CHAPTER

1

Proportions, Rates, and Ratios

Most introductory textbooks in statistics and biostatistics start with methods for organizing, summarizing, and presenting continuous data–numbers measured on a continuous scale; for example, measurements of height, weight, blood pressure, or cholesterol level. We have decided, however, to adopt a different starting point because our focused areas are research in biomedical sciences, and health decisions are more frequently based on proportions, ratios, or rates. In addition, it is easier to learn methods for categorical data and, therefore, to build knowledge and confidence. In this first chapter we will see how the concepts of proportion, rate, and ratios appeal to common sense, and learn their meaning and uses. You will be able to apply what you learn quickly to tackle real-life data, concepts, and applications—concepts such as case–control study, measures of morbidity, and mortality—before moving on to more complicated methods for continuous data in Chapter 2.

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 Not-improved. The resulting data are called "binary data" or "dichotomous data." 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 always, 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 individuals, the number "x" of positive outcomes is not sufficient; the group size "n", or total number of observations, should also be recorded and used. The number "x" tells us very little and only becomes meaningful after adjusting for the size "n" of the group; in

Health and Numbers: A Problems-Based Introduction to Biostatistics, Third Edition, by Chap T. Le

other words, the two figures "x" and "n" are often combined into a *statistic*, called *proportion*:

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

The term "statistic" means a summarized figure from observed data; a summary, some calculated number. It can be easily seen that $0 \le p \le 1$ because the number "x" of positive outcomes cannot be negative and, as the size of a subgroup, it cannot be greater than "n". This proportion "p" is sometimes expressed as a percentage and is calculated as follows:

$$\% = \left(\frac{x}{n}\right) 100\%$$

Example 1.1

A study, published several years ago by the Urban Coalition of Minneapolis and the University of Minnesota Adolescent Health Program, surveyed 12,915 students in Grades 7 through 12 in Minneapolis and St. Paul public schools. The report said 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 they had not seen a doctor or a dentist in the last two years, followed by 17.7% of American Indians, 16.1% of Blacks, and 10% of Hispanics. Among Whites, it was 6.5%.

Proportion is a number used to *describe* a group of individuals according to a dichotomous characteristic under investigation. The following are a few illustrations of its use in the health sciences.

1.1.1. Comparative Studies

Comparative studies are intended to show possible *differences* between two or more groups. For example, the same survey of Example 1.1 also provided the following figures concerning boys in the surveyed 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 American Indians.

In addition to surveys which are cross-sectional, as seen in the above example, 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 the exposure to a suspected *risk factor*. They are commonly referred to as *case–control studies*. In a 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 individuals 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 it 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:

(a) Diagnoses since 1970 at the single large hospital in Brunswick.
(b) Diagnoses during 1975–1976 at three major hospitals in Savannah.
(c) 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 as follows:

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

The exposure under investigation, "Shipbuilding," refers to employment in shipyards during World War II. By 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 may be 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 *risk factor*; however, the term would be also even used in studies where the factor under investigation has beneficial effects.

In an examination of the smokers in the above data set, 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, we have

(1a) For the controls,

$$\begin{aligned}\text{Proportion of exposure} &= \frac{45}{315}\\ &= .143 \text{ or } 14.3\%\end{aligned}$$

(2a) For the cases,

$$\begin{aligned}\text{Proportion of exposure} &= \frac{84}{397}\\ &= .212 \text{ or } 21.2\%\end{aligned}$$

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

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

(1b) For the controls,

$$\begin{aligned}\text{Proportion of exposure} &= \frac{35}{238}\\ &= .147 \text{ or } 14.7\%\end{aligned}$$

(2b) For the cases,

$$\begin{aligned}\text{Proportion of exposure} &= \frac{11}{61}\\ &= .180 \text{ or } 18.0\%\end{aligned}$$

Again, the results also reveal different exposure histories: the proportion of exposure among cases was higher than that among controls.

The above analyses also show that the difference (cases versus controls) between proportions of exposure among smokers, that is,

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

is different from the difference (cases versus controls) between proportions of 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 above calculation shows that the possible effects of employment in shipyards (as a suspected risk factor) are different for smokers and nonsmokers. This difference of differences (6.9% versus 3.3%), if confirmed, is called an "interaction" or an *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 example is provided in the following example concerning glaucomatous blindness.

