothesis, the second sample. One would then order the observed and generated values of the statistic and decide which values are more extreme; this would lead to a rejection region for the null hypothesis. From this, a p-value could be computed. These abstract ideas are illustrated by the following examples.

Example 8.9. As mentioned above, for fixed observed values, the two-sample t-test is a monotone function of the value of $\overline{x}-\overline{y}$, the difference in the means of the two samples. Suppose that we have the $\overline{x} - \overline{y}$ observed. One might then generate random permutations and compute the values of $\overline{x} - \overline{y}$. Suppose that we generate n such values. For a two-sided test, let us order the *absolute* values of the statistic, including both our random sample under the null hypothesis and the actual observation, giving us $n + 1$ values. Suppose that the actual observed value of the statistic from the data is the kth-order statistic, where we have ordered the absolute values from smallest to largest. Larger values tend to give more evidence against the null hypothesis of equal means. Suppose that we would reject for all observations as large as the kth-order statistic or larger. This corresponds to a p-value of $(n + 2 - k)/(n + 1)$.

One problem that we have not discussed yet is the method for generating the random permutation and $\overline{x} - \overline{y}$ values. This is usually done by computer. The computer generates random permutations by using what are called *random number generators* (see Note 8.10). A study using the generation of random quantities by computer is called a *Monte Carlo study*, for the gambling establishment at Monte Carlo with its random gambling devices and games. Note that by using Monte Carlo permutations, we can avoid the need to generate all possible permutations! This makes permutation tests feasible for large numbers of observations.

Another type of example comes about when one does not know how to compute the distribution theoretically under the null hypothesis.

Example 8.10. This example will not give all the data but will describe how a Monte Carlo test was used. In the Coronary Artery Surgery Study (CASS [1981], Alderman et al. [1982]), a study was made of the reasons people that were treated by coronary bypass surgery or medical therapy. Among 15 different institutions, it was found that many characteristics affected the assignments of patients to surgical therapy. A multivariate statistical analysis of a type described later in this book (linear discriminant analysis) was used to identify factors related to choice of therapy and to estimate the probability that someone would have surgery. It was clear that the sites differed in the percentage of people assigned to surgery, but it was also clear that the clinical sites had patient populations with different characteristics. Thus, one could not immediately conclude that the clinics had different philosophies of assignment to therapy merely by running a χ^2 test. Conceivably, the differences between clinics could be accounted for by the different characteristics of the patient populations. Using the estimated probability that each patient would or would not have surgery, the total number of surgical cases was distributed among the clinics using a Monte Carlo technique. The corresponding χ^2 test for the observed and expected values was computed for each of these randomly generated assignments under the null hypothesis of no clinical difference. This was done 1000 times. The actual observed value for the statistic turned out to be larger than any of the 1000 simulations. Thus, the estimated p-value for the significance of the conjecture that the clinics had different methods of assigning people to therapy was less than 1/1001. It was thus concluded that the clinics had different philosophies by which they assigned people to medical or surgical therapy.

We now turn to other possible uses of the Monte Carlo technique.

8.10.2 The Bootstrap

The motivation for distribution-free statistical procedures is that we need to know the distribution of a statistic when the frequency distribution F of the data is not known a priori. A very ingenious way around this problem is given by the *bootstrap*, a procedure due in its full maturity to Efron [1979], although special cases and related ideas had been around for many years.

The idea behind the bootstrap is that although we do not know F, we have a good estimate of it in the empirical frequency distribution F_n. If we can estimate the distribution of our statistic when data are sampled from F_n, we should have a good approximation to the distribution of the statistic when data are sampled from the true, unknown F. We can create data sets sampled from F_n simply by resampling the observed data: We take a sample of size n from our data set of size n (replacing the sampled observation each time). Some observations appear once, others twice, others not at all.

The bootstrap appears to be too good to be true (the name emphasizes this, coming from the concept of "lifting yourself by your bootstraps"), but both empirical and theoretical analysis confirm that it works in a fairly wide range of cases. The two main limitations are that it works only for independent observations and that it fails for certain extremely nonrobust statistics (the only simple examples being the maximum and minimum). In both cases there are more sophisticated variants of the bootstrap that relax these conditions.

Because it relies on approximating F by F_n the bootstrap is a large-sample method that is only asymptotically distribution-free, although it is successful in smaller samples than, for example, the t-test for nonnormal data. Efron and Tibshirani [1986, 1993] are excellent references; much of the latter is accessible to the nonstatistician. Davison and Hinckley [1997] is a more advanced book covering many variants on the idea of resampling. The Web appendix to this chapter links to more demonstrations and examples of the bootstrap.

Example 8.11. We illustrate the bootstrap by reexamining the confidence interval for $P[X < Y]$ generated in Example 8.5. Recall that we were comparing treadmill times for normal subjects and those with three-vessel disease. The observed $P[X < Y]$ was $15/80 = 0.1875$. In constructing a bootstrap sample we sample 8 observations from the normal and 10 from the three-vessel disease data and compute U/mn for the sample. Repeating this 1000 times gives an estimate of the distribution of $P[X < Y]$. Taking the upper and lower $\alpha/2$ percentage points of the distribution gives an approximate 95% confidence interval. In this case the confidence interval is $[0, 0.41]$. Figure 8.3 shows a histogram of the bootstrap distribution with the normal approximation from Example 8.5 overlaid on it.

Comparing this to the interval generated from the normal approximation, we see that both endpoints of the bootstrap interval are slightly higher, and the bootstrap interval is not quite symmetric about the observed value, but the two intervals are otherwise very similar. The bootstrap technique requires more computer power but is more widely applicable: It is less conservative in large samples and may be less liberal in small samples.

