1

DESCRIPTIVE METHODS FOR CATEGORICAL DATA

Most introductory textbooks in statistics and biostatistics start with methods for summarizing and presenting continuous data. We have decided, however, to adopt a different starting point because our focused areas are in the biomedical sciences, and health decisions are frequently based on proportions, ratios, or rates. In this first chapter we will see how these concepts appeal to common sense, and learn their meaning and uses.

1.1 PROPORTIONS

Many outcomes can be classified as belonging to one of two possible categories: presence and absence, nonwhite and white, male and female, improved and nonimproved. Of course, one of these two categories is usually identified as of primary interest: for example, presence in the presence and absence classification, nonwhite in the white and nonwhite classification. We can, in general, relabel the two outcome categories as positive (+) and negative (−). An outcome is *positive* if the primary category is observed and is *negative* if the other category is observed.

It is obvious that, in the summary to characterize observations made on a group of people, the number x of positive outcomes is not sufficient; the group size n, or total number of observations, should also be recorded. The number x tells us very little and becomes meaningful only after adjusting for the size n of the group; in other words, the two figures x and n are often combined into a *statistic*, called a *proportion*:

$$p = \frac{x}{n}.$$

Introductory Biostatistics, Second Edition. Chap T. Le and Lynn E. Eberly.

Companion website: www.wiley.com/go/Le/Biostatistics

The term *statistic* means a summarized quantity from observed data. Clearly, $0 \le p \le 1$. This proportion p is sometimes expressed as a percentage and is calculated as follows:

$$\text{percentage}(\%) = \frac{x}{n}(100).$$

Example 1.1
A study published by the Urban Coalition of Minneapolis and the University of Minnesota Adolescent Health Program surveyed 12 915 students in grades 7–12 in Minneapolis and St. Paul public schools. The report stated that minority students, about one-third of the group, were much less likely to have had a recent routine physical checkup. Among Asian students, 25.4% said that they had not seen a doctor or a dentist in the last two years, followed by 17.7% of Native Americans, 16.1% of blacks, and 10% of Hispanics. Among whites, it was 6.5%.

Proportion is a number used to describe a group of people according to a *dichotomous*, or *binary, characteristic* under investigation. It is noted that characteristics with multiple categories can have a proportion calculated per category, or can be dichotomized by pooling some categories to form a new one, and the concept of proportion applies. The following are a few illustrations of the use of proportions in the health sciences.

1.1.1 Comparative Studies

Comparative studies are intended to show possible differences between two or more groups; Example 1.1 is such a typical comparative study. The survey cited in Example 1.1 also provided the following figures concerning boys in the group who use tobacco at least weekly. Among Asians, it was 9.7%, followed by 11.6% of blacks, 20.6% of Hispanics, 25.4% of whites, and 38.3% of Native Americans.

In addition to surveys that are cross-sectional, as seen in Example 1.1, data for comparative studies may come from different sources; the two fundamental designs being retrospective and prospective. *Retrospective studies* gather past data from selected cases and controls to determine differences, if any, in *exposure* to a suspected *risk factor*. These are commonly referred to as *case–control studies*; each such study is focused on a particular disease. In a typical case–control study, cases of a specific disease are ascertained as they arise from population-based registers or lists of hospital admissions, and controls are sampled either as disease-free persons from the population at risk or as hospitalized patients having a diagnosis other than the one under study. The advantages of a retrospective study are that it is economical and provides answers to research questions relatively quickly because the cases are already available. Major limitations are due to the inaccuracy of the exposure histories and uncertainty about the appropriateness of the control sample; these problems sometimes hinder retrospective studies and make them less preferred than prospective studies. The following is an example of a retrospective study in the field of occupational health.

Example 1.2
A case–control study was undertaken to identify reasons for the exceptionally high rate of lung cancer among male residents of coastal Georgia. Cases were identified from these sources:

1. Diagnoses since 1970 at the single large hospital in Brunswick;
2. Diagnoses during 1975–1976 at three major hospitals in Savannah;
3. Death certificates for the period 1970–1974 in the area.

Controls were selected from admissions to the four hospitals and from death certificates in the same period for diagnoses other than lung cancer, bladder cancer, or chronic lung cancer. Data are tabulated separately for smokers and nonsmokers in Table 1.1. The exposure under investigation, "shipbuilding," refers to employment in shipyards during World War II. By using a separate tabulation, with the first half of the table for nonsmokers and the second half for smokers, we treat *smoking* as a potential confounder. A *confounder* is a factor, an exposure by itself, not under investigation but related to the disease (in this case, lung cancer) and the exposure (shipbuilding); previous studies have linked smoking to lung cancer, and construction workers are more likely to be smokers. The term *exposure* is used here to emphasize that employment in shipyards is a suspected *risk* factor; however, the term is also used in studies where the factor under investigation has beneficial effects.

In an examination of the smokers in the data set in Example 1.2, the numbers of people employed in shipyards, 84 and 45, tell us little because the sizes of the two groups, cases and controls, are different. Adjusting these absolute numbers for the group sizes (397 cases and 315 controls), we have:

1. For the smoking controls,

$$\text{proportion with exposure} = \frac{45}{315} = 0.143 \text{ or } 14.3\%.$$

2. For the smoking cases,

$$\text{proportion with exposure} = \frac{84}{397} = 0.212 \text{ or } 21.2\%.$$

TABLE 1.1

Smoking	Shipbuilding	Cases	Controls
No	Yes	11	35
	No	50	203
Yes	Yes	84	45
	No	313	270

The results reveal different exposure histories: the proportion in shipbuilding among cases was higher than that among controls. It is *not* in any way conclusive proof, but it is a good *clue*, indicating a possible *relationship* between the disease (lung cancer) and the exposure (shipbuilding).

Similar examination of the data for nonsmokers shows that, by taking into consideration the numbers of cases and controls, we have the following figures for shipbuilding employment:

1. For the non-smoking controls,

$$\text{proportion with exposure} = \frac{35}{238} = 0.147 \quad \text{or} \quad 14.7\%.$$

2. For the non-smoking cases,

$$\text{proportion with exposure} = \frac{11}{61} = 0.180 \text{ or } 18.0\%.$$

The results for non-smokers also reveal different exposure histories: the proportion in shipbuilding among cases was again higher than that among controls.

The analyses above also show that the case-control difference in the proportions with the exposure among smokers, that is,

$$21.2 - 14.3 = 6.9\%,$$

is different from the case-control difference in the proportions with the exposure among nonsmokers, which is:

$$18.0 - 14.7 = 3.3\%.$$

The differences, 6.9% and 3.3%, are *measures* of the strength of the relationship between the disease and the exposure, one for each of the two strata: the two groups of smokers and nonsmokers, respectively. The calculation above shows that the possible effects of employment in shipyards (as a suspected risk factor) are different for smokers and nonsmokers. This difference of differences, if confirmed, is called a *three-term interaction* or *effect modification*, where smoking alters the effect of employment in shipyards as a risk for lung cancer. In that case, *smoking* is not only a confounder, it is an *effect modifier*, which modifies the effects of shipbuilding (on the possibility of having lung cancer).

Another illustration is provided in the following example concerning glaucomatous blindness.

TABLE 1.2

	Population	Cases	Cases per 100 000
White	32 930 233	2832	8.6
Nonwhite	3 933 333	3227	82.0

Example 1.3
Counts of persons registered blind from glaucoma are listed in Table 1.2.

For these *disease registry data*, direct calculation of a proportion results in a very tiny fraction, that is, the number of cases of the disease per person at risk. For convenience, in Table 1.2, this is multiplied by 100 000, and hence the result expresses the number of cases per 100 000 people. This data set also provides an example of the use of proportions as disease *prevalence*, which is defined as:

$$\text{prevalence} = \frac{\text{number of diseased persons at the time of investigation}}{\text{total number of persons examined}}.$$

Disease prevalence and related concepts are discussed in more detail in Section 1.2.2.

For blindness from glaucoma, calculations in Example 1.3 reveal a striking difference between the races: The blindness prevalence among nonwhites was over eight times that among whites. The number "100 000" was selected arbitrarily; any power of 10 would be suitable so as to obtain a result between 1 and 100, sometimes between 1 and 1000; it is easier to state the result "82 cases per 100 000" than to say that the prevalence is 0.00082.

1.1.2 Screening Tests

Other uses of proportions can be found in the evaluation of *screening tests* or *diagnostic procedures*. Following these procedures, using clinical observations or laboratory techniques, people are classified as healthy or as falling into one of a number of disease categories. Such tests are important in medicine and epidemiologic studies and may form the basis of early interventions. Almost all such tests are imperfect, in the sense that healthy persons will occasionally be classified wrongly as being ill, while some people who are really ill may fail to be detected. That is, *misclassification* is unavoidable. Suppose that each person in a large population can be classified as truly positive or negative for a particular disease; this true diagnosis may be based on more refined methods than are used in the test, or it may be based on evidence that emerges after the passage of time (e.g., at autopsy). For each class of people, diseased and healthy, the test is applied, with the results depicted in Figure 1.1.

The two proportions fundamental to evaluating diagnostic procedures are sensitivity and specificity. *Sensitivity* is the proportion of diseased people detected as positive by the test:

$$\text{sensitivity} = \frac{\text{number of diseased persons who test positive}}{\text{total number of diseased persons}}.$$

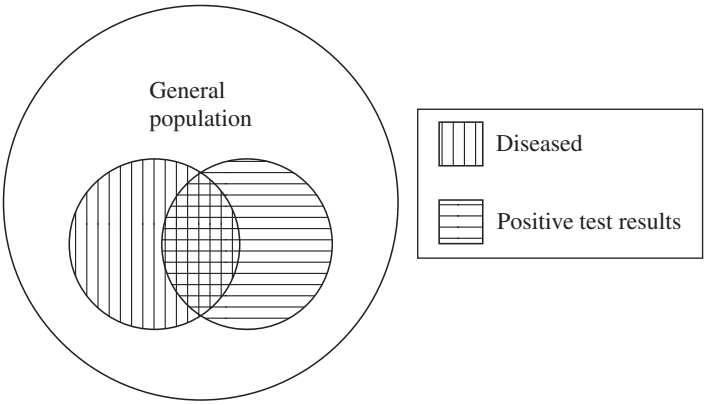

FIGURE 1.1 Graphical display of a screening test.

The corresponding errors are *false negatives*. *Specificity* is the proportion of healthy people detected as negative by the test:

$$\text{specificity} = \frac{\text{number of healthy persons who test negative}}{\text{total number of healthy persons}}.$$

The corresponding errors are *false positives*.

Clearly, it is desirable that a test or screening procedure be highly sensitive and highly specific. However, the two types of errors go in opposite directions; for example, an effort to increase sensitivity may lead to more false positives, and vice versa.

Example 1.4

A cytological test was undertaken to screen women for cervical cancer. Consider a group of 24 103 women consisting of 379 women whose cervices are abnormal (to an extent sufficient to justify concern with respect to possible cancer) and 23 724 women whose cervices are acceptably healthy. A test was applied and results are tabulated in Table 1.3. (This study was performed with a rather old test and is used here only for illustration.)

TABLE 1.3

	Test		
True	−	+	Total
−	23 362	362	23 724
+	225	154	379

The calculations

$$\text{sensitivity} = \frac{154}{379} = 0.406 \text{ or } 40.6\%$$

$$\text{specificity} = \frac{23\ 362}{23\ 724} = 0.985 \text{ or } 98.5\%$$

show that the test is highly specific (98.5%) but not very sensitive (40.6%); among the 379 women with the disease, more than half (59.4%) had false negatives. The implications of the use of this test are:

1. If a woman without cervical cancer is tested, the result would almost surely be negative, *but*
2. If a woman with cervical cancer is tested, the chance is that the disease would go undetected because 59.4% of these cases would result in false negatives.