Example 1.3

Persons registered blind from glaucoma

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

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

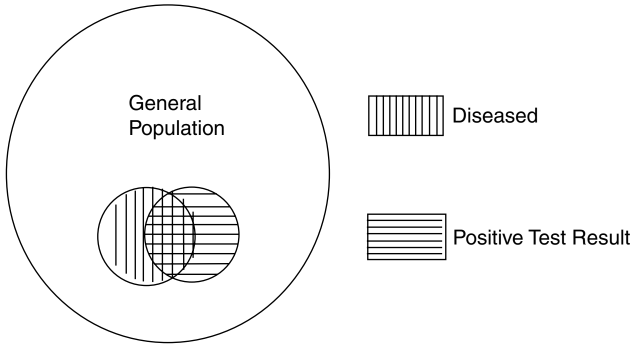
$$\text{Prevalence} = \frac{\text{Number of diseased individuals at the time of investigation}}{\text{Total number of individuals tested}}$$

More details on disease prevalence and related concepts are in Section 1.2.2.

For blindness from glaucoma, calculations in this example reveal a striking difference between the races: the blindness prevalence among Nonwhites was over 8 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 saying that the prevalence was .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, clinical observation or laboratory techniques, individuals 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 individuals would occasionally be classified wrongly as being ill, while some individuals who are really ill may fail to be detected. That is, misclassification—in either way—is unavoidable. Suppose that each individual 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 which emerges after passage of time, for instance at autopsy. For each class of individuals, diseased and healthy, the test is applied and results are depicted as follows:



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

$$\text{Sensitivity} = \frac{\text{Number of diseased individuals who screen positive}}{\text{Total number of diseased individuals}}$$

(the corresponding errors are *false negatives*), where as the specificity is the proportion of healthy individuals detected as negative:

$$\text{Specificity} = \frac{\text{Number of healthy individuals who screen negative}}{\text{Total number of healthy individuals}}$$

(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.

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 the table below (This study was performed for a rather old test; it is used here only for illustration.):

	Test		
Truth	Negative	Positive	Total
Negative	23,362	362	23,724
Positive	225	154	379

The following calculations

$$\begin{aligned}\text{Sensitivity} &= \frac{154}{379}\\ &= .406 \quad \text{or} \quad 40.6\%\end{aligned}$$

$$\begin{aligned}\text{Specificity} &= \frac{23{,}362}{23{,}724}\\ &= .985 \quad \text{or} \quad 98.5\%\end{aligned}$$

show that the above test is highly specific (98.5%) but not very sensitive (40.6%); there were more than half (100% − 40.6% = 59.4%) 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 lead to 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% American Indians, and 4% other or unknown.

1.1.3. Displaying Proportions

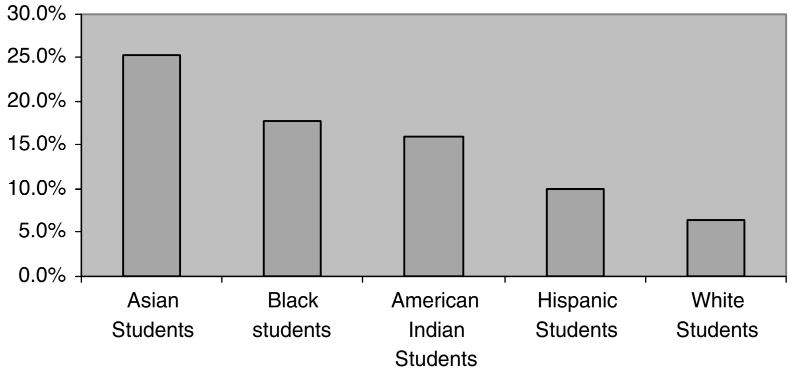
Perhaps the most effective and most convenient way of organizing, summarizing, and then presenting data, especially discrete or categorical data (observations or outcomes belong to one of two or several categories), is through the use of *graphs*. Graphs convey the information, the general patterns in a set of data, at a single glance. Graphs are often easier to read than *tables*; the most informative graphs are simple and self-explanatory. Of course, in order to achieve that objective, graphs should be carefully constructed. 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 with *title* and *label*, and they should require little or no additional explanation in the main text.