Related resampling ideas appear elsewhere in the book. The idea of splitting a sample to estimate the effect of a model in an unbiased manner is discussed in Chapters 11 and 13 and elsewhere. Systematically omitting part of a sample, estimating values, and testing on the omitted part is used; if one does this, say for all subsets of a certain size, a *jackknife* procedure is being used (see Efron [1982]; Efron and Tibshirani [1993]).

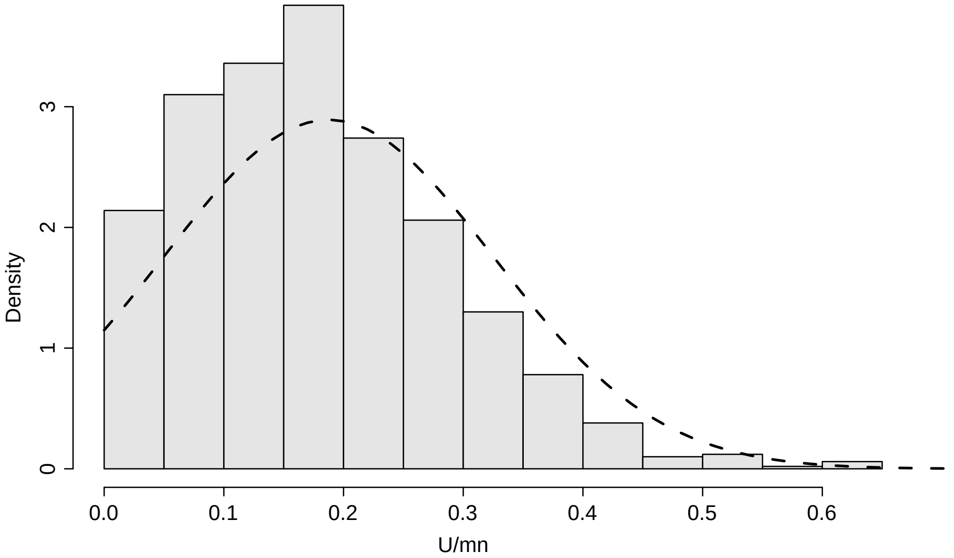

Figure 8.3 Histogram of bootstrap distribution of U/mn and positive part of normal approximation (dashed line). (Data from CASS [1981]; see Example 8.5.)

8.10.3 Empirical Evaluation of the Behavior of Statistics: Modeling and Evaluation

Monte Carlo generation on a computer is also useful for studying the behavior of statistics. For example, we know that the χ^2-statistic for contingency tables, as discussed in Chapter 7, has approximately a χ^2-distribution for large samples. But is the distribution approximately χ^2 for smaller samples? In other words, is the statistic fairly robust with respect to sample size? What happens when there are small numbers of observations in the cells? One way to evaluate small-sample behavior is a Monte Carlo study (also called a *simulation study*). One can generate multinomial samples with the two traits independent, compute the χ^2-statistic, and observe, for example, how often one would reject at the 5% significance level. The Monte Carlo simulation would allow evaluation of how large the sample needs to be for the asymptotic χ^2 critical value to be useful.

Monte Carlo simulation also provides a general method for estimating power and sample size. When designing a study one usually wishes to calculate the probability of obtaining statistically significant results under the proposed alternative hypothesis. This can be done by simulating data from the alternative hypothesis distribution and performing the planned test. Repeating this many times allows the power to be estimated. For example, if 910 of 1000 simulations give a statistically significant result, the power is estimated to be 91%. In addition to being useful when no simple formula exists for the power, the simulation approach is helpful in concentrating the mind on the important design factors. Having to simulate the possible results of a study makes it very clear what assumptions go into the power calculation.

Another use of the Monte Carlo method is to model very complex situations. For example, you might need to design a hospital communications network with many independent inputs. If you knew roughly the distribution of calls from the possible inputs, you could simulate by Monte Carlo techniques the activity of a proposed network if it were built. In this manner, you could see whether or not the network was often overloaded. As another example, you could model the hospital system of an area under the assumption of new hospitals being added and various assumptions about the case load. You could also model what might happen in catastrophic circumstances (*provided* that realistic assumptions could be made). In general, the modeling and simulation approach gives one method of evaluating how changes in an environment might affect other factors without going through the expensive and potentially catastrophic exercise of actually building whatever is to be simulated. Of course, such modeling depends *heavily* on the skill of the people constructing the model, the realism of the assumptions they make, and whether or not the probabilistic assumptions used correspond approximately to the real-life situation.

A starting reference for learning about Monte Carlo ideas is a small booklet by Hoffman [1979]. More theoretical texts are Edgington [1987] and Ripley [1987] .

*8.11 ROBUST TECHNIQUES

Robust techniques cover more than the field of nonparametric and distribution-free statistics. In general, distribution-free statistics give robust techniques, but it is possible to make more classical methods robust against certain violations of assumptions.

We illustrate with three approaches to making the sample mean robust. Another approach discussed earlier, which we shall not discuss again here, is to use the sample median as a measure of location. The three approaches are modifications of the traditional mean statistic $\overline{x}$. Of concern in computing the sample mean is the effect that an outlier will have. An observation far away from the main data set can have an enormous effect on the sample mean. One would like to eliminate or lessen the effect of such outlying and possibly spurious observations.

An approach that has been suggested is the α-trimmed mean. With the α-trimmed mean, we take some of the largest and smallest observations and drop them from each end. We then compute the usual sample mean on the data remaining.

Definition 8.11. The *α-trimmed mean* of n observations is computed as follows: Let k be the smallest integer greater than or equal to αn. Let $X_{(i)}$ be the order statistics of the sample. The α-trimmed mean drops approximately a proportion α of the observations from both ends of the distribution. That is,

$$\alpha\text{-trimmed mean} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} X_{(i)}$$

We move on to the two other ways of modifying the mean, and then illustrate all three with a data set. The second method of modifying the mean is called *Winsorization*. The α-trimmed mean drops the largest and smallest observations from the samples. In the Winsorized mean, such observations are included, but the large effect is reduced. The approach is to shrink the smallest and largest observations to the next remaining observations, and count them as if they had those values. This will become clearer with the example below.