Finally, it is important to note that throughout this section, proportions have been defined so that both the numerator and the denominator are counts or frequencies, and the numerator corresponds to a subgroup of the larger group involved in the denominator, resulting in a number between 0 and 1 (or between 0 and 100%). It is straightforward to generalize this concept for use with characteristics having more than two outcome categories; for each category we can define a proportion, and these category-specific proportions add up to 1 (or 100%).

Example 1.5
An examination of the 668 children reported living in crack/cocaine households shows 70% blacks, followed by 18% whites, 8% Native Americans, and 4% other or unknown.

1.1.3 Displaying Proportions

Perhaps the most effective and most convenient way of presenting data, especially discrete data, is through the use of graphs. Graphs convey the information, the general patterns in a set of data, at a single glance. Therefore, graphs are often easier to read than tables; the most informative graphs are simple and self-explanatory. Of course, to achieve that objective, graphs should be constructed carefully. Like tables, they should be clearly labeled and units of measurement and/or magnitude of quantities should be included. Remember that graphs must tell their own story; they should be complete in themselves and require little or no additional explanation.

Bar Charts Bar charts are a very popular type of graph used to display several proportions for quick comparison. In applications suitable for bar charts, there are several groups and we investigate one binary characteristic. In a bar chart, the various

groups are represented along the horizontal axis; they may be arranged alphabetically, by the size of their proportions, or on some other rational basis. A vertical bar is drawn above each group such that the height of the bar is the proportion associated with that group. The bars should be of equal width and should be separated from one another so as not to imply continuity.

Example 1.6
We can present the data set on children without a recent physical checkup (Example 1.1) by a bar chart, as shown in Figure 1.2.

Pie Charts Pie charts are another popular type of graph. In applications suitable for pie charts, there is only one group but we want to decompose it into several categories. A pie chart consists of a circle; the circle is divided into wedges that correspond to the magnitude of the proportions for various categories. A pie chart shows the differences between the sizes of various categories or subgroups as a decomposition of the total. It is suitable, for example, for use in presenting a budget, where we can easily see the difference between United States expenditures on health care and defense. In other words, a bar chart is a suitable graphic device when we have several groups, each associated with a different proportion; whereas a pie chart is more suitable when we have one group that is divided into several categories. The proportions of various categories in a pie chart should add up to 100%. Like bar charts, the categories in a pie chart are usually arranged by the size of the proportions. They may also be arranged alphabetically or on some other rational basis.

Example 1.7
We can present the data set on children living in crack households (Example 1.5) by a pie chart as shown in Figure 1.3.

Another example of the pie chart's use is for presenting the proportions of deaths due to different causes.

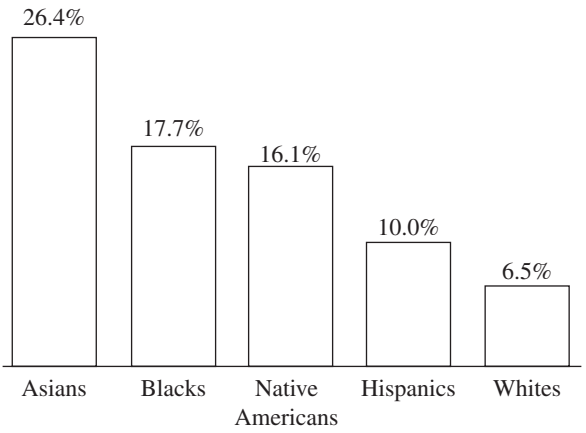

FIGURE 1.2 Children without a recent physical checkup.

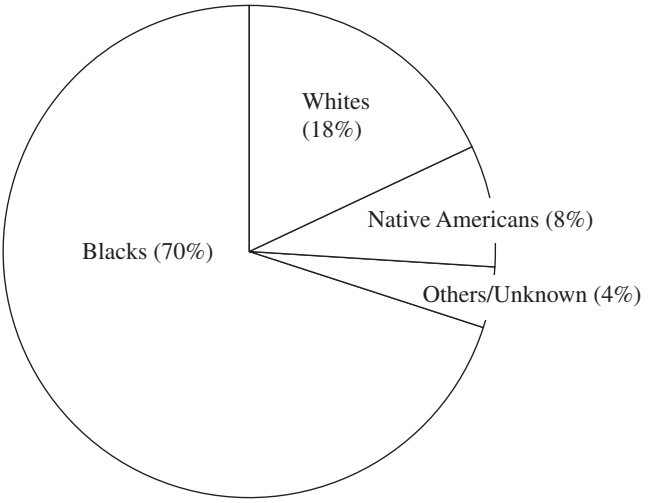

FIGURE 1.3 Children living in crack households.

TABLE 1.4

Cause of death	Number of deaths
Heart disease	12 378
Cancer	6448
Cerebrovascular disease	3958
Accidents	1814
Others	8088
Total	32 686

Example 1.8
Table 1.4 lists the number of deaths due to a variety of causes among Minnesota residents for the year 1975. After calculating the proportion of deaths due to each cause: for example,

$$\text{deaths due to cancer} = \frac{6448}{32\ 686} = 0.197 \text{ or } 19.7\%$$

we can present the results as in the pie chart shown in Figure 1.4.

Line Graphs A line graph is similar to a bar chart, but the horizontal axis represents time. In the applications most suitable to use line graphs, one binary characteristic is observed repeatedly over time. Different "groups" are consecutive years, so that a line graph is suitable to illustrate how certain proportions change over time. In a line graph, the proportion associated with each year is represented by a point at the appropriate height; the points are then connected by straight lines.

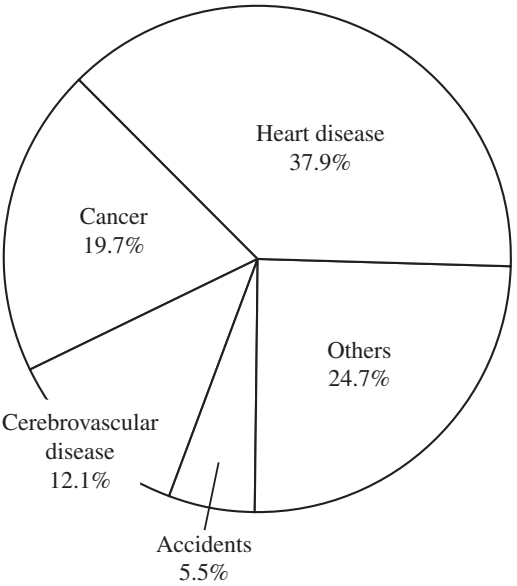

FIGURE 1.4 Causes of death for Minnesota residents, 1975.

TABLE 1.5

Year	Crude death rate per 100 000
1984	792.7
1985	806.6
1986	809.3
1987	813.1

Example 1.9
Between the years 1984 and 1987, the crude death rates for women in the United States were as listed in Table 1.5. The change in crude death rate for American women can be represented by the line graph shown in Figure 1.5.

In addition to their use with proportions, line graphs can be used to describe changes in the number of occurrences and in continuous measurements.

Example 1.10
The line graph shown in Figure 1.6 displays the trend in rates of malaria reported in the United States between 1940 and 1989 (proportion × 100 000 as above).

1.2 RATES

The term *rate* is somewhat confusing: sometimes it is used interchangeably with the term *proportion* as defined in Section 1.1; sometimes it refers to a quantity of a very different nature. In Section 1.2.1, on the *change rate*, we cover this special use, and in

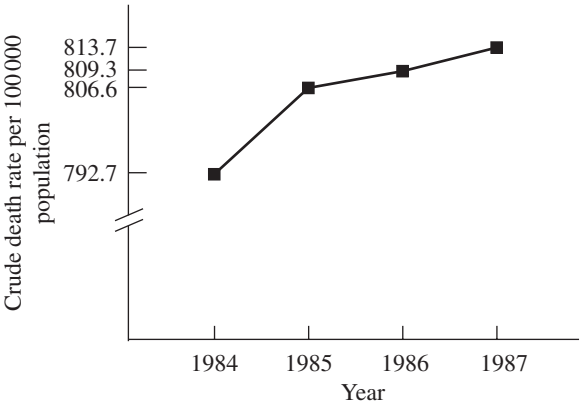

FIGURE 1.5 Death rates for United States women, 1984–1987.

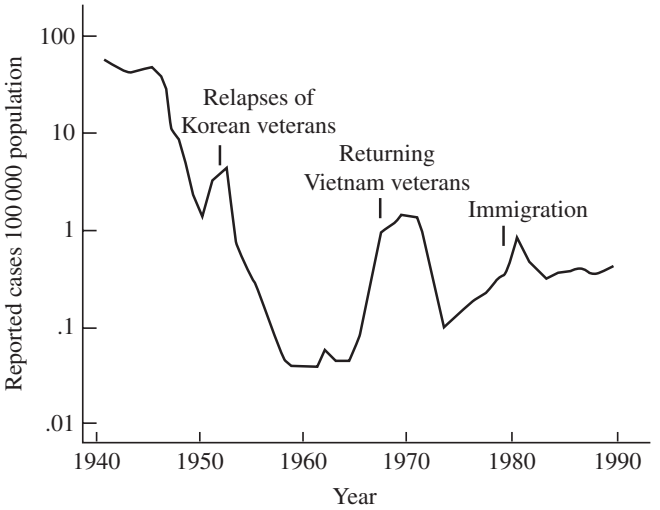

FIGURE 1.6 Malaria rates in the United States, 1940–1989.

the next two Sections, 1.2.2 and 1.2.3, we focus on *rates* used interchangeably with *proportions* as measures of morbidity and mortality. Even when they refer to the same things – measures of morbidity and mortality – there is some degree of difference between these two terms. In contrast to the static nature of proportions, rates are aimed at measuring the occurrences of events during or after a certain time period.

1.2.1 Changes

Familiar examples of rates include their use to describe changes after a certain period of time. The *change rate* is defined by:

$$\text{change rate}\,(\%) = \frac{\text{new value} - \text{old value}}{\text{old value}} \times 100.$$

In general, change rates could exceed 100%. *They are not proportions* (a proportion is a number between 0 and 1 or between 0 and 100%). Change rates are used primarily for description and are not involved in common *statistical analyses*.

Example 1.11

The following is a typical paragraph of a *news report*:

A total of 35 238 new AIDS cases was reported in 1989 by the Centers for Disease Control (CDC), compared to 32 196 reported during 1988. The 9% increase is the smallest since the spread of AIDS began in the early 1980s. For example, new AIDS cases were up 34% in 1988 and 60% in 1987. In 1989, 547 cases of AIDS transmissions from mothers to newborns were reported, up 17% from 1988; while females made up just 3971 of the 35 238 new cases reported in 1989, that was an increase of 11% over 1988.

In Example 1.11:

1. The change rate for new AIDS cases was calculated as

$$\frac{35\ 238 - 32\ 196}{32\ 196} \times 100 = 9.4\%$$

   (this was *rounded down* to the reported figure of 9% in the news report).

2. For the new AIDS cases transmitted from mothers to newborns, we have

$$17\% = \frac{547 - (1988\ \text{cases})}{1988\ \text{cases}} \times 100$$

   leading to

$$1988\ \text{cases} = \frac{547}{1.17} = 468$$

   (a figure obtainable, as shown above, but usually not reported because of redundancy).

   Similarly, the number of new AIDS cases for the year 1987 is calculated as follows:

$$34\% = \frac{32\ 196 - (1987\ \text{cases})}{1987\ \text{total}} \times 100$$

   or

$$1987\ \text{cases} = \frac{32\,196}{1.34} = 24\,027.$$

3. Among the 1989 new AIDS cases, the proportion of females is

$$\frac{3971}{35\ 238} = 0.113 \quad \text{or} \quad 11.3\%$$

and the proportion of males is

$$\frac{35\ 238 - 3971}{35\ 238} = 0.887 \quad \text{or} \quad 88.7\%.$$

The proportions of females and males add up to 1.0 or 100%.

1.2.2 Measures of Morbidity and Mortality

The field of vital statistics makes use of some special applications of rates, three types of which are commonly mentioned: crude, specific, and adjusted (or standardized). Unlike change rates, these measures are proportions. *Crude rates* are computed for an entire large group or population; they disregard factors such as age, gender, and race. *Specific rates* consider these differences among subgroups or categories of diseases. *Adjusted* or *standardized rates* are used to make valid summary comparisons between two or more groups possessing (for example) different age distributions.