Bar Charts

Bar charts are a very popular type of graph used to *display several proportions*—one for each of several groups—for quick comparison. In a bar chart, the various groups are represented along the horizontal axis; they may be arranged alphabetically, or 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 kids without a recent physical checkup (Example 1.1) by a bar chart as follows:



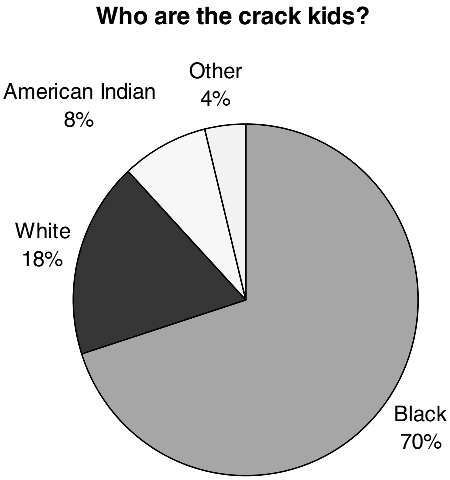
Adolescents without a recent physical checkup

Pie Charts

Pie charts are another popular type of graph. 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 expenditures on health care and defense in the United States. 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, which 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 the crack kids of Example 1.5 by a pie chart as follows:



Example 1.8

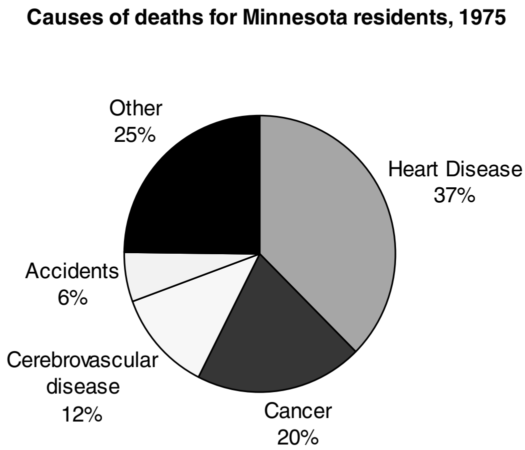
The following table provides the number of deaths due to different causes among Minnesota residents for the year 1975.

Cause of Death	Number of Deaths
Heart disease	12,378
Cancers	6,448
Cerebrovascular disease	3,958
Accidents	1,814
Others	8,088
Total	32,686

After calculating the proportion of deaths due to each cause, for example

$$\begin{aligned}\text{Deaths due to cancers} &= \frac{6{,}448}{32{,}686} \\ &= .197 \text{ or } 19.7\%\end{aligned}$$

We can present the results as in the following pie chart:



Line Graphs

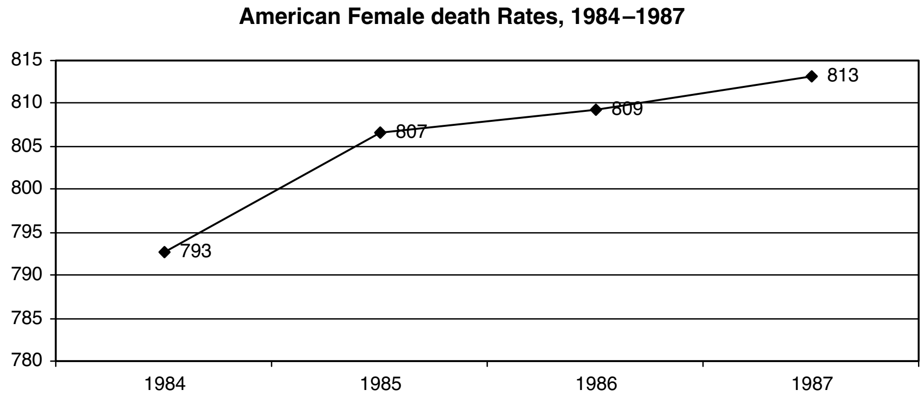
A *line graph* is similar to a bar chart but the horizontal axis represents *time*. Different "groups" are consecutive years so that a line graph is suitable to illustrate how certain *proportions change over time*. In a line graph, the proportion associated with each year is represented by a point at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