Definition 8.12. The *α-Winsorized mean* is computed as follows. Let k be the smallest integer greater than or equal to αn. The α-Winsorized mean is

$$\alpha\text{-Winsorized mean} = \frac{1}{n}\left[(k+1)(X_{(k+1)} + X_{(n-k)}) + \sum_{i=k+2}^{n-k-1} X_{(i)}\right]$$

The third method is to weight observations differentially. In general, we would want to weight the observations at the ends or tails less and those in the middle more. Thus, we will base the weights on the order statistics where the weights for the first few order statistics and the last few order statistics are typically small. In particular, we define the weighted mean to be

$$\text{weighted mean} = \frac{\sum_{i=1}^{n} W_i X_{(i)}}{\sum_{i=1}^{n} W_i}, \qquad \text{where } W_i \geq 0$$

Problem 8.26 shows that the α-trimmed mean and the α-Winsorized mean are examples of weighted means with appropriately chosen weights.

Example 8.12. We compute the mean, median, 0.1-trimmed mean, and 0.1-Winsorized mean for the female treadmill data of Example 8.6.

$$\text{mean} = \bar{x} = \frac{30 + \cdots + 936}{20} = 708$$

$$\text{median} = \frac{X_{(10)} + X_{(11)}}{2} = 750$$

Now $0.1 \times 20 = 2$, so $k = 2$.

$$\alpha\text{-trimmed mean} = \frac{570 + \cdots + 900}{16} = 734.6$$

$$\alpha\text{-Winsorized mean} = \frac{1}{20}(3(579 + 900) + 618 + \cdots + 840) = 734.7$$

Note that the median and both robust mean estimates are considerably higher than the sample mean $\bar{x}$. This is because of the small outlier of 30.

The Winsorized mean was intended to give outlying observations the same influence on the estimate as the most extreme of the interior estimates. In fact, the trimmed mean does this and the Winsorized mean gives outlying observations rather more influence. This, combined with the simplicity of the trimmed mean, makes it more attractive.

Robust techniques apply in a much more general context than shown here, and indeed are more useful in other situations. In particular, for regression and multiple regression (subjects of subsequent chapters in this book), a large amount of statistical theory has been developed for making the procedures more robust [Huber, 1981].

*8.12 FURTHER READING AND DIRECTIONS

There are several books dealing with nonparametric statistics. Among these are Lehmann and D'Abrera [1998] and Kraft and van Eeden [1968]. Other books deal exclusively with nonparametric statistical techniques. Three that are accessible on a mathematical level suitable for readers of this book are Marascuilo and McSweeney [1977], Bradley [1968], and Siegel and Castellan [1990].

A book that gives more of a feeling for the mathematics involved at a level above this text but which does not require calculus is Hajek [1969]. Another very comprehensive text that outlines much of the theory of statistical tests but is on a somewhat more advanced mathematical level, is Hollander and Wolfe [1999]. Finally, a comprehensive text on robust methods, written at a very advanced mathematical level, is Huber [2003].

In other sections of this book we give nonparametric and robust techniques in more general settings. They may be identified by one of the words *nonparametric, distribution-free*, or *robust* in the title of the section.

NOTES

8.1 Definitions of Nonparametric and Distribution-Free

The definitions given in this chapter are close to those of Huber [2003]. Bradley [1968] states that "roughly speaking, a nonparametric test is a test which makes no hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test is one which makes no assumptions about the precise form of the sampled population."

8.2 Relative Efficiency

The statements about relative efficiency in this chapter refer to asymptotic relative efficiency [Bradley, 1968; Hollander and Wolfe, 1999; Marascuilo and McSweeney, 1977]. For two possible *estimates*, the asymptotic relative efficiency of A to B is the limit of the ratio of the variance of B to the variance of A as the sample size increases. For two possible *tests*, first select a sequence of alternatives such that as n becomes large, the power (probability of rejecting the null hypothesis) for test A converges to a fixed number greater than zero and less than 1. Let this number be C. For each member of the sequence, find sample sizes n_A and n_B such that both tests have (almost) power C. The limit of the ratio n_B to n_A is the asymptotic relative efficiency. Since the definition is for large sample sizes (asymptotic), for smaller sample sizes the efficiency may be more or less than the figures we have given. Both Bradley [1968] and Hollander and Wolfe [1999] have considerable information on the topic.

8.3 Crossover Designs for Drugs

These are subject to a variety of subtle differences. There may be carryover effects from the drugs. Changes over time—for example, extreme weather changes—may make the second part of the crossover design different than the first. Some drugs may permanently change the subjects in some way. Peterson and Fisher [1980] give many references germane to randomized clinical trials.

8.4 Signed Rank Test

The values of the ranks are known; for n observations, they are the integers $1 - n$. The only question is the sign of the observation associated with each rank. Under the null hypothesis, the sign is equally likely to be plus or minus. Further, knowing the rank of an observation based on the absolute values does not predict the sign, which is still equally likely to be plus or minus independently of the other observations. Thus, all 2^n patterns of plus and minus signs are equally likely. For $n = 2$, the four patterns are:

Ranks	1	2	1	2	1	2	1	2
Signs	−	−	+	−	−	+	+	+
S	0		1		2		3	

So $P[S \leq 0] = 1/4$, $P[S \leq 1] = 1/2$, $P[S \leq 2] = 3/4$, and $P[S \leq 3] = 1$.