The annual *crude death rate* is defined as the number of deaths in a calendar year divided by the population on 1 July of that year (which is usually an estimate); the quotient is often multiplied by 1000 or other suitable power of 10, resulting in a number between 1 and 100 or between 1 and 1000. For example, the 1980 population of California was 23 000 000 (as estimated on 1 July) and there were 190 237 deaths during 1980, leading to

$$\begin{aligned}\text{crude death rate} &= \frac{190\ 247}{23\ 000\ 000} \times 1000 \\ &= 8.3 \text{ deaths per } 1000 \text{ persons per year.}\end{aligned}$$

The age- and cause-specific death rates are defined similarly.

As for *morbidity*, the disease prevalence, as defined in Section 1.1, is a proportion used to describe the population at a certain point in time, whereas *incidence* is a rate used in connection with new cases:

$$\text{incidence rate} = \frac{\begin{array}{c}\text{number of persons who developed the disease} \\ \text{over a defined period of time (one year, say)}\end{array}}{\begin{array}{c}\text{number of persons initially without the disease} \\ \text{who were followed for the defined period of time}\end{array}}.$$

In other words, the prevalence presents a snapshot of the population's morbidity experience at a certain time point, whereas the incidence is aimed to investigate new onset morbidity. For example, the 35 238 new AIDS cases in Example 1.11 and the

national population without AIDS at the start of 1989 could be combined according to the formula above to yield an incidence of AIDS for the year.

Another interesting use of rates is in connection with *cohort studies*, epidemiological designs in which one enrolls a group of persons and follows them over certain periods of time; examples include occupational mortality studies, among others. The cohort study design focuses on a particular exposure rather than a particular disease as in case–control studies. Advantages of a longitudinal approach include the opportunity for more accurate measurement of exposure history and a careful examination of the time relationships between exposure and any disease under investigation. Each member of a cohort belongs to one of three types of termination:

1. Subjects still alive on the analysis date;
2. Subjects who died on a known date within the study period;
3. Subjects who are lost to follow-up after a certain date (these cases are a potential source of bias; effort should be expended on reducing the number of subjects in this category).

The contribution of each member is the length of follow-up time from enrollment to his or her termination. The quotient, defined as the number of deaths observed for the cohort, divided by the total of all members' follow-up times (in person-years, say) is the *rate* to characterize the mortality experience of the cohort:

$$\text{follow-up death rate} = \frac{\text{number of deaths}}{\text{total person-years}}.$$

Rates may be calculated for total deaths and for separate causes of interest, and they are usually multiplied by an appropriate power of 10, say 1000, to result in a single- or double-digit figure: for example, deaths per 1000 months of follow-up. Follow-up death rates may be used to measure the effectiveness of medical treatment programs.

Example 1.12

In an effort to provide a complete analysis of the survival of patients with end-stage renal disease (ESRD), data were collected for a sample that included 929 patients who initiated hemodialysis for the first time at the Regional Disease Program in Minneapolis, Minnesota, between 1 January 1976 and 30 June 1982; all patients were followed until 31 December 1982. Of these 929 patients, 257 are diabetics; among the 672 nondiabetics, 386 are classified as low risk (without co-morbidities such as arteriosclerotic heart disease, peripheral vascular disease, chronic obstructive pulmonary disease, and cancer). Results from these two subgroups are listed in Table 1.6. (Only some summarized figures are given here for illustration; details such as numbers of deaths and total treatment months for subgroups are not included.) For example, for low-risk patients over 60 years of age, there were 38 deaths during 2906 treatment months, leading to

$$\frac{38}{2906} \times 1000 = 13.08 \text{ deaths per 1000 treatment months.}$$

TABLE 1.6

Group	Age	Deaths/1000 treatment months
Low risk	1–45	2.75
	46–60	6.93
	61+	13.08
Diabetic	1–45	10.29
	46–60	12.52
	61+	22.16

TABLE 1.7

	Alaska			Florida		
Age group	Number of deaths	Persons	Deaths per 100 000	Number of deaths	Persons	Deaths per 100 000
0–4	162	40 000	405.0	2 049	546 000	375.3
5–19	107	128 000	83.6	1 195	1 982 000	60.3
20–44	449	172 000	261.0	5 097	2 676 000	190.5
45–64	451	58 000	777.6	19 904	1 807 000	1 101.5
65+	444	9 000	4 933.3	63 505	1 444 000	4 397.9
Total	1 615	407 000	396.8	91 760	8 455 000	1 085.3

1.2.3 Standardization of Rates

Crude rates, as measures of morbidity or mortality, can be used for population description and may be suitable for investigations of their variations over time; however, comparisons of crude rates are often invalid because the populations may be different with respect to an important characteristic such as age, gender, or race (these are potential *confounders*). To overcome this difficulty, an adjusted (or standardized) rate is used in the comparison; the adjustment removes the difference between populations in composition with respect to a confounder.

Example 1.13
Table 1.7 provides mortality data for Alaska and Florida for the year 1977.

Example 1.13 shows that the 1977 crude death rate per 100 000 population for Alaska was 396.8 and for Florida was 1085.7, almost a three-fold difference. However, a closer examination shows the following:

1. Alaska had higher age-specific death rates for four of the five age groups, the only exception being 45–64 years.
2. Alaska had a higher percentage of its population in the younger age groups.

The findings make it essential to adjust the death rates of the two states for age distribution in order to make a valid comparison. A simple way to achieve this, called

the *direct method*, is to apply, to a common *standard population*, age-specific rates observed from the two populations under investigation. For this purpose, the United States population as of the last decennial census is frequently used. The procedure consists of the following steps:

1. The standard population is enumerated by the same age groups.
2. The expected number of deaths in the standard population is computed for each age group of each of the two populations being compared. For example, for age group 0–4, the United States population for 1970 was 84 416 (per million in total population); therefore, we have:

   a. Alaska rate = 405.0 per 100 000. The expected number of deaths is:

$$\frac{(84\ 416)(405.0)}{100\ 000} = 341.9$$
$$\simeq 342$$

   (≃ means "almost equal to").

   b. Florida rate = 375.3 per 100 000. The expected number of deaths is:

$$\frac{(84\ 416)(375.3)}{100\ 000} = 316.8$$
$$\simeq 317$$

   which is lower than the expected number of deaths for Alaska obtained for the same age group.
3. Obtain the total number of deaths expected by adding expected deaths across the age groups.
4. The age-adjusted death rate is:

$$\text{adjusted rate} = \frac{\text{total number of deaths expected}}{\text{total standard population}} \times 100\ 000.$$

The calculations are detailed in Table 1.8.

The age-adjusted death rate per 100 000 population for Alaska is 788.6 and for Florida is 770.6. These age-adjusted rates are much closer than as shown by the crude rates, and the adjusted rate for Florida is *lower*. It is important to keep in mind that any population could be chosen as "standard," and because of this, an adjusted rate is artificial; it does not reflect data from an actual population. The numerical values of the adjusted rates depend in large part on the choice of the standard population. They have real meaning only as relative comparisons.

The advantage of using the United States population as the standard is that we can adjust death rates of many states and compare them with each other.

TABLE 1.8

Age group	1970 United States standard million	Alaska		Florida	
		Age-specific rate	Expected deaths	Age-specific rate	Expected deaths
0–4	84 416	405.0	342	375.3	317
5–19	294 353	83.6	246	60.3	177
20–44	316 744	261.0	827	190.5	603
45–64	205 745	777.6	1600	1101.5	2266
65+	98 742	4933.3	4871	4397.9	4343
Total	1 000 000		7886		7706

TABLE 1.9

Age group	Alaska population (used as standard)	Florida	
		Rate/100 000	Expected number of deaths
0–4	40 000	375.3	150
5–19	128 000	60.3	77
20–44	172 000	190.5	328
45–64	58 000	1101.5	639
65+	9 000	4397.9	396
Total	407 000		1590

Any population could be selected and used as a standard. In Example 1.13 it does not mean that there were only one million people in the United States in 1970; it only presents the *age distribution* of one million United States residents for that year. If all we want to do is to compare Florida with Alaska, we could choose either state as the standard and adjust the death rate of the other; this practice would save half the labor. For example, if we choose Alaska as the standard population, the adjusted death rate for Florida is calculated as shown in Table 1.9. The new adjusted rate,

$$\frac{(1590)(100\ 000)}{407\ 000} = 390.7 \text{ per } 100\ 000$$

is not the same as that obtained using the 1970 United States population as the standard (it was 770.6), but it also shows that after age adjustment, the death rate in Florida (390.7 per 100 000) is somewhat lower than that of Alaska (396.8 per 100 000; there is no need for adjustment here because we use Alaska's population as the standard population).

1.3 RATIOS

In many cases, such as disease prevalence and disease incidence, proportions and rates are defined very similarly, and the terms *proportions* and *rates* may even be used interchangeably. *Ratio* is a completely different term; it is a computation of the form

$$\text{ratio} = \frac{a}{b}$$

where a and b are *similar quantities* measured from *different groups* or under *different circumstances*. An example is the male/female ratio of smoking rates; such a ratio is positive but may exceed 1.0.

1.3.1 Relative Risk

One of the most often used ratios in epidemiological studies is *relative risk*, a concept for the comparison of two groups or populations with respect to a certain unwanted event (e.g., disease or death). The traditional method of expressing it in prospective studies is simply the ratio of the incidence rates:

$$\text{relative risk} = \frac{\text{disease incidence in group 1}}{\text{disease incidence in group 2}}.$$

However, the ratio of disease prevalences as well as follow-up death rates can also be formed. Usually, group 2 is under standard conditions–such as nonexposure to a certain risk factor – against which group 1 (exposed) is measured. A relative risk greater than 1.0 indicates harmful effects, whereas a relative risk below 1.0 indicates beneficial effects. For example, if group 1 consists of smokers and group 2 of nonsmokers, we have a *relative risk due to smoking*. Using the data on end-stage renal disease (ESRD) of Example 1.12, we can obtain the relative risks due to diabetes (Table 1.10). All three numbers are greater than 1.0 (indicating higher mortality for diabetics) and form a decreasing trend with increasing age.

1.3.2 Odds and Odds Ratio

The *relative risk*, also called the *risk ratio*, is an important index in epidemiological studies because in such studies it is often useful to measure the *increased* risk (if any) of incurring a particular disease if a certain factor is present. In cohort studies such an index is obtained readily by observing the experience of groups of subjects with

TABLE 1.10

Age group	Relative risk
1–45	3.74 = 10.29/2.75
46–60	1.81= 12.52/6.93
61+	1.69 = 22.16/13.08

TABLE 1.11

	Disease		
Factor	+	−	Total
+	A	B	$A + B$
−	C	D	$C + D$
Total	$A + C$	$B+D$	$N = A + B + C + D$

and without the factor, as shown in Table 1.11. In a case–control study the data do not present an immediate answer to this type of question, and we now consider how to obtain a useful shortcut solution.

Suppose that each subject in a large study, at a particular time, is classified as positive or negative according to some risk factor, and as having or not having a certain disease under investigation. For any such categorization the population may be enumerated in a 2×2 table (Table 1.11). The entries A, B, C, and D in the table are sizes of the four combinations of disease presence/ absence and factor presence/ absence, and the number N at the lower right corner of the table is the total population size. The relative risk is

$$\begin{aligned} \text{RR} &= \frac{A}{A+B} \div \frac{C}{C+D} \\ &= \frac{A\left(C+D\right)}{C\left(A+B\right)}. \end{aligned}$$

In many situations, the number of subjects classified as disease positive is small compared to the number classified as disease negative; that is,

$$\begin{aligned} C + D &\simeq D \\ A + B &\simeq B \end{aligned}$$

and therefore the relative risk can be approximated as follows:

$$\begin{aligned} \text{RR} &\simeq \frac{AD}{BC} \\ &= \frac{A\,/\,B}{C\,/\,D} \\ &= \frac{A\,/\,C}{B\,/\,D} \end{aligned}$$

where the slash denotes division. The resulting ratio, AD/BC, is an approximate relative risk, but it is often referred to as an *odds ratio* because:

1. A/B and C/D are the *odds* in favor of having disease from groups with or without the factor, respectively.

2. A/C and B/D are the odds in favor of having been exposed to the factors from groups with or without the disease. These two odds can easily be estimated using case–control data, by using sample frequencies. For example, the odds A/C can be estimated by a/c, where a is the number of exposed cases and c the number of nonexposed cases in the sample of cases used in a case–control design.