Example 1.9

Between the years 1984 and 1987, the *crude death rates* for females in the United States are as follows:

Year	Crude Death Rate Per 100,000 Population
1984	792.7
1985	806.6
1986	809.3
1987	813.1

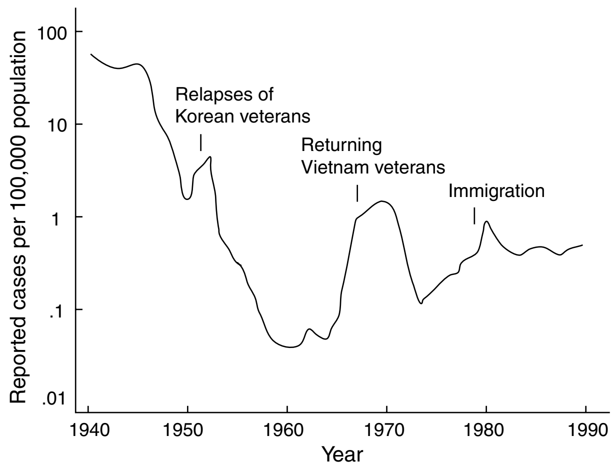
The change in crude death rate, per 100,000 population, for U.S. females can be represented by the following line graph:



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

Example 1.10

The following line graph displays the trend in the reported rates of malaria that occurred in the United States between 1940 and 1989 (proportion times 100,000 as above).



Malaria rates in the U.S., 1940–1989

1.2. RATES

The term *rate* is kind of confusing; sometimes it is used interchangingly with the term *proportion* as defined in the previous section, sometimes it refers to a quantity of very different nature. Section on the *change rate* covers this special use and the next two Sections 1.2.2 and 1.2.3 focus on rates used interchangingly with proportions as *measures of morbidity and mortality*. Even when they refer to the same things—measures of morbidity and mortality, there is still some degree of difference between these two terms; as 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 0 and 100%). *Change rates* are mostly used for description and 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 were 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 this example:

(1) The change rate for new AIDS cases was calculated as

$$\frac{35{,}238 - 32{,}196}{32{,}196} \times (10)\% = 9.4\%$$

(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

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

leading to

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

(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:

$$34\% = \frac{32{,}196 - (\text{Number of cases in 1987})}{\text{Number of cases in 1987}} \times (100)\%$$

or

$$\text{Number of cases in 1987} = \frac{32{,}196}{1.34} = 24{,}027$$

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

$$\frac{3{,}971}{35{,}238} = .113 \quad \text{or} \quad 11.3\%$$

and the proportion of males is

$$\frac{35{,}238 - 3{,}971}{35{,}238} = .887 \quad \text{or} \quad 88.7\%$$

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 some special applications of rates, three kinds 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, sex, 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 different age distributions.

The annual *crude death rate* is defined as the number of deaths in a calendar year, divided by the population on July 1 of that year (which is usually an estimate); the quotient is often multiplied by 1000, or another suitable power of 10, resulting in a number between 1 and 100, or 1 and 1000. For example, the 1980 population of California was 23,000,000 (as estimated by July 1) and there were 190,247 deaths during 1980, leading to

$$\text{Crude death rate} = \frac{190{,}247}{23{,}000{,}000} \times (1000) = 8.3 \text{ deaths per 1000 persons per year}$$

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

As for *morbidity*, the *disease prevalence*, as defined in the previous section, 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:

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

In other words, prevalence presents a snapshot of the population's morbidity experience at certain time point whereas the *incidence* is aimed to investigate possible time trends. For example, the 35,238 new AIDS cases in Example 1.7 and the national population without AIDS at the start of 1989 could be combined according to the above formula to yield an incidence of AIDS for the year.