8.5 Quantile Test

If the alternative hypothesis of interest is an increase in extreme values of the outcome variable, a more powerful rank test can be based on the number of values above a given threshold. That is, the outcome value X_i is recoded to 1 if it is above the threshold and 0 if it is below the threshold. This recoding reduces the data to a 2×2 table, and Fisher's exact test can be used to make the comparison (see Section 6.3). Rather than prespecifying a threshold, one could specify that the threshold was to be, say, the 90th percentile of the combined sample. Again the data would be recoded to 1 for an observation in the top 10%, 0 for other observations, giving a 2 × 2 table. It is important that either a threshold or a percentile be specified in advance. Selecting the threshold that gives the largest difference in proportions gives a test related to the Kolmogorov–Smirnov test, and when proper control of the implicit multiple comparisons is made, this test is not particularly powerful.

8.6 Transitivity

One disadvantage of the rank tests is that they are not necessarily *transitive*. Suppose that we conclude from the Mann–Whitney test that group A has larger values than group B, and group B has larger values than group C. It would be natural to assume that group A has larger values than group C, but the Mann–Whitney test could conclude the reverse—that C was larger than A. This fact is important in the theory of elections, where different ways of running elections are generally equivalent to different rank tests. It implies that candidate A could beat B, B could beat C, and C could beat A in fair two-way runoff elections, a problem noted in the late eighteenth century by Condorcet. Many interesting issues related to nontransitivity were discussed in Martin Gardner's famous "Mathematical Games" column in *Scientific American* of December 1970, October 1974, and November 1997.

The practical importance of nontransitivity is unclear. It is rare in real data, so may largely be a philosophical issue. On the other hand, it does provide a reminder that the rank-based tests are not just a statistical garbage disposal that can be used for any data whose distribution is unattractive.

8.7 Kolmogorov–Smirnov Statistic Is a Rank Statistic

We illustrate one technique used to show that the Kolmogorov–Smirnov statistic is a rank test. Looking at Figure 8.2, we could slide both curves along the x-axis without changing the value of the maximum difference, D. Since the curves are horizontal, we can stretch them along the axis (as long as the order of the jumps does not change) and not change the value of D. Place the first jump at 1, the second at 2, and so on. We have placed the jumps then at the ranks! The height of the jumps depends on the sample size. Thus, we can compute D from the ranks (and knowing which group have the rank) and the sample sizes. Thus, D is nonparametric and distribution-free.

8.8 One-Sample Kolmogorov–Smirnov Tests and One-Sided Kolmogorov–Smirnov Tests

It is possible to compare one sample to a hypothesized distribution. Let F be the empirical cumulative distribution function of a sample. Let H be a hypothesized distribution function. The statistic

$$D = \max_x |F(x) - H(x)|$$

is the one-sample statistic. If H is continuous, critical values are tabulated for this nonparametric test in the tables already cited in this chapter. An approximation to the p-value for the one-sample Kolmogorov–Smirnov test is

$$P(D > d) \leq 2e^{-2d^2/n}$$

This is conservative regardless of sample size, the value of d, the presence or absence of ties, and the true underlying distribution F, and is increasingly accurate as the p-value decreases. This approximation has been known for a long time, but the fact that it is guaranteed to be conservative is a recent, very difficult mathematical result [Massart, 1990].

The Kolmogorov–Smirnov two-sample statistic was based on the largest difference between two empirical cumulative distribution functions; that is,

$$D = \max_x |F(x) - G(x)|$$

where F and G are the two empirical cumulative distribution functions. Since the absolute value is involved, we are not differentiating between F being larger and G being larger. If we had hypothesized as an alternative that the F population took on larger values in general, F would tend to be less than G, and we could use

$$D^+ = \max_x (G(x) - F(x))$$

Such one-sided Kolmogorov–Smirnov statistics are used and tabulated. They also are nonparametric rank tests for use with one-sided alternatives.

8.9 More General Rank Tests

The theory of tests based on ranks is well developed [Hajek, 1969; Hajek and Sidak, 1999; Huber, 2003]. Consider the two-sample problem with groups of size n and m, respectively. Let $R_i (i = 1, 2, \ldots, n)$ be the ranks of the first sample. Statistics of the following form, with a a function of R_i, have been studied extensively.

$$S = \frac{1}{n} \sum_{i=1}^{n} a(R_i)$$

The $a(R_i)$ may be chosen to be efficient in particular situations. For example, let $a(R_i)$ be such that a standard normal variable has probability $R_i/(n+m+1)$ of being less than or equal to this value. Then, when the usual two-sample t-test normal assumptions hold, the relative efficiency is 1. That is, this rank test is as efficient as the t-test for large samples. This test is called the *normal scores test* or *van der Waerden test.*

8.10 Monte Carlo Technique and Pseudorandom Number Generators

The term *Monte Carlo technique* was introduced by the mathematician Stanislaw Ulam [1976] while working on the Manhattan atomic bomb project.

Computers typically do not generate random numbers; rather, the numbers are generated in a sequence by a specific computer algorithm. Thus, the numbers are called *pseudorandom numbers*. Although not random, the sequence of numbers need to appear random. Thus, they are tested in part by statistical tests. For example, a program to generate random integers from zero to nine may have a sequence of generated integers tested by the χ^2 goodness-of-fit test to see that the "probability" of each outcome is 1/10. A generator of uniform numbers on the interval (0, 1) can have its empirical distribution compared to the uniform distribution by the one-sample Kolmogorov–Smirnov test (Note 8.8). The subject of pseudorandom number generators is very deep both philosophically and mathematically. See Chaitin [1975] and Dennett [1984, Chaps. 5 and 6] for discussions of some of the philosophical issues, the former from a mathematical viewpoint.

Computer and video games use pseudorandom number generation extensively, as do computer security systems. A number of computer security failures have resulted from poor-quality pseudorandom number generators being used in encryption algorithms. One can generally assume that the generators provided in statistical packages are adequate for statistical (not cryptographic) purposes, but it is still useful to repeat complex simulation experiments with a different generator if possible. A few computer systems now have "genuine" random number generators that collect and process randomness from sources such as keyboard and disk timings.