For the many diseases that are rare, the terms *relative risk* and *odds ratio* are used interchangeably because of the above-mentioned approximation. Of course, it is totally acceptable to draw conclusions on an odds ratio without invoking this approximation for a disease that is not rare. The relative risk is an important epidemiological index used to measure seriousness, or the magnitude of the harmful effect of suspected risk factors. For example, if we have

$$\text{RR} = 3.0$$

we can say that people exposed have a risk of contracting the disease that is approximately three times the risk of those unexposed. A perfect 1.0 indicates no effect, and beneficial factors result in relative risk values which are smaller than 1.0. From data obtained by a case–control or retrospective study, it is impossible to calculate the relative risk that we want, but if it is reasonable to assume that the disease is rare (prevalence is less than 0.05, say), we can calculate the odds ratio as a stepping stone and use it as an approximate relative risk. In these cases, we interpret the odds ratio calculated just as we would interpret the relative risk.

Example 1.14

The role of smoking in the etiology of pancreatitis has been recognized for many years. To provide estimates of the quantitative significance of these factors, a hospital-based study was carried out in eastern Massachusetts and Rhode Island between 1975 and 1979. Ninety-eight patients who had a hospital discharge diagnosis of pancreatitis were included in this unmatched case–control study. The control group consisted of 451 patients admitted for diseases other than those of the pancreas and biliary tract. Risk factor information was obtained from a standardized interview with each subject, conducted by a trained interviewer.

Some data for the males are given in Table 1.12. For these data for this example, the approximate relative risks, or odds ratios, are calculated as follows:

1. For ex-smokers relative to never-smokers,

$$\begin{aligned} \text{RR}_e &\simeq \frac{13/2}{80/56} \\ &= \frac{(13)(56)}{(80)(2)} \\ &= 4.55. \end{aligned}$$

(The subscript e in RR_e indicates that we are calculating the relative risk (RR) for ex-smokers.)

TABLE 1.12

Use of cigarettes	Cases	Controls
Never	2	56
Ex-smokers	13	80
Current smokers	38	81
Total	53	217

2. For current smokers relative to never-smokers,

$$\begin{aligned} \mathrm{RR}_c &\simeq \frac{38/2}{81/56} \\ &= \frac{(38)(56)}{(81)(2)} \\ &= 13.14. \end{aligned}$$

(The subscript c in RR_c indicates that we are calculating the relative risk (RR) for current smokers.)

In these calculations, the nonsmokers (who never smoke) are used as *references*. These values indicate that the risk of having pancreatitis for current smokers is approximately 13.14 times the risk for people who never smoke. The effect for ex-smokers is smaller (4.55 times) but is still very high (compared to 1.0, the no-effect baseline for relative risks and odds ratios). In other words, if the smokers were to quit smoking, they would reduce their own risk (from 13.14 times to 4.55 times) but *not* to the normal level for people who never smoke.

1.3.3 Generalized Odds for Ordered $2 \times k$ Tables

In this section we provide an interesting generalization of the concept of odds ratios to ordinal outcomes which are sometime used in biomedical research. Readers, especially beginners, may decide to skip it without loss of continuity; if so, the corresponding end-of-chapter Exercises should be skipped accordingly: 1.24(b), 1.25(c), 1.26(b), 1.27(b, c), 1.35(c), 1.38(c), and 1.45(b).

We can see this possible generalization by noting that an odds ratio can be interpreted as an odds but for a different event. For example, consider again the same 2×2 table as used in Section 1.3.2 (Table 1.11). Consider each of the A exposed cases, and imagine pairing each of them with each of the D unexposed controls; likewise consider each of the C unexposed cases and imagine pairing each of them with each of the B exposed controls. Then the number of these case–control pairs with different exposure histories is $(AD + BC)$; among them, AD pairs with an exposed case and BC pairs with an exposed control. Therefore AD/BC, the odds ratio of Section 1.3.2, can be seen as the odds of finding a pair with an exposed case among the discordant pairs (a *discordant pair* is a case–control pair with *different* exposure histories).

The interpretation above of the concept of an odds ratio as an odds can be generalized as follows. The aim here is to present an efficient method for use with *ordered*

TABLE 1.13

	Extent of injury received			
Seat belt	None	Minor	Major	Death
Yes	75	160	100	15
No	65	175	135	25

TABLE 1.14

	Column level				
Row	1	2	⋯	k	Total
1	a_1	a_2	⋯	a_k	A
2	b_1	b_2	⋯	b_k	B
Total	n_1	n_2	⋯	n_k	N

$2 \times k$ *contingency tables*, tables with two rows and k columns having a certain natural ordering. Let us first consider an example concerning the use of seat belts in automobiles. Each accident in this example is classified according to whether a seat belt was used and to the severity of injuries received: none, minor, major, or death (Table 1.13).

To compare the extent of injury from those who used seat belts with those who did not, we can calculate the percentage of seat belt users in each injury group, and we see that it decreases from level "none" to level "death:"

$$\text{None}: \quad \frac{75}{75+65} = 54\%$$

$$\text{Minor}: \quad \frac{160}{160+175} = 48\%$$

$$\text{Major}: \quad \frac{100}{100+135} = 43\%$$

$$\text{Death}: \quad \frac{15}{15+25} = 38\%.$$

What we are seeing here is a *trend* or a statistical *association* indicating that the lower the percentage of seat belt users, the more severe the injury.

We now present the concept of *generalized odds*, a special statistic specifically formulated to measure the strength of such a trend, and will use the same example and another one to illustrate its use. In general, consider an ordered $2 \times k$ table with the frequencies shown in Table 1.14.

The number of *concordances* is calculated by:

$$C = a_1\left(b_2 + \cdots + b_k\right) + a_2\left(b_3 + \cdots + b_k\right) + \cdots + a_{k-1}b_k$$

(the term *concordance pair* as used above corresponds to a less severe injury for the seat belt user). The number of *discordances* is

$$D = b_1\left(a_2 + \cdots + a_k\right) + b_2\left(a_3 + \cdots + a_k\right) + \cdots + b_{k-1}a_k.$$

To *measure* the degree of association, we use the index C/D and call it the generalized odds; if there are only two levels of injury, this new index is reduced to the familiar odds ratio. When data are properly arranged, by an a priori hypothesis, the products in the number of concordance pairs C (e.g., a_1b_2) go from upper left to lower right, and the products in the number of discordance pairs D (e.g., b_1a_2) go from lower left to upper right. In that a priori hypothesis, column 1 is associated with row 1; in the example above, the use of seat belt (yes, first row) is hypothesized to be associated with less severe injury (none, first column). Under this hypothesis, the resulting generalized odds is greater than 1.0.

Example 1.15

For the study above on the use of seat belts in automobiles, we have from the data shown in Table 1.13,

$$\begin{aligned} C &= (75)(175+135+25)+(160)(135+25)+(100)(25) \\ &= 53\,225 \\ D &= (65)(160+100+15)+(175)(100+15)+(135)(15) \\ &= 40\,025 \end{aligned}$$

leading to generalized odds of

$$\begin{aligned} \theta &= \frac{C}{D} \\ &= \frac{53\,225}{40\,025} \\ &= 1.33. \end{aligned}$$

That is, *given two people with different levels of injury*, the (generalized) odds that the more severely injured person did not wear a seat belt is 1.33. In other words, the people with the more severe injuries would be more likely than the people with less severe injuries to be those who did not use a seat belt.

The following example shows the use of generalized odds in case–control studies with an ordinal risk factor.

Example 1.16

A case–control study of the epidemiology of preterm delivery, defined as one with less than 37 weeks of gestation, was undertaken at Yale–New Haven Hospital in Connecticut during 1977. The study population consisted of 175 mothers of singleton

TABLE 1.15

Age	Cases	Controls
14–17	15	16
18–19	22	25
20–24	47	62
25–29	56	122
≥30	35	78

preterm infants and 303 mothers of singleton full-term infants. Table 1.15 gives the distribution of mother's age. We have:

$$\begin{aligned} C &= (15)(25+62+122+78)+(22)(62+122+78) \\ &\quad +(47)(122+78)+(56)(78) \\ &= 23\ 837 \\ D &= (16)(22+47+56+35)+(25)(47+56+35) \\ &\quad +(62)(56+35)+(122)(35) \\ &= 15\ 922, \end{aligned}$$

leading to generalized odds of:

$$\begin{aligned} \theta &= \frac{C}{D} \\ &= \frac{23\ 837}{15\ 922} \\ &= 1.5. \end{aligned}$$

This means that the odds that the younger mother has a preterm delivery is 1.5. In other words, the younger mothers would be more likely to have a preterm delivery.

The next example shows the use of generalized odds for contingency tables with more than two rows and more than two columns of data.

Example 1.17

Table 1.16 shows the results of a survey in which each subject of a sample of 282 adults was asked to indicate which of three policies he or she favored with respect to smoking in public places. We have:

$$\begin{aligned} C &= (15)(100+30+44+23)+(40)(30+23)+(100)(23) \\ &= 8380 \\ D &= (5)(100+30+40+10)+(44)(30+10)+(100)(10) \\ &= 4410. \end{aligned}$$

TABLE 1.16

Highest education level	Policy favored: No restrictions on smoking	Policy favored: Smoking allowed in designated areas only	Policy favored: No smoking at all	Total
Grade school	15	40	10	65
High school	15	100	30	145
College graduate	5	44	23	72
Total	35	184	63	300

leading to generalized odds of:

$$\begin{aligned}\theta &= \frac{C}{D} \\ &= \frac{8380}{4410} \\ &= 1.9.\end{aligned}$$

This means that the odds that the more educated person favors more restriction for smoking in public places is 1.9. In other words, people with more education would prefer more restriction on smoking in public places.

1.3.4 Mantel–Haenszel Method

In most investigations we are concerned with one primary outcome, such as a disease, and are focusing on one primary (risk) factor, such as an exposure with a possible harmful effect. There are situations, however, where an investigator may want to adjust for a confounder that could influence the outcome of a statistical analysis. A *confounder*, or *confounding variable*, is a variable that may be associated with either the disease or exposure or both. For example, in Example 1.2, a case–control study was undertaken to investigate the relationship between lung cancer and employment in shipyards during World War II among male residents of coastal Georgia. In this case, smoking is a possible counfounder; it has been found to be associated with lung cancer and it may be associated with employment because construction workers are more likely to be smokers. Specifically, we want to know:

1. Among smokers, whether or not shipbuilding and lung cancer are related;
2. Among nonsmokers, whether or not shipbuilding and lung cancer are related.

In fact, the original data were tabulated separately for three smoking levels (nonsmoking, moderate smoking, heavy smoking); in Example 1.2, the last two tables were combined and presented together for simplicity. Assume that the confounder, smoking, is not an effect modifier (i.e., smoking does not alter the relationship between lung cancer and shipbuilding), and that we do not want to reach separate

conclusions, one at each level of smoking. In those cases, we want to pool data for a combined decision. When both the disease and the exposure are binary, a popular method used to achieve this task is the *Mantel–Haenszel method*. This method provides one single estimate for the *common odds ratio* and can be summarized as follows:

1. We form 2×2 tables, one at each level of the confounder.
2. At each level of the confounder, we have the data listed in Table 1.17.