Another interesting use of rates is in connection with *cohort studies*. Cohort studies are *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 the 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 observed number of deaths for the cohort, divided by the total follow-up times (in *person-years*, say) is the *rate* to characterize the mortality experience of the cohort:

$$\text{Follow-up death rate} = \frac{\text{Number } of \text{ deaths}}{\text{Total person-years}}$$

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-digit or double-digit figure; for example, deaths per 1000 months of follow-up. Follow-up death rates maybe used to measure 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 comorbidities such as arteriosclerotic heart disease, peripheral vascular disease, chronic obstructive pulmonary, and cancer). Results from these two subgroups were as follows (only some summarized figures are given here for illustration; details such as numbers of deaths and total treatment months for subgroups are not included—some detailed data can be found in Exercise 1.30):

Group	Age Group	Deaths/1000 Treatment Months
Low risk	1–45	2.75
	46–60	6.93
	61 +	13.08
Diabetics	1–45	10.29
	46–60	12.52
	61 +	22.16

For example, for the low-risk patients over 60 years of age, there were 38 deaths during 2906 treatment months leading to:

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

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, the comparisons of crude rates are often invalid because the populations may be different with respect to an important characteristic such as age, sex or race (these are *potential confounders*). To overcome this difficulty, adjusted (or standardized) rates are used in the comparison; the adjustment removes the difference in composition with respect to a confounder, which is often *age*.

Example 1.13

The following table provides mortality data for Alaska and Florida for the year 1977.

	Alaska			Florida		
Age Group	No. of Deaths	Persons	Deaths/100,000	No. of Deaths	Persons	Deaths/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

(continued)

(*Continued*)

	Alaska			Florida		
Age Group	No. of Deaths	Persons	Deaths/100,000	No. of Deaths	Persons	Deaths/100,000
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	1101.5
65 +	444	9,000	4933.3	63,505	1,444,000	4397.9
Total	1613	407,000	396.3	91,750	8,455,000	1085.2

This example shows that the 1977 crude death rate per 100,000 population for Alaska was 396.8 and for Florida was 1085.7—almost a threefold difference. However, a closer examination shows:

(1) Alaska had higher *age-specific death rates* for 4 of the 5 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 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 population of the United States as of the last decennial census is frequently used. The procedure consists of the following steps:

(1) The standard population is listed by the same age groups.
(2) *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 U.S. population for 1970 was 84,416 (per 1 million); therefore, we have:
   (i) Alaska rate = 405.0 per 100,000
   The expected number of deaths is

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

   (ii) Florida rate = 375.3 per 100,000
   The expected number of deaths is

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

   (iii) The expected number of deaths for Florida is lower than the expected number of deaths for Alaska obtained for the same age group.

(iv) Obtain total number of expected deaths.
(v) Age-adjusted death rate is

$$\text{Adjusted death rate} = \frac{\text{Total number of expected deaths}}{\text{Standard population size}} \times (100{,}000)$$

Detailed calculations are, using the U.S. population in 1970 as the common standard:

		Alaska		Florida	
Age Group	Standard	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

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 values of the adjusted rates depend in large part on the *choice of the standard population*. They have real meaning only for comparisons.

The advantage of using the U.S. population as the standard is that we can adjust death rates of many states and compare them with each other. Any population could be selected and used as "standard." In the above example, it does not mean that there were only 1 million people in the United States in the year 1970; it only presents the *age distribution* of 1 million U.S. residents for that year. If all we want to do is to compare Florida versus Alaska, we could choose either one of the states as standard and adjust the death rate of the other; this practice would save half of the labor. For example, if we choose Alaska as the standard population then the adjusted death rate for the state of Florida is calculated as follows:

		Florida	
Age Group	(Alaska as) Standard	Specific Rate/100,000	Expected Deaths
0–4	40,000	375.3	150
5–9	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

The new adjusted rate

$$\frac{(1{,}590)(100{,}000)}{407{,}000} = 390.7 \text{ per } 100{,}000$$

is not the same as that obtained using the 1970 U.S. population as 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 two terms, proportion and rate 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-to-female ratio of smoking rates; such a ratio is positive but may well exceed 1.0.