PROBLEMS

8.1 The following data deal with the treatment of essential hypertension (*essential* is a technical term meaning that the cause is unknown; a synonym is *idiopathic*) and is from a paper by Vlachakis and Mendlowitz [1976]. Seventeen patients received treatments C, A, and B, where C is the control period, A is propranolol+phenoxybenzamine, and B is propranolol + phenoxybenzamine + hydrochlorothiazide. Each patient received C first, then either A or B, and finally, B or A. The data in Table 8.6 consist of the systolic blood pressure in the recumbent position.

Table 8.6 Blood Pressure Data for Problem 8.1

Patient	C	A	B	Patient	C	A	B
1	185	148	132	10	180	132	136
2	160	128	120	11	176	140	135
3	190	144	118	12	200	165	144
4	192	158	115	13	188	140	115
5	218	152	148	14	200	140	126
6	200	135	134	15	178	135	140
7	210	150	128	16	180	130	130
8	225	165	140	17	150	122	132
9	190	155	138				

(a) Take the differences between the systolic blood pressures on treatments A and C. Use the sign test to test for a treatment A effect (two-sided test; give the p-value).

(b) Take the differences between treatments B and C. Use the sign test to test for a treatment B effect (one-sided test; give the p-value).

(c) Take the differences between treatments B and A. Test for a treatment difference using the sign test (two-sided test; give the p-value).

8.2 Several population studies have demonstrated an inverse correlation of sudden infant death syndrome (SIDS) rate with birthweight. The occurrence of SIDS in one of a pair of twins provides an opportunity to test the hypothesis that birthweight is a major determinant of SIDS. The set of data in Table 8.7 was collected by D. R. Peterson of the

Table 8.7 Birthweight Data for Problem 8.2

Dizygous Twins		Monozygous Twins		Dizygous Twins		Monozygous Twins	
SIDS	Non-SIDS	SIDS	Non-SIDS	SIDS	Non-SIDS	SIDS	Non-SIDS
1474	2098	1701	1956	2381	2608	1956	1588
3657	3119	2580	2438	2892	2693	2296	2183
3005	3515	2750	2807	2920	3232	3232	2778
2041	2126	1956	1843	3005	3005	1446	2268
2325	2211	1871	2041	2268	2325	1559	1304
2296	2750	2296	2183	3260	3686	2835	2892
3430	3402	2268	2495	3260	2778	2495	2353
3515	3232	2070	1673	2155	2552	1559	2466
1956	1701	1786	1843	2835	2693		
2098	2410	3175	3572	2466	1899		
3204	2892	2495	2778	3232	3714		

Department of Epidemiology, University of Washington, consists of the birthweights of each of 22 dizygous twins and each of 19 monozygous twins.

(a) For the dizygous twins test the alternative hypothesis that the SIDS child of each pair has the lower birthweight by taking differences and using the sign test. Find the one-sided p-value.

(b) As in part (a), but do the test for the monozygous twins.

(c) As in part (a), but do the test for the combined data set.

8.3 The following data are from Dobson et al. [1976]. Thirty-six patients with a confirmed diagnosis of phenylketonuria (PKU) were identified and placed on dietary therapy before reaching 121 days of age. The children were tested for IQ (Stanford–Binet test) between the ages of 4 and 6; subsequently, their normal siblings of closest age were also tested with the Stanford–Binet. The 15 pairs shown in Table 8.8 are the first 15 listed in the paper. The null hypothesis is that the PKU children, on average, have the same IQ as their siblings. Using the sign test, find the two-sided p-value for testing against the alternative hypothesis that the IQ levels differ.

Table 8.8 PKU/IQ Data for Problem 8.3

Pair	IQ of PKU Case	IQ of Sibling	Pair	IQ of PKU Case	IQ of Sibling
1	89	77	9	110	88
2	98	110	10	90	91
3	116	94	11	76	99
4	67	91	12	71	93
5	128	122	13	100	104
6	81	94	14	108	102
7	96	121	15	74	82
8	116	114			

8.4 Repeat Problem 8.1 using the signed rank test rather than the sign test. Test at the 0.05 significance level.

8.5 Repeat Problem 8.2, parts (a) and (b), using the signed rank test rather than the sign test. Test at the 0.05 significance level.

8.6 Repeat Problem 8.3 using the signed rank test rather than the sign test. Test at the 0.05 significance level.

8.7 Bednarek and Roloff [1976] deal with the treatment of apnea (a transient cessation of breathing) in premature infants using a drug called aminophylline. The variable of interest, "average number of apneic episodes per hour," was measured before and after treatment with the drug. An episode was defined as the absence of spontaneous breathing for more than 20 seconds, or less if associated with bradycardia or cyanosis. Table 8.9 details the response of 13 patients to aminophylline treatment at 16 hours compared with 24 hours before treatment (in apneic episodes per hour).

(a) Use the sign test to examine a treatment effect (give the two-sided p-value).

(b) Use the signed rank test to examine a treatment effect (two-sided test at the 0.05 significance level).

Table 8.9 Before/After Treatment Data for Problem 8.7

Patient	24 Hours Before	16 Hours After	Before–After (Difference)
1	1.71	0.13	1.58
2	1.25	0.88	0.37
3	2.13	1.38	0.75
4	1.29	0.13	1.16
5	1.58	0.25	1.33
6	4.00	2.63	1.37
7	1.42	1.38	0.04
8	1.08	0.50	0.58
9	1.83	1.25	0.58
10	0.67	0.75	−0.08
11	1.13	0.00	1.13
12	2.71	2.38	0.33
13	1.96	1.13	0.83

8.8 The following data from Schechter et al. [1973] deal with sodium chloride preference as related to hypertension. Two groups, 12 normal and 10 hypertensive subjects, were isolated for a week and compared with respect to Na^+ intake. The average daily Na^+ intakes are listed in Table 8.10. Compare the average daily Na^+ intake of the hypertensive subjects with that of the normal volunteers by means of the Wilcoxon two-sample test at the 5% significance level.