Since we assume that the confounder is not an effect modifier, the odds ratio is constant across its levels. The odds ratio at each level is estimated by ad/bc; the Mantel–Haenszel procedure pools data across levels of the confounder to obtain a combined estimate (a kind of weighted average of level-specific odds ratios):

$$\mathrm{OR}_{\mathrm{MH}} = \frac{\sum ad/n}{\sum bc/n}.$$

Example 1.18

A case–control study was conducted to identify reasons for the exceptionally high rate of lung cancer among male residents of coastal Georgia as first presented in Example 1.2. The primary risk factor under investigation was employment in shipyards during World War II, and data are tabulated separately for three levels of smoking (Table 1.18).

There are three 2×2 tables, one for each level of smoking. We begin with the 2×2 table for nonsmokers (Table 1.19). We have for the nonsmokers:

$$\frac{ad}{n} = \frac{(11)(203)}{299} = 7.47$$

$$\frac{bc}{n} = \frac{(35)(50)}{299} = 5.85.$$

TABLE 1.17

	Disease classification		
Exposure	+	−	Total
+	a	b	r_1
−	c	d	r_2
Total	c_1	c_2	n

TABLE 1.18

Smoking	Shipbuilding	Cases	Controls
No	Yes	11	35
	No	50	203
Moderate	Yes	70	42
	No	217	220
Heavy	Yes	14	3
	No	96	50

TABLE 1.19

Shipbuilding	Cases	Controls	Total
Yes	11 (a)	35 (b)	46 (r_1)
No	50 (c)	203 (d)	253 (r_2)
Total	61 (c_1)	238 (c_2)	299 (n)

The process is repeated for each of the other two smoking levels. For moderate smokers:

$$\frac{ad}{n} = \frac{(70)(220)}{549}$$
$$= 28.05$$
$$\frac{bc}{n} = \frac{(42)(217)}{549}$$
$$= 16.60.$$

For heavy smokers:

$$\frac{ad}{n} = \frac{(14)(50)}{163}$$
$$= 4.29$$
$$\frac{bc}{n} = \frac{(3)(96)}{163}$$
$$= 1.77.$$

These results are combined to obtain an estimate for the common odds ratio:

$$\text{OR}_{\text{MH}} = \frac{7.47 + 28.05 + 4.29}{5.85 + 16.60 + 1.77}$$
$$= 1.64.$$

This combined estimate of the odds ratio, 1.64, represents an approximate increase of 64% in lung cancer risk for those employed in the shipbuilding industry.

The following is a similar example aiming at the possible effects of oral contraceptive use on myocardial infarction. The presentation has been shortened, showing only key calculations.

Example 1.19
A case–control study was conducted to investigate the relationship between myocardial infarction (MI) and oral contraceptive use (OC). The data, stratified by cigarette smoking, are listed in Table 1.20. Application of the Mantel–Haenszel procedure yields the results shown in Table 1.21. The combined odds ratio estimate is

$$\begin{aligned} \mathrm{OR}_{\mathrm{MH}} &= \frac{3.57+18.84}{2.09+12.54} \\ &= 1.53 \end{aligned}$$

representing an approximate increase of 53% in myocardial infarction risk for oral contraceptive users.

1.3.5 Standardized Mortality Ratio

In a cohort study, the follow-up death rates are calculated and used to describe the mortality experience of the cohort under investigation. However, the observed mortality of the cohort is often compared with that expected from the death rates of the national population (used as *standard* or reference or *baseline*). The basis of this method is the comparison of the observed number of deaths, d, from the cohort with the mortality that would have been expected if the group had experienced death rates similar to those of the national population of which the cohort is a part. Let e denote the expected number of deaths; then the comparison is based on the following ratio, called the *standardized mortality ratio:*

$$\mathrm{SMR} = \frac{d}{e}.$$

TABLE 1.20

Smoking	OC Use	Cases	Controls
No	Yes	4	52
	No	34	754
Yes	Yes	25	83
	No	171	853

TABLE 1.21

	Smoking	
	No	Yes
ad/n	3.57	18.84
bc/n	2.09	12.54

The expected number of deaths is calculated using published national life tables, and the calculation can be approximated as follows:

$$e \simeq \lambda T$$

where T is the total follow-up time (person-years) from the cohort and λ the annual death rate (per person) from the referenced population. Of course, the annual death rate of the referenced population changes with age. Therefore, what we actually do in research is more complicated, although based on the same idea. First, we subdivide the cohort into many age groups, then calculate the product λT for each age group using the correct age-specific rate for that group, and add up the results.

Example 1.20
Some 7000 British workers exposed to vinyl chloride monomer were followed for several years to determine whether their mortality experience differed from those of the general population. The data in Table 1.22 are for deaths from cancers and are tabulated separately for four groups based on years since entering the industry. This data display shows some interesting features:

1. For the group with 1–4 years since entering the industry, we have a death rate that is substantially less than that of the general population (SMR = 0.445, or 44.5%). This phenomenon, known as the *healthy worker effect*, is probably a consequence of a selection factor whereby workers are necessarily in better health (than people in the general population) at the time of their entry into the workforce.
2. We see an attenuation of the healthy worker effect (i.e., a decreasing trend) with the passage of time, so that the cancer death rates show a slight excess after 15 years. (Vinyl chloride exposures are known to induce a rare form of liver cancer and to increase rates of brain cancer.)

Taking the ratio of two standardized mortality ratios is another way of expressing relative risk. For example, the risk of the 15+ years group is 1.58 times the risk of the 5–9 years group, since the ratio of the two corresponding mortality ratios is:

$$\frac{111.8}{70.6} = 1.58.$$

TABLE 1.22

	Years since entering the industry				
Deaths from cancers	1–4	5–9	10–14	15+	Total
Observed	9	15	23	68	115
Expected	20.3	21.3	24.5	60.8	126.8
SMR (%)	44.5	70.6	94.0	111.8	90.7

Similarly, the risk of the 15+ years group is 2.51 times the risk of the 1–4 years group because the ratio of the two corresponding mortality ratios is:

$$\frac{111.8}{44.5} = 2.51.$$

1.4 NOTES ON COMPUTATIONS

Much of this book is concerned with arithmetic procedures for data analysis, some with rather complicated formulas. In many biomedical investigations, particularly those involving large quantities of data, the analysis (e.g., regression analysis in Chapter 9) gives rise to difficulties in computational implementation. In these investigations it will be necessary to use statistical software specially designed to do these jobs. Most of the calculations described in this book can be carried out readily using statistical packages, and any student or practitioner of data analysis will find the use of such packages essential.

Methods of survival analysis (first half of Chapter 13), for example, and nonparametric methods (Sections 2.4 and 7.4), and methods of multiple regression analysis (Section 9.2 and Chapters 10, 11, and 12) may best be handled by a specialized package such as SAS or R; coding instructions for these two packages are included in our examples where they were used. However, students and investigators contemplating the use of one of these commercial programs should read the specifications for each program before choosing the options necessary or suitable for any particular procedure. But these sections are exceptions; many calculations described in this book can be carried out readily using Microsoft Excel, software available in every personal computer. Notes on the use of Excel are included in separate sections at the end of each chapter.

A *worksheet* or spreadsheet is a (blank) sheet where you do your work. An Excel *file* holds a stack of worksheets in a *workbook*. You can *name* a sheet, put data on it and *save*; later, you can *open* and use it. You can *move* or *size* your windows by *dragging* the borders. You can also scroll up and down, or left and right, through an Excel worksheet using the scroll bars on the right side and at the bottom.

An Excel worksheet consists of *grid lines* forming *columns* and *rows*; columns are *lettered* and rows are *numbered*. The intersection of each column and row is a box called a *cell*. Every cell has an *address*, also called a *cell reference*; to refer to a cell, enter the column letter followed by the row number. For example, the intersection of *column C* and *row 3* is *cell C3*. Cells hold numbers, text, or formulas. To refer to a range of cells, enter the cell in the upper left corner of the range followed by a colon (:) and then the lower right corner of the range. For example, A1:B20 refers to the first 20 rows in both columns A and B.

You can click a cell to make it *active* (for use); an active cell is where you enter or edit your data, and it is identified by a *heavy border*. You can also *define* or *select* a *range* by left-clicking on the upper leftmost cell and dragging the mouse to the lower rightmost cell. To move around inside a selected range, press Tab or Enter to move forward one cell at a time.

Excel is software designed to handle numbers; so get a project and start typing. Conventionally, files for data analysis use rows for subjects and columns for factors. For example, you conduct a survey using a 10-item questionaire and receive returns from 75 people; your data require a file with 75 rows and 10 columns – not counting labels for columns (factor names) and rows (subject ID). If you make an error, it can be fixed (hit the *Del* key, which deletes the cell contents). You can change your mind again, deleting the delete, by clicking the *Undo* button (reverse curved arrow). Remember, you can widen your columns by double-clicking their right borders.

The formula bar (near the top, next to an = sign) is a common way to provide the content of an active cell. Excel executes a formula from left to right and performs multiplication (*) and division (/) before addition (+) and subtraction (−). Parentheses can/should be used to change the order of calculations. To use formulas (e.g., for data transformation), do it in one of two ways: (1) click the cell you want to fill, then type an = sign followed by the formula in the formula bar (e.g., click C5, then type = A5 + B5); or (2) click the cell you want to fill, then click the *paste* function icon, f*, which will give you – in a box – a list of Excel functions available for your use.

The *cut and paste* procedure greatly simplifies the typing necessary to form a chart or table or to write numerous formulas. The procedure involves highlighting the cells that contain the information you want to copy, clicking on the *cut* button (scissors icon) or *copy* button (two-page icon), selecting the cell(s) in which the information is to be placed, and clicking on the *paste* button (clipboard and page icon).

The *select and drag* is very efficient for data transformation. Suppose that you have the weight and height of 15 men (weights are in C6:C20 and heights are in D6:D20) and you need their body mass index. You can use the formula bar, for example, clicking E6 and typing = C6/(D6∧2). The content of E6 now has the body mass index for the first man in your sample, but you do not have to repeat this process 15 times. Notice that when you click on E6, there is a little box in the lower right corner of the cell boundary. If you move the mouse over this box, the cursor changes to a smaller plus sign. If you click on this box, you can then drag the mouse over the remaining cells (E7 to E20), and when you release the button, the cells will be filled with the calculated body mass index values.

Bar and Pie Charts Forming a bar chart or pie chart to display proportions is a very simple task. Just click any blank cell to start; you'll be able to move your chart to any location when done. With data ready, click the *ChartWizard* icon (the one with multiple colored bars on the *standard toolbar* near the top). A box appears with choices including bar chart, pie chart, and line chart; the list is on the left side. Choose your chart type and follow instructions. There are many choices, including three dimensions. You can put data and charts side by side for an impressive presentation.

Rate Standardization This is a good problem to practice with Excel: Use it as a *calculator*. Recall this example:

- Florida's rate = 1085.3
- Alaska's rate = 396.8

- If Florida has Alaska's population (see Table 1.9), you can:
  - Use *formula* to calculate the first number of expected deaths.
  - Use *drag and fill* to obtain other numbers of expected deaths.
  - Select the last column, then click the *Autosum* icon (Σ) to obtain the total number of deaths expected.

Forming 2 × 2 Tables Recall that in a data file you have one row for each subject and one column for each variable. Suppose that *two* of those variables are categorical, say *binary*, and you want to form a 2 × 2 table so you can study their relationship: for example, to calculate an *odds ratio*.

- *Step 0:* Create a dummy factor; call it, say, *fake*, and fill up that column with "*1*" (you can enter "1" for the first subject, then *select* and *drag*).
- *Step 1: Activate* a cell (by clicking it), then click *Data* (on the bar above the standard toolbar, near the end); when a box appears, choose *Pivot-Table Report*. Click "*next*" (to indicate that data are here, in Excel, then *highlight* the area containing your data (including variable names on the first row – use the mouse; or you could identify the *range of cells* – say, C5:E28) as a response to a question on *range*. Then click "*next*" to bring in the *PivotTable Wizard*, which shows two groups of things:
  a. A *frame* for a 2 × 2 table with places identified as *row, column*, and *data*;
  b. Names of the factors you chose: say, *exposure, disease*, and *fake*.
- *Step 2: Drag* exposure to merge with row (or column), *drag* disease to merge with column (or row), and *drag* fake to merge with data. Then click *Finish*; a 2 × 2 table appears in the active cell you identified, complete with *cell frequencies, row and column totals*, and *grand total*.

Note: If you have another *factor* in addition to exposure and disease available in the data set, even a column for *names* or *IDs*, there is no need to create the *dummy factor*. Complete step 1; then, in step 2, *drag* that third factor, say *ID*, to merge with "*data*" in the frame shown by the *PivotTable Wizard*; it appears as *sum ID*. Click on that item, then choose *count* (to replace *sum*).

EXERCISES

1.1 Self-reported injuries among left- and right-handed people were compared in a survey of 1986 college students in British Columbia, Canada. Of the 180 left-handed students, 93 reported at least one injury, and 619 of the 1716 right-handed students reported at least one injury in the same period. Arrange the data in a 2 × 2 table and calculate the proportion of people with at least one injury during the period of observation for each group.

1.2 A study was conducted to evaluate the hypothesis that tea consumption and premenstrual syndrome are associated. A group of 188 nursing students and

64 tea factory workers were given questionnaires. The prevalence of premenstrual syndrome was 39% among the nursing students and 77% among the tea factory workers. How many people in each group have premenstrual syndrome? Arrange the data in a 2×2 table.

1.3 The relationship between prior condom use and tubal pregnancy was assessed in a population-based case–control study at Group Health Cooperative of Puget Sound during 1981–1986. The results are shown in Table E1.3. Compute the group size and the proportion of subjects in each group who never used condoms.

TABLE E1.3

Condom use	Cases	Controls
Never	176	488
Always	51	186

1.4 Epidemic keratoconjunctivitis (EKC) or "shipyard eye" is an acute infectious disease of the eye. A case of EKC is defined as an illness:

- Consisting of redness, tearing, and pain in one or both eyes for more than three days' duration;
- Diagnosed as EKC by an ophthalmologist.

In late October 1977, physician A (one of the two ophthalmologists providing the majority of specialized eye care to the residents of a central Georgia county; population 45 000) saw a 27-year-old nurse who had returned from a vacation in Korea with severe EKC. She received symptomatic therapy and was warned that her eye infection could spread to others; nevertheless, numerous cases of an illness similar to hers soon occurred in the patients and staff of the nursing home (nursing home A) where she worked (these people came to physician A for diagnosis and treatment). Table E1.4 provides the exposure history of 22 persons with EKC between 27 October 1977 and 13 January 1978 (when the outbreak stopped after proper control techniques were initiated). Nursing home B, included in this table, is the only other area chronic-care facility. Compute and compare the proportions of cases from the two nursing homes. What would be your conclusion?

TABLE E1.4

Exposure cohort	Number exposed	Number of cases
Nursing home A	64	16
Nursing home B	238	6

1.5 In August 1976, tuberculosis was diagnosed in a high school student (index case) in Corinth, Mississippi. Subsequently, laboratory studies revealed that the student's disease was caused by drug-resistant tubercule bacilli. An epidemiologic

investigation was conducted at the high school. Table E1.5 gives the rate of positive tuberculin reactions, determined for various groups of students according to degree of exposure to the index case.

(a) Compute and compare the proportions of positive cases for the two exposure levels. What would be your conclusion?

(b) Calculate the odds ratio associated with high exposure. Does this result support your conclusion in part (a)?

TABLE E1.5

Exposure level	Number tested	Number positive
High	129	63
Low	325	36

1.6 Consider the data taken from a study that attempts to determine whether the use of electronic fetal monitoring (EFM) during labor affects the frequency of cesarean section deliveries. Of the 5824 infants included in the study, 2850 were electronically monitored and 2974 were not. The outcomes are listed in Table E1.6.

(a) Compute and compare the proportions of cesarean delivery for the two exposure groups. What would be your conclusion?

(b) Calculate the odds ratio associated with EFM exposure. Does this result support your conclusion in part (a)?

TABLE E1.6

	EFM exposure		
Cesarean delivery	Yes	No	Total
Yes	358	229	587
No	2492	2745	5237
Total	2850	2974	5824

1.7 A study was conducted to investigate the effectiveness of bicycle safety helmets in preventing head injury. The data consist of a random sample of 793 persons who were involved in bicycle accidents during a one-year period (Table E1.7).

TABLE E1.7

	Wearing helmet		
Head injury	Yes	No	Total
Yes	17	218	235
No	130	428	558
Total	147	646	793

(a) Compute and compare the proportions of head injury for the group with helmets versus the group without helmets. What would be your conclusion?

(b) Calculate the odds ratio associated with not using helmet. Does this result support your conclusion in part (a)?

1.8 A case–control study was conducted in Auckland, New Zealand, to investigate the effects among regular drinkers of alcohol consumption on both nonfatal myocardial infarction and coronary death in the 24 hours after drinking. Data were tabulated separately for men and women (Table E1.8).

(a) Refer to the myocardial infarction data. Calculate separately for men and women the odds ratio associated with drinking.

(b) Compare the two odds ratios in part (a). When the difference is confirmed properly, we have an effect modification.

(c) Refer to coronary deaths. Calculate separately for men and women the odds ratio associated with drinking.

(d) Compare the two odds ratios in part (c). When the difference is confirmed properly, we have an effect modification.

TABLE E1.8

		Myocardial infarction		Coronary death	
	Drink during the last 24 h	Controls	Cases	Controls	Cases
Men	No	197	142	135	103
	Yes	201	136	159	69
Women	No	144	41	89	12
	Yes	122	19	76	4

1.9 Data taken from a study to investigate the effects of smoking on cervical cancer are stratified by the number of sexual partners (Table E1.9).

(a) Calculate the odds ratio associated with smoking separately for the two groups, those with zero or one partner and those with two or more partners.

(b) Compare the two odds ratios in part (a). When the difference is confirmed properly, we have an effect modification.

(c) Assuming that the odds ratios for the two groups, those with zero or one partner and those with two or more partners, are equal (in other words, the number of partners is not an effect modifier), calculate the Mantel–Haenszel estimate of this common odds ratio.

TABLE E1.9

		Cancer	
Number of partners	Smoking	Yes	No
Zero or one	Yes	12	21
	No	25	118
Two or more	Yes	96	142
	No	92	150

1.10 Table E1.10 provides the proportions of currently married women having an unplanned pregnancy. (Data are tabulated for several different methods of contraception.) Display these proportions in a bar chart.

TABLE E1.10

Method of contraception	Proportion with unplanned pregnancy
None	0.431
Diaphragm	0.149
Condom	0.106
IUD	0.071
Pill	0.037

1.11 Table E1.11 summarizes the coronary heart disease (CHD) and lung cancer mortality rates per 1000 person-years by the number of cigarettes smoked per day at baseline for men participating in the Multiple Risk Factor Intervention Trial (MRFIT, a very large controlled clinical trial focusing on the relationship between smoking and cardiovascular diseases). For each cause of death, display the rates in a bar chart.

TABLE E1.11

		CHD deaths		Lung cancer deaths	
	Total	*N*	Rate/1000 person-years	*N*	Rate/1000 person-years
Never-smokers	1859	44	2.22	0	0
Ex-smokers	2813	73	2.44	13	0.43
Smokers					
1–19 cigarettes/day	856	23	2.56	2	0.22
20–39 cigarettes/day	3747	173	4.45	50	1.29
≥40 cigarettes/day	3591	115	3.08	54	1.45

1.12 Table E1.12 provides data taken from a study on the association between race and use of medical care by adults experiencing chest pain in the past year. Display the proportions of the three response categories for each group, blacks and whites, in a separate pie chart.

TABLE E1.12

Response	Black	White
MD seen in past year	35	67
MD seen not in past year	45	38
MD never seen	78	39
Total	158	144

1.13 The frequency distribution for the number of cases of pediatric AIDS between 1983 and 1989 is shown in Table E1.13. Display the trend of numbers of cases using a line graph.

TABLE E1.13

Year	Number of cases	Year	Number of cases
1983	122	1987	1 412
1984	250	1988	2 811
1985	455	1989	3 098
1986	848		

1.14 A study was conducted to investigate the changes between 1973 and 1985 in women's use of three preventive health services. The data were obtained from the National Health Survey; women were divided into subgroups according to age and race. The percentages of women receiving a breast exam within the past two years are given in Table E1.14. Separately for each group, blacks and whites, display the proportions of women receiving a breast exam within the past two years in a bar chart so as to show the relationship between the examination rate and age. Mark the midpoint of each age group on the horizontal axis and display the same data using a line graph.

TABLE E1.14

	Breast examination within past two years	
Age and race	1973	1985
Total	65.5	69.6
Black	61.7	74.8
White	65.9	69.0
20–39 years	77.5	77.9
Black	77.0	83.9
White	77.6	77.0
40–59 years	62.1	66.0
Black	54.8	67.9
White	62.9	65.7
60–79 years	44.3	56.2
Black	39.1	64.5
White	44.7	55.4

1.15 Consider the data shown in Table E1.15. Calculate the sensitivity and specificity of x-ray as a screening test for tuberculosis.

TABLE E1.15

	Tuberculosis		
X-ray	No	Yes	Total
Negative	1739	8	747
Positive	51	22	73
Total	1790	30	1820

1.16 Sera from a T-lymphotropic virus type (HTLV-I) risk group (prostitute women) were tested with two commercial "research" enzyme-linked immunoabsorbent assays (EIA) for HTLV-I antibodies. These results were compared with a gold standard, and outcomes are shown in Table E1.16. Calculate and compare the sensitivity and specificity of these two EIAs.

TABLE E1.16

	Dupont's EIA		Cellular product's EIA	
True	Positive	Negative	Positive	Negative
Positive	15	1	16	0
Negative	2	164	7	179

1.17 Table E1.17 provides the number of deaths for several leading causes among Minnesota residents for the year 1991.

(a) Calculate the percentage of total deaths from each cause, and display the results in a pie chart.

(b) From the death rate (per 100 000 population) for heart disease, calculate the population for Minnesota for the year 1991.

(c) From the result of part (b), fill in the missing death rates (per 100 000 population) in the table.

TABLE E1.17

Cause of death	Number of deaths	Rate per 100 000 population
Heart disease	10 382	294.5
Cancer	8 299	?
Cerebrovascular disease	2 830	?
Accidents	1 381	?
Other causes	11 476	?
Total	34 368	?

1.18 The survey described in Example 1.1, continued in Section 1.1.1, provided percentages of boys from various ethnic groups who use tobacco at least weekly. Display these proportions in a bar chart similar to the one in Figure 1.2.

1.19 A case–control study was conducted relating to the epidemiology of breast cancer and the possible involvement of dietary fats, vitamins, and other nutrients. It included 2024 breast cancer cases admitted to Roswell Park Memorial Institute, Erie County, New York, from 1958 to 1965. A control group of 1463 was chosen from patients having no neoplasms and no pathology of gastrointestinal or reproductive systems. The primary factors being investigated were vitamins A and E (measured in international units per month). Data for 1500 women over 54 years of age are given in Table E1.19. Calculate the odds ratio associated with a decrease (exposure is low consumption) in ingestion of foods containing vitamin A.

TABLE E1.19

Vitamin A (IU/month)	Cases	Controls
≤150 500	893	392
>150 500	132	83
Total	1025	475

1.20 Refer to the data set in Table 1.1 (see Example 1.2).

(a) Calculate the odds ratio associated with employment in shipyards for nonsmokers.

(b) Calculate the same odds ratio for smokers.

(c) Compare the results of parts (a) and (b). When the difference is confirmed properly, we have a three-term interaction or effect modification, where smoking alters the effect of employment in shipyards as a risk for lung cancer.