Table 8.10 Sodium Data for Problem 8.8

Normal	Hypertensive	Normal	Hypertensive
10.2	92.8	45.8	34.7
2.2	54.8	63.6	62.2
0.0	51.6	1.8	11.0
2.6	61.7	0.0	39.1
0.0	250.8	3.7	
43.1	84.5	0.0	

8.9 During July and August 1976, a large number of Legionnaires attending a convention died of a mysterious and unknown cause. Epidemiologists have talked of "an outbreak of Legionnaires' disease." Chen et al. [1977] examined the hypothesis of nickel contamination as a toxin. They examined the nickel levels in the lungs of nine cases and nine controls. The authors point out that contamination at autopsy is a possibility. The data are as follows (μg per 100 g dry weight):

Legionnaire Cases	65	24	52	86	120	82	399	87	139
Control Cases	12	10	31	6	5	5	29	9	12

Note that there was no attempt to match cases and controls. Use the Wilcoxon test at the one-sided 5% level to test the null hypothesis that the numbers are samples from similar populations.

Table 8.11 Plasma iPGE Data for Problem 8.10

Patient Number	Mean Plasma iPGE (pg/mL)	Mean Serum Calcium (ml/dL)
	Patients with Hypercalcemia	
1	500	13.3
2	500	11.2
3	301	13.4
4	272	11.5
5	226	11.4
6	183	11.6
7	183	11.7
8	177	12.1
9	136	12.5
10	118	12.2
11	60	18.0
	Patients without Hypercalcemia	
12	254	10.1
13	172	9.4
14	168	9.3
15	150	8.6
16	148	10.5
17	144	10.3
18	130	10.5
19	121	10.2
20	100	9.7
21	88	9.2

8.10 Robertson et al. [1976] discuss the level of plasma prostaglandin E (iPGE in pg/mL) in patients with cancer with and without hypercalcemia. The data are given in Table 8.11. Note that the variables are "mean plasma iPGE" and "mean serum Ca" levels; presumably more than one assay was carried out for each patient's level. The number of such tests for each patient is not indicated, nor is the criterion for the number. Using the Wilcoxon two-sample test, test for differences between the two groups in:

(a) Mean plasma iPGE.

(b) Mean serum Ca.

8.11 Sherwin and Layfield [1976] present data about protein leakage in the lungs of male mice exposed to 0.5 part per million of nitrogen dioxide (NO_2). Serum fluorescence data were obtained by sacrificing animals at various intervals. Use the two-sided Wilcoxon test, 0.05 significance level, to look for differences between controls and exposed mice.

(a) At 10 days:

Controls	143	169	95	111	132	150	141
Exposed	152	83	91	86	150	108	78

(b) At 14 days:

Controls	76	40	119	72	163	78
Exposed	119	104	125	147	200	173

8.12 Using the data of Problem 8.8:

(a) Find the value of the Kolmogorov–Smirnov statistic.

(b) Plot the two empirical distribution functions.

(c) Do the curves differ at the 5% significance level? For sample sizes 10 and 12, the 10%, 5%, and 1% critical values for mnD are 60, 66, and 80, respectively.

8.13 Using the data of Problem 8.9:

(a) Find the value of the Kolmogorov–Smirnov statistic.

(b) Do you reject the null hypothesis at the 5% level? For $m = 9$ and $n = 9$, the 10%, 5%, and 1% critical values of mnD are 54, 54, and 63, respectively.

8.14 Using the data of Problem 8.10:

(a) Find the value of the Kolmogorov–Smirnov statistic for both variables.

(b) What can you say about the p-value? For $m = 10$ and $n = 11$, the 10%, 5%, and 1% critical values of mnD are 57, 60, and 77, respectively.

8.15 Using the data of Problem 8.11:

(a) Find the value of the Kolmogorov–Smirnov statistic.

(b) Do you reject at 10%, 5%, and 1%, respectively? Do this for parts (a) and (b) of Problem 8.11. For $m = 7$ and $n = 7$, the 10%, 5%, and 1% critical values of mnD are 35, 42, and 42, respectively. The corresponding critical values for $m = 6$ and $n = 6$ are 30, 30, and 36.

8.16 Test at the 0.05 significance level for a significant improvement with the cream treatment of Example 8.2.

(a) Use the sign test.

(b) Use the signed rank test.

(c) Use the t-test.

8.17 Use the expression of colostrum data of Example 8.2, and test at the 0.10 significance level the null hypothesis of no treatment effect.

(a) Use the sign test.

(b) Use the signed rank test.

(c) Use the usual t-test.

8.18 Test the null hypothesis of no treatment difference from Example 8.2 using each of the tests in parts (a), (b), and (c).

(a) The Wilcoxon two-sample test.

(b) The Kolmogorov–Smirnov two-sample test. For $m = n = 19$, the 20%, 10%, 5%, 1%, and 0.1% critical values for mnD are 133, 152, 171, 190, and 228, respectively.

(c) The two-sample t-test.

Compare the two-sided p-values to the extent possible. Using the data of Example 8.2, examine each treatment.

(d) Nipple-rolling vs. masse cream.

(e) Nipple-rolling vs. expression of colostrum.

(f) Masse cream vs. expression of colostrum.

8.19 As discussed in Chapter 3, Winkelstein et al. [1975] studied systolic blood pressures of three groups of Japanese men: native Japanese, first-generation immigrants to the United States (Issei), and second-generation Japanese in the United States (Nisei). The data are listed in Table 8.12. Use the asymptotic Wilcoxon two-sample statistic to test:

(a) Native Japanese vs. California Issei.