(d) Assuming that the odds ratios for the two groups, nonsmokers and smokers, are equal (in other words, smoking is not an effect modifier), calculate the Mantel–Haenszel estimate of this common odds ratio.

1.21 Although cervical cancer is not a major cause of death among American women, it has been suggested that virtually all such deaths are preventable. In an effort to find out who is being screened for the disease, data from the 1973 National Health Interview (a sample of the United States population) were used to examine the relationship between Pap testing and some socioeconomic factors. Table E1.21 provides the percentages of women who reported never having had a Pap test. (These are from metropolitan areas.)

(a) Calculate the odds ratios associated with race (black versus white) among

(i) 25–44 years old, nonpoor

(ii) 45–64 years old, nonpoor

(iii) 65+ years old, nonpoor.

TABLE E1.21

Age and income	White	Black
25–44 years		
Poor	13.0	14.2
Nonpoor	5.9	6.3
45–64 years		
Poor	30.2	33.3
Nonpoor	13.2	23.3
65 years and over		
Poor	47.4	51.5
Nonpoor	36.9	47.4

Briefly discuss a possible effect modification, if any.

(b) Calculate the odds ratios associated with income (poor versus nonpoor) among

(i) 25–44 years old, black

(ii) 45–64 years old, black

(iii) 65+ years old, black.

Briefly discuss a possible effect modification, if any.

(c) Calculate the odds ratios associated with race (black versus white) among

(i) 65+ years old, poor

(ii) 65+ years old, nonpoor.

Briefly discuss a possible effect modification.

1.22 Since incidence rates of most cancers rise with age, this must always be considered a confounder. Stratified data for an unmatched case–control study are given in Table E1.22. The disease was esophageal cancer among men and the risk factor was alcohol consumption.

(a) Calculate separately for the three age groups the odds ratio associated with *high* alcohol consumption.

(b) Compare the three odds ratios in part (a). When the difference is confirmed properly, we have an effect modification.

(c) Assuming that the odds ratios for the three age groups are equal (in other words, age is not an effect modifier), calculate the Mantel–Haenszel estimate of this common odds ratio.

TABLE E1.22

	Daily alcohol consumption	
Age	80+ g	0–79 g
25–44 years		
Cases	5	5
Controls	35	270
45–64 years		
Cases	67	55
Controls	56	277
65+ years		
Cases	24	44
Controls	18	129

1.23 Postmenopausal women who develop endometrial cancer are on the whole heavier than women who do not develop the disease. One possible explanation is that heavy women are more exposed to endogenous estrogens which are produced in postmenopausal women by conversion of steroid precursors to active estrogens in peripheral fat. In the face of varying levels of endogenous estrogen

production, one might ask whether the carcinogenic potential of exogenous estrogens would be the same in all women. A case–control study has been conducted to examine the relation among weight, replacement estrogen therapy, and endometrial cancer. The results are shown in Table E1.23.

(a) Calculate separately for the three weight groups the odds ratio associated with estrogen replacement.

(b) Compare the three odds ratios in part (a). When the difference is confirmed properly, we have an effect modification.

(c) Assuming that the odds ratios for the three weight groups are equal (in other words, weight is not an effect modifier), calculate the Mantel–Haenszel estimate of this common odds ratio.

TABLE E1.23

	Estrogen replacement	
Weight (kg)	Yes	No
<57		
Cases	20	12
Controls	61	183
57–75		
Cases	37	45
Controls	113	378
>75		
Cases	9	42
Controls	23	140

1.24 The role of menstrual and reproductive factors in the epidemiology of breast cancer has been reassessed using pooled data from three large case–control studies of breast cancer from several Italian regions (Negri et al., 1988). Table E1.24 summarizes data for age at menopause and age at first live birth.

TABLE E1.24

	Cases	Controls
Age at first live birth (years)		
<22	621	898
22–24	795	909
25–27	791	769
≥28	1043	775
Age at menopause (years)		
<45	459	543
45–49	749	803
≥50	1378	1167

(a) For each of the two factors (age at first live birth and age at menopause), choose the lowest level as the reference and calculate the odds ratio associated with each other level.

(b) For each of the two factors (age at first live birth and age at menopause), calculate the generalized odds and give your interpretation. How does this result compare with those in part (a)?

1.25 Risk factors of gallstone disease were investigated in male self-defense officials who received, between October 1986 and December 1990, a retirement health examination at the Self-Defense Forces Fukuoka Hospital, Fukuoka, Japan. Some of the data are shown in Table E1.25.

(a) For each of the three factors (smoking, alcohol, body mass index), rearrange the data into a 3 × 2 table; the other column is for those without gallstones.

(b) For each of the three 3 × 2 tables in part (a), choose the lowest level as the reference and calculate the odds ratio associated with each other level.

(c) For each of the three 3 × 2 tables in part (a), calculate the generalized odds and give your interpretation. How does this result compare with those in part (b)?

TABLE E1.25

	Number of men surveyed	
Factor	Total	Number with gallstones
Smoking		
Never	621	11
Past	776	17
Current	1342	33
Alcohol		
Never	447	11
Past	113	3
Current	2179	47
Body mass index (kg/m^2)		
<22.5	719	13
22.5–24.9	1301	30
≥25.0	719	18

1.26 Data were collected from 2197 white ovarian cancer patients and 8893 white controls in 12 different United States case–control studies conducted by various investigators in the period 1956–1986. These were used to evaluate the relationship of invasive epithelial ovarian cancer to reproductive and menstrual characteristics, exogenous estrogen use, and prior pelvic surgeries. Data related to unprotected intercourse and to history of infertility are shown in Table E1.26.

(a) For each of the two factors (duration of unprotected intercourse and history of infertility, treating the latter as ordinal: no history, history but no drug

TABLE E1.26

	Cases	Controls
Duration of unprotected intercourse (years)		
<2	237	477
2–9	166	354
10–14	47	91
≥15	133	174
History of infertility		
No	526	966
Yes		
No drug use	76	124
Drug use	20	11

use, and history with drug use), choose the lowest level as the reference, and calculate the odds ratio associated with each other level.

(b) For each of the two factors (duration of unprotected intercourse and history of infertility, treating the latter as ordinal: no history, history but no drug use, and history with drug use), calculate the generalized odds and give your interpretation. How does this result compare with those in part (a)?

1.27 Postneonatal mortality due to respiratory illnesses is known to be inversely related to maternal age, but the role of young motherhood as a risk factor for respiratory morbidity in infants has not been explored thoroughly. A study was conducted in Tucson, Arizona, aimed at the incidence of lower respiratory tract illnesses during the first year of life. In this study, over 1200 infants were enrolled at birth between 1980 and 1984. The data shown in Table E1.27 are concerned with wheezing lower respiratory tract illnesses (wheezing LRI: no/yes).

(a) For each of the two groups, boys and girls, choose the lowest age group as the reference and calculate the odds ratio associated with each age group.

(b) For each of the two groups, boys and girls, calculate the generalized odds and give your interpretation. How does this result compare with those in part (a)?

(c) Compare the two generalized odds in part (b) and draw your conclusion.

TABLE E1.27

	Boys		Girls	
Maternal age (years)	No	Yes	No	Yes
<21	19	8	20	7
21–25	98	40	128	36
26–30	160	45	148	42
>30	110	20	116	25

1.28 An important characteristic of glaucoma, an eye disease, is the presence of classical visual field loss. Tonometry is a common form of glaucoma screening, whereas, for example, an eye is classified as positive if it has an intraocular pressure of 21 mmHg or higher at a single reading. Given the data shown in Table E1.28, calculate the sensitivity and specificity of this screening test.

TABLE E1.28

	Test result		
Field loss	Positive	Negative	Total
Yes	13	7	20
No	413	4567	4980

1.29 From the information in the news report quoted in Example 1.11, calculate:

(a) The number of new AIDS cases for the years 1987 and 1986.

(b) The number of cases of AIDS transmission from mothers to newborns for 1988.

1.30 In an effort to provide a complete analysis of the survival of patients with end-stage renal disease (ESRD), data were collected for a sample that included 929 patients who initiated hemodialysis for the first time at the Regional Disease Program in Minneapolis, Minnesota between 1 January 1976 and 30 June 1982; all patients were followed until 31 December 1982. Of these 929 patients, 257 are diabetics; among the 672 nondiabetics, 386 are classified as low risk (without comorbidities such as arteriosclerotic heart disease, peripheral vascular disease, chronic obstructive pulmonary, and cancer). For the low-risk ESRD patients, we have the follow-up data shown in Table E1.30 (in addition to those in Example 1.12). Compute the follow-up death rate for each age group and the relative risk for the group "70+" versus "51–60."

TABLE E1.30

Age (years)	Deaths	Treatment months
21–30	4	1012
31–40	7	1387
41–50	20	1706
51–60	24	2448
61–70	21	2060
70+	17	846

1.31 Mortality data for the state of Georgia for the year 1977 are given in part (a) of Table E1.31.

(a) From this mortality table, calculate the crude death rate for Georgia.

(b) From Table E1.31 and the mortality data for Alaska and Florida for the year 1977 (Table 1.7), calculate the age-adjusted death rate for Georgia and

TABLE E1.31a

Age (years)	Deaths	Population
0–4	2 483	424 600
5–19	1 818	1 818 000
20–44	3 656	1 126 500
45–64	12 424	870 800
65+	21 405	360 800

TABLE E1.31b

Age (years)	Group Population
0–4	84 416
5–19	294 353
20–44	316 744
45–64	205 745
65+	98 742
Total	1 000 000

compare to those for Alaska and Florida, with the United States population given in Table 1.8, reproduced here as part (b) in Table E1.31, being used as the standard.

(c) Calculate the age-adjusted death rate for Georgia with the Alaska population serving as the standard population. How does this adjusted death rate compare to the crude death rate of Alaska?

1.32 Refer to the same set of mortality data as in Exercise 1.31. Calculate and compare the age-adjusted death rates for the states of Alaska and Florida, with the Georgia population serving as the standard population. How do mortality in these two states compare to mortality in the state of Georgia?

1.33 Some 7000 British workers exposed to vinyl chloride monomer were followed for several years to determine whether their mortality experience differed from those of the general population. In addition to data for deaths from cancers as seen in Example 1.20 (Table 1.23), the study also provided the data shown in Table E1.33 for deaths due to circulatory disease. Calculate the SMR for each subgroup and the relative risk for group "15+" versus group "1–4."

TABLE E1.33

	Years since entering the industry				
Deaths	1–4	5–9	10–14	15+	Total
Observed	7	25	38	110	180
Expected	32.5	35.6	44.9	121.3	234.1

1.34 A long-term follow-up study of diabetes has been conducted among Pima Indian residents of the Gila River Indian Community of Arizona since 1965. Subjects of this study who were at least five years old and of at least half Pima ancestry were examined approximately every two years; examinations included measurements of height and weight and a number of other factors. Table E1.34 relates diabetes incidence rate (new cases/1000 person-years) to body mass index (a measure of obesity defined as weight/height^2). Display these rates by means of a bar chart.

TABLE E1.34

Body mass index	Incidence rate
<20	0.8
20–25	10.9
25–30	17.3
30–35	32.6
35–40	48.5
≥40	72.2

1.35 In the course of selecting controls for a study to evaluate the effect of caffeine-containing coffee on the risk of myocardial infarction among women 30–49 years of age, a study noted appreciable differences in coffee consumption among hospital patients admitted for illnesses not known to be related to coffee use. Among potential controls, the coffee consumption of patients who had been admitted to hospital by conditions having an acute onset (such as fractures) was compared to that of patients admitted for chronic disorders (Table E1.35).

(a) Each of the 6815 subjects above is considered as belonging to one of the three groups defined by the number of cups of coffee consumed per day (the three columns). Calculate for each of the three groups the proportion of subjects admitted because of an acute onset. Display these proportions by means of a bar chart.

(b) For those admitted because of their chronic conditions, express their coffee consumption by means of a pie chart.

(c) Calculate the generalized odds and give your interpretation. Exposure is defined as having an acute condition.