(b) Native Japanese vs. California Nisei.

(c) California Issei vs. California Nisei.

Table 8.12 Blood Pressure Data for Problem 8.19

Blood Pressure (mmHg)	Native Japanese	Issei	Nisei
<106	218	4	23
106–114	272	23	132
116–124	337	49	290
126–134	362	33	347
136–144	302	41	346
146–154	261	38	202
156–164	166	23	109
>166	314	52	112

***8.20** Rascati et al. [2001] report a study of medical costs for children with asthma in which children prescribed steroids had a higher mean cost than other children, but lower costs according to a Wilcoxon rank-sum test. How can this happen, and what conclusions should be drawn?

***8.21** An outlier is an observation far from the rest of the data. This may represent valid data or a mistake in experimentation, data collection, or data entry. At any rate, a few outlying observations may have an extremely large effect. Consider a one-sample t-test of mean zero based on 10 observations with

$$\overline{x} = 10 \quad \text{and} \quad s^2 = 1$$

Suppose now that one observation of value x is added to the sample.

(a) Show that the value of the new sample mean, variance, and t-statistic are

$$\overline{x} = \frac{100 + x}{11}$$

$$s^2 = \frac{10x^2 - 200x + 1099}{11 \times 10}$$

$$t = \frac{100 + x}{\sqrt{x^2 - 20x + 109.9}}$$

***(b)** Graph t as a function of x.

(c) For which values of x would one reject the null hypothesis of mean zero? What does the effect of an outlier (large absolute value) do in this case?

(d) Would you reject the null hypothesis without the outlier?

(e) What would the graph look like for the Wilcoxon signed rank test? For the sign test?

***8.22** Using the ideas of Note 8.4 about the signed rank test, verify the values shown in Table 8.13 when $n = 4$.

Table 8.13 Signed-Rank Test Data for Problem 8.23

s	$P[S \le s]$	s	$P[S \le s]$
0	0.062	6	0.688
1	0.125	7	0.812
2	0.188	8	0.875
3	0.312	9	0.938
4	0.438	10	1.000
5	0.562		

Source: Owen [1962]; by permission of Addison-Wesley Publishing Company.

***8.23** The Wilcoxon two-sample test depends on the fact that under the null hypothesis, if two samples are drawn without ties, all $\binom{n+m}{n}$ arrangements of the n ranks from the first sample, and the m ranks from the second sample, are equally likely. That is, if $n = 1$ and $m = 2$, the three arrangements

$$\begin{array}{llll} \mathbf{1} & 2 & 3; & W = 1 \\ 1 & \mathbf{2} & 3; & W = 2 \\ 1 & 2 & \mathbf{3}; & W = 3 \end{array}$$

are equally likely. Here, the rank from population 1 appears in bold type.

(a) If $n = 2$ and $m = 4$, graph the distribution function of the Wilcoxon two-sample statistic when the null hypothesis holds.

(b) Find $E(W)$. Does it agree with equation (5)?

(c) Find var(W). Does it agree with equation (6)?

***8.24** (Permutation Two-Sample t-Test) To use the permutation two-sample t-test, the text (in Section *8.9) used the fact that for $n + m$ fixed values, the t-test was a monotone function of $\overline{x} - \overline{y}$. To show this, prove the following equality:

$$t = \frac{1}{\sqrt{\dfrac{(n+m)\left(\sum_i x_i^2 + \sum_i y_i^2\right) - \left(\sum_i x_i + \sum_i y_i\right)^2 - nm(\overline{x} - \overline{y})^2}{nm(n+m-2)(\overline{x} - \overline{y})^2}}}$$

Note that the first two terms in the numerator of the square root are constant for all permutations, so t is a function of $\overline{x} - \overline{y}$.

***8.25** (One-Sample Randomization t-Test) For the randomization one-sample t-test, the paired x_i and y_i values give $\overline{x}-\overline{y}$ values. Assume that the $|x_i - y_i|$ are known but the signs are random, independently $+$ or $-$ with probability 1/2. The $2^n (i = 1, 2, \ldots, n)$ patterns of pluses and minuses are equally likely.

(a) Show that the one-sample t-statistic is a monotone function of $\overline{x-y}$ when the $|x_i - y_i|$ are known. Do this by showing that

$$t = \frac{\overline{x-y}}{\sqrt{\left[-n(\overline{x-y})^2 + \sum_i (x_i - y_i)^2\right]/n(n-1)}}$$

(b) For the data

i	X_i	Y_i
1	1	2
2	3	1
3	1	5

compute the eight possible randomization values of t. What is the two-sided randomization p-value for the t observed?

***8.26** (Robust Estimation of the Mean) Show that the α-trimmed mean and the α-Winsorized mean are weighted means by explicitly showing the weights W_i that are given the two means.

***8.27** (Robust Estimation of the Mean)

(a) For the combined data for SIDS in Problem 8.2, compute (**i**) the 0.05 trimmed mean; (**ii**) the 0.05 Winsorized mean; (**iii**) the weighted mean with weights $W_i = i(n + 1 - i)$, where n is the number of observations.

(b) The same as in Problem 8.27(a), but do this for the non-SIDS twins.

REFERENCES

Alderman, E., Fisher, L. D., Maynard, C., Mock, M. B., Ringqvist, I., Bourassa, M. G., Kaiser, G. C., and Gillespie, M. J. [1982]. Determinants of coronary surgery in a consecutive patient series from geographically dispersed medical centers: the Coronary Artery Surgery Study. *Circulation*, **66**: 562–568.

Bednarek, E., and Roloff, D. W. [1976]. Treatment of apnea of prematurity with aminophylline. *Pediatrics*, **58**: 335–339.