TABLE E1.35

	Cups of coffee per day			
Reason for admission	0	1–4	≥5	Total
Acute conditions	340	457	183	980
Chronic conditions	2440	2527	868	5835

1.36 In a seroepidemiologic survey of health workers representing a spectrum of exposure to blood and patients with hepatitis B virus (HBV), it was found that infection increased as a function of contact. Table E1.36 provides data for hospital workers with uniform socioeconomic status at an urban teaching hospital in Boston, Massachusetts.

TABLE E1.36

Personnel	Exposure	*n*	HBV positive
Physicians	Frequent	81	17
	Infrequent	89	7
Nurses	Frequent	104	22
	Infrequent	126	11

(a) Calculate the proportion of HBV-positive workers in each subgroup.

(b) Calculate the odds ratios associated with frequent contacts (compared to infrequent contacts). Do this separately for physicians and nurses.

(c) Compare the two ratios obtained in part (b). A large difference would indicate a three-term interaction or effect modification, where frequent exposure effects are different for physicians and nurses.

(d) Assuming that the odds ratios for the two groups, physicians and nurses, are equal (in other words, type of personnel is not an effect modifier), calculate the Mantel–Haenszel estimate of this common odds ratio.

1.37 The results of the Third National Cancer Survey have shown substantial variation in lung cancer incidence rates for white males in Allegheny County, Pennsylvania, which may be due to different smoking rates. Table E1.37 gives the percentages of current smokers by age for two study areas.

(a) Display the age distribution for Lawrenceville by means of a pie chart.

(b) Display the age distribution for South Hills by means of a pie chart. How does this chart compare to the one in part (a)?

(c) Display the smoking rates for Lawrenceville and South Hills, side by side, by means of a bar chart.

TABLE E1.37

	Lawrenceville		South Hills	
Age (years)	*n*	%	*n*	%
35–44	71	54.9	135	37.0
45–54	79	53.2	193	28.5
55–64	119	43.7	138	21.7
≥65	109	30.3	141	18.4
Total	378	46.8	607	27.1

1.38 Prematurity, which ranks as the major cause of neonatal morbidity and mortality, has traditionally been defined on the basis of a birth weight under 2500 g. But this definition encompasses two distinct types of infants: infants who are small because they are born early, and infants who are born at or near term but are small because their growth was retarded. *Prematurity* has now been replaced by *low birth weight* to describe the second type, and *preterm* to characterize the first type (babies born before 37 weeks of gestation).

A case–control study of the epidemiology of preterm delivery was undertaken at Yale–New Haven Hospital in Connecticut during 1977. The study population consisted of 175 mothers of singleton preterm infants and 303 mothers of singleton full-term infants. Table E1.38 gives the distribution of age (Table E1.38a) and socioeconomic status (Table E1.38b).

(a) Refer to the age data and choose the "≥30" group as the reference. Calculate the odds ratio associated with every other age group. Is it true, in general, that the younger the mother, the higher the risk?

(b) Refer to the socioeconomic data and choose the "lower" group as the reference. Calculate the odds ratio associated with every other group. Is it true, in general, that the poorer the mother, the higher the risk?

(c) Refer to the socioeconomic data. Calculate the generalized odds and give your interpretation. Does this support the conclusion in part (b)?

TABLE E1.38a

Age (years)	Cases	Controls
14–17	15	16
18–19	22	25
20–24	47	62
25–29	56	122
≥30	35	78

TABLE E1.38b

Socioeconomic level	Cases	Controls
Upper	11	40
Upper middle	14	45
Middle	33	64
Lower middle	59	91
Lower	53	58
Unknown	5	5

1.39 Sudden infant death syndrome (SIDS), also known as sudden unexplained death, crib death, or cot death, claims the lives of an alarming number of apparently normal infants every year. In a study at the University of Connecticut

School of Medicine, significant associations were found between SIDS and certain demographic characteristics. Some of the summarized data are given in Table E1.39. (Expected deaths are calculated using Connecticut infant mortality data for 1974–1976.)

(a) Calculate the standardized mortality ratio (SMR) for each subgroup.

(b) Compare males with females and blacks with whites.

TABLE E1.39

	Number of deaths	
	Observed	Expected
Gender		
Male	55	45
Female	35	45
Race		
Black	23	11
White	67	79

1.40 Adult male residents of 13 counties in western Washington in whom testicular cancer had been diagnosed during 1977–1983 were interviewed over the telephone regarding their history of genital tract conditions, including vasectomy. For comparison, the same interview was given to a sample of men selected from the population of these counties by dialing telephone numbers at random. The data in Table E1.40 are tabulated by religious background. Calculate the odds ratio associated with vasectomy for each religious group. Is there any evidence of an effect modification? If not, calculate the Mantel–Haenszel estimate of the common odds ratio.

TABLE E1.40

Religion	Vasectomy	Cases	Controls
Protestant	Yes	24	56
	No	205	239
Catholic	Yes	10	6
	No	32	90
Others	Yes	18	39
	No	56	96

1.41 The role of menstrual and reproductive factors in the epidemiology of breast cancer has been reassessed using pooled data from three large case–control studies of breast cancer from several Italian regions. The data are summarized

in Table E1.24 for age at menopause and age at first live birth. Find a way (or ways) to summarize the data further so as to express the observation that the risk of breast cancer is lower for women with younger ages at first live birth and younger ages at menopause.

1.42 In 1979 the United States Veterans Administration conducted a health survey of 11 230 veterans. The advantages of this survey are that it includes a large random sample with a high interview response rate, and it was done before the public controversy surrounding the issue of the health effects of possible exposure to Agent Orange. The data in Table E1.42 relate Vietnam service to eight posttraumatic stress disorder symptoms among the 1787 veterans who entered military service between 1965 and 1975. Calculate the odds ratio for each symptom.

TABLE E1.42

	Service in Vietnam	
Symptom	Yes	No
Nightmares		
Yes	197	85
No	577	925
Sleep problems		
Yes	173	160
No	599	851
Troubled memories		
Yes	220	105
No	549	906
Depression		
Yes	306	315
No	465	699
Temper control problems		
Yes	176	144
No	595	868
Life goal association		
Yes	231	225
No	539	786
Omit feelings		
Yes	188	191
No	583	821
Confusion		
Yes	163	148
No	607	864

1.43 It has been hypothesized that dietary fiber decreases the risk of colon cancer, whereas meats and fats are thought to increase this risk. A large study was undertaken to confirm these hypotheses. Fiber and fat consumptions are classified as low or high, and data are tabulated separately in Table E1.43 for males

TABLE E1.43

	Males		Females	
Diet	Cases	Controls	Cases	Controls
Low fat, high fiber	27	38	23	39
Low fat, low fiber	64	78	82	81
High fat, high fiber	78	61	83	76
High fat, low fiber	36	28	35	27

and females ("low" means below median). For each group (males and females), using "low fat, high fiber" as the reference, calculate the odds ratio associated with every other dietary combination. Is there any evidence of an effect modification (interaction between consumption of fat and consumption of fiber)?

1.44 Data are compiled in Table E1.44 from different studies designed to investigate the accuracy of death certificates. The results of 5373 autopsies were compared to the causes of death listed on the certificates. Find a graphical way to display the downward trend of accuracy over time.

TABLE E1.44

	Accurate certificate		
Date of study	Yes	No	Total
1955–1965	2040	694	2734
1970–1971	437	203	640
1975–1978	1128	599	1727
1980	121	151	272

1.45 A study was conducted to ascertain factors that influence a physician's decision to transfuse a patient. A sample of 49 attending physicians was selected. Each physician was asked a question concerning the frequency with which an unnecessary transfusion was given because another physician suggested it. The same question was asked of a sample of 71 residents. The data are shown in Table E1.45.

TABLE E1.45

	Frequency of unnecessary transfusion				
Type of physician	Very Frequently (every week)	Frequently (every two weeks)	Occasionally (every month)	Rarely (every two months)	Never
Attending	1	1	3	31	13
Resident	2	13	28	23	5

(a) Choose "never" as the reference and calculate the odds ratio associated with each other frequency and "residency."

(b) Calculate the generalized odds and give your interpretation. Does this result agree with those in part (a)?

1.46 When a patient is diagnosed as having cancer of the prostate, an important question in deciding on a treatment strategy is whether or not the cancer has spread to the neighboring lymph nodes. The question is so critical in prognosis and treatment that it is customary to operate on the patient (i.e., perform a laparotomy) for the sole purpose of examining the nodes and removing tissue samples to examine under the microscope for evidence of cancer. However, certain variables that can be measured without surgery are predictive of the nodal involvement; and the purpose of the study presented here was to examine the data for 53 prostate cancer patients receiving surgery to determine which of five preoperative variables are predictive of nodal involvement. Table E1.46 presents the complete data set. For each of the 53 patients, there are two continuous independent variables, age at diagnosis and level of serum acid phosphatase (×100; called "acid"), and three binary variables: x-ray reading, pathology reading (grade) of a biopsy of the tumor obtained by needle before surgery, and a rough measure of the size and location of the tumor (stage) obtained by palpation with the fingers via the rectum. For these three binary independent variables, a value of 1 signifies a positive or more serious state and a 0 denotes a negative or less serious finding. In addition, the sixth column presents the finding at surgery—the primary outcome of interest, which is binary, a value of 1 denoting nodal involvement, and a value of 0 denoting no nodal involvement found at surgery. In this exercise we investigate the effects of the three binary preoperative variables (x-ray, grade, stage); the effects of the two continuous factors (age, acid phosphatase) will be studied in an exercise in Chapter 2.

(a) Arrange the data on *nodes* and *x-ray* into a 2 × 2 table, calculate the odds ratio associated with x-ray and give your interpretation.

(b) Arrange the data on *nodes* and *grade* into a 2 × 2 table, calculate the odds ratio associated with grade and give your interpretation.

(c) Arrange the data on *nodes* and *stage* into a 2 × 2 table, calculate the odds ratio associated with stage and give your interpretation.

If you use Microsoft Excel to solve this problem, 2 × 2 tables can be formed using *PivotTable Wizard* in the *Data* menu. A SAS program for part (a), for example, would include these intructions:

```
PROC FREQ;
   TABLES NODES*XRAY/ OR;
```

TABLE E1.46

X-ray	Grade	Stage	Age	Acid	Nodes	X-ray	Grade	Stage	Age	Acid	Nodes
0	1	1	64	40	0	0	0	0	60	78	0
0	0	1	63	40	0	0	0	0	52	83	0
1	0	0	65	46	0	0	0	1	67	95	0
0	1	0	67	47	0	0	0	0	56	98	0
0	0	0	66	48	0	0	0	1	61	102	0
0	1	1	65	48	0	0	0	0	64	187	0
0	0	0	60	49	0	1	0	1	58	48	1
0	0	0	51	49	0	0	0	1	65	49	1
0	0	0	66	50	0	1	1	1	57	51	1
0	0	0	58	50	0	0	1	0	50	56	1
0	1	0	56	50	0	1	1	0	67	67	1
0	0	1	61	50	0	0	0	1	67	67	1
0	1	1	64	50	0	0	1	1	57	67	1
0	0	0	56	52	0	0	1	1	45	70	1
0	0	0	67	52	0	0	0	1	46	70	1
1	0	0	49	55	0	1	0	1	51	72	1
0	1	1	52	55	0	1	1	1	60	76	1
0	0	0	68	56	0	1	1	1	56	78	1
0	1	1	66	59	0	1	1	1	50	81	1
1	0	0	60	62	0	0	0	0	56	82	1
0	0	0	61	62	0	0	0	1	63	82	1
1	1	1	59	63	0	1	1	1	65	84	1
0	0	0	51	65	0	1	0	1	64	89	1
0	1	1	53	66	0	0	1	0	59	99	1
0	0	0	58	71	0	1	1	1	68	126	1
0	0	0	63	75	0	1	0	0	61	136	1
0	0	1	53	76	0						

Note: This is a very long data file; its electronic copy is available in Web-based form from www.wiley.com/go/Le/Biostatistics.