Beyer, W. H. (ed.) [1990]. *CRC Handbook of Tables for Probability and Statistics*. 2nd ed. CRC Press, Boca Raton, FL.

Bradley, J. V. [1968]. *Distribution-Free Statistical Tests*. Prentice Hall, Englewood Cliffs, NJ.

Brown, M. S., and Hurlock, J. T. [1975]. Preparation of the breast for breast-feeding. *Nursing Research*, **24**: 448–451.

CASS [1981]. (Principal investigators of CASS and their associates; Killip, T. (ed.); Fisher, L. D., and Mock, M. (assoc. eds.) National Heart, Lung and Blood Institute Coronary Artery Surgery Study. *Circulation*, **63**: part II, I–1 to I–81. Used with permission from the American Heart Association.

Chaitin, G. J. [1975]. Randomness and mathematical proof, *Scientific American*, **232**(5): 47–52.

Chen, J. R., Francisco, R. B., and Miller, T. E. [1977]. Legionnaires' disease: nickel levels. *Science*, **196**: 906–908.

Church, J. D., and Harris, B. [1970]. The estimation of reliability from stress–strength relationships. *Technometrics*, **12**: 49–54.

Davison, A. C., and Hinckley, D. V. [1997]. *Bootstrap Methods and Their Application*. Cambridge University Press, New York.

Dennett, D. C. [1984]. *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press, Cambridge, MA.

Dobson, J. C., Kushida, E., Williamson, M., and Friedman, E. [1976]. Intellectual performance of 36 phenylketonuria patients and their nonaffected siblings. *Pediatrics*, **58**: 53–58.

Edgington, E. S. [1995]. *Randomization Tests*, 3rd ed. Marcel Dekker, New York.

Efron, B. [1979]. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**: 1–26.

Efron, B. [1982]. *The Jackknife, Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Efron, B., and Tibshirani, R. [1986]. The bootstrap (with discussion). *Statistical Science*, **1**: 54–77.

Efron, B., and Tibshirani, R. [1993]. *An Introduction to the Bootstrap*. Chapman & Hall, London.

Hajek, J. [1969]. *A Course in Nonparametric Statistics*. Holden-Day, San Francisco.

Hajek, J., and Sidak, Z. [1999]. *Theory of Rank Tests*. 2nd ed. Academic Press, New York.

Hoffman, D. T. [1979]. *Monte Carlo: The Use of Random Digits to Simulate Experiments*. Models and monographs in undergraduate mathematics and its Applications, Unit 269, EDC/UMAP, Newton, MA.

Hollander, M., and Wolfe, D. A. [1999]. *Nonparametric Statistical Methods*, 2nd ed. Wiley, New York.

Huber, P. J. [2003]. *Robust Statistics*. Wiley, New York.

Johnson, R. A., Verill, S., and Moore D. H. [1987]. Two-sample rank tests for detecting changes that occur in a small proportion of the treated population. *Biometrics*, **43**: 641–655

Kraft, C. H., and van Eeden, C. [1968]. *A Nonparametric Introduction to Statistics*. Macmillan, New York.

Lehmann, E. L., and D'Abrera, H. J. M. [1998]. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. [2002]. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, **23**: 151–169.

Marascuilo, L. A., and McSweeney, M. [1977]. *Nonparametric and Distribution-Free Methods for the Social Sciences*. Brooks/Cole, Scituate, MA.

Massart, P. [1990]. The tight constant in the Dvoretsky-Kiefer-Wolfowitz inequality. *Annals of Probability*, **18**: 897–919.

Odeh, R. E., Owen, D. B., Birnbaum, Z. W., and Fisher, L. D. [1977]. *Pocket Book of Statistical Tables*. Marcel Dekker, New York.

Owen, D. B. [1962]. *Handbook of Statistical Tables*. Addison-Wesley, Reading, MA.

Peterson, A. P., and Fisher, L. D. [1980]. Teaching the principles of clinical trials design. *Biometrics*, **36**: 687–697.

Rascati, K. L., Smith, M. J., and Neilands, T. [2001]. Dealing with skewed data: an example using asthma-related costs of Medicaid clients. *Clinical Therapeutics*, **23**: 481–498.

Ripley B. D. [1987]. *Stochastic Simulation*. Wiley, New York.

Robertson, R. P., Baylink, D. J., Metz, S. A., and Cummings, K. B. [1976]. Plasma prostaglandin in patients with cancer with and without hypercalcemia. *Journal of Clinical Endocrinology and Metabolism*, **43**: 1330–1335.

Schechter, P. J., Horwitz, D., and Henkin, R. I. [1973]. Sodium chloride preference in essential hypertension. *Journal of the American Medical Association*, **225**: 1311–1315.

Sherwin, R. P., and Layfield, L. J. [1976]. Protein leakage in the lungs of mice exposed to 0.5 ppm nitrogen dioxide: a fluorescence assay for protein. *Archives of Environmental Health*, **31**: 116–118.

Siegel, S., and Castellan, N. J., Jr. [1990]. *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. McGraw-Hill, New York.

Ulam, S. M. [1976]. *Adventures of a Mathematician*. Charles Scribner's Sons, New York.

U.S. EPA [1994]. *Statistical Methods for Evaluating the Attainment of Cleanup Standards*, Vol. 3, *Reference-Based Standards for Soils and Solid Media*. EPA/600/R-96/005. Office of Research and Development, U.S. EPA, Washington, DC.

Vlachakis, N. D., and Mendlowitz, M. [1976]. Alpha- and beta-adrenergic receptor blocking agents combined with a diuretic in the treatment of essential hypertension. *Journal of Clinical Pharmacology*, **16**: 352–360.

Winkelstein, W., Jr., Kazan, A., Kato, H., and Sachs, S. T. [1975]. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii, and California: blood pressure distributions. *American Journal of Epidemiology*, **102**: 502–513.