1.3.1. Relative Risk

One of the most often used ratios in epidemiological studies is the *Relative Risk* (RR), a concept for the comparison of two groups or populations with respect to a certain unwanted event (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, 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, which is greater than 1.0 indicates harmful effects whereas a relative risk, which is less than 1.0 indicates beneficial effects. For example, if group 1 consists of smokers and group 2 nonsmokers, then we have a *relative risk due to smoking*. Using the data on ESRD of Example 1.12, we can obtain the following relative risks due to diabetes:

Age Group	Relative Risk
1–45	3.74
46–60	1.81
61 +	1.69

All three numbers are greater than 1 (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 *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 readily obtained by observing the experience of groups of subjects with and without the factor as shown above. 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 an useful short-cut 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, as follows:

	Disease		
Factor	Yes (+)	No (−)	Total
Exposed (+)	A	B	A + B
Unexposed (−)	C	D	C + D
Total	A + C	B + D	N = A + B + C + D

The entries A, B, C and D in the table are sizes of the four combinations of disease presence-and-absence and factor presence-and-absence and the number N at the lower right corner of the table is the total population size. The relative risk is

$$\begin{aligned} \mathrm{RR} &= \frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}} \div \frac{\mathrm{C}}{\mathrm{C}+\mathrm{D}} \\ &= \frac{\mathrm{A}(\mathrm{C}+\mathrm{D})}{\mathrm{C}(\mathrm{A}+\mathrm{B})} \end{aligned}$$

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

$$\mathrm{C}+\mathrm{D} \cong \mathrm{D}$$

$$\mathrm{A}+\mathrm{B} \cong \mathrm{B}$$

($\cong$ means "almost equal to") and, therefore, the relative risk can be approximated as follows:

$$\begin{aligned} \mathrm{RR} &\cong \frac{\mathrm{AD}}{\mathrm{BC}} \\ &= \frac{\mathrm{A}/\mathrm{B}}{\mathrm{C}/\mathrm{D}} = \frac{\mathrm{A}/\mathrm{C}}{\mathrm{B}/\mathrm{D}} \end{aligned}$$

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

(i) A/B and C/D are the *odds* in favor of having disease from groups with or without the factor;
(ii) A/C and B/D are the odds in favor of having exposed to the factors from groups with or without the disease.

The two odds in (ii) can be easily 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 exposed controls in the sample of cases 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. 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

$$\mathrm{RR} = 3.0$$

then we can say that the exposed individuals have a risk of contracting the disease, which is approximately three times the risk of unexposed individuals. 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.05, say), then we can calculate the *odds ratio* as a "stepping stone" and use it as an approximate *relative risk* (we use the notation "$\cong$", meaning *almost equal to*, for this purpose). In these cases, we interpret the calculated odds ratio just as we would do with the relative risk.

Example 1.14

The role of smoking in the etiology of pancreatitis has been recognized for many years. In order 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. The following are some data for the males:

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

For the above data for this example, the approximate relative risks or odds ratios are calculated as follows:

(i) For ex-smokers,

$$\begin{aligned} RR_e &= \frac{13/2}{80/56} \\ &= \frac{(13)(56)}{(80)(2)} \\ &= 4.55 \end{aligned}$$

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

(ii) For current smokers,

$$\begin{aligned} RR_c &= \frac{38/2}{81/56} \\ &= \frac{(38)(56)}{(81)(2)} \\ &= 13.14 \end{aligned}$$

(The subscript "c" in the notation RR_c indicates that we are calculating the 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 same risk for people who never smoke. The effect for ex-smokers is smaller (4.55 times) but it is still very high (as compared to 1.0—the no-effect baseline for relative risks and odds ratios). In other words, if the smokers 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. The Mantel–Haenszel Method

In most in investigation, we are concerned with one primary outcome, such as a disease, and 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 a 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 confounder; it has been found to be associated with lung cancer and it may be associated with employment because construction workers are likely to be smokers. Specifically, we want to know the following:

(i) Among smokers, whether or not shipbuilding and lung cancer are related.
(ii) Among nonsmokers, whether or not shipbuilding and lung cancer are related.

In fact, the original data were tabulated separately for three smoking levels (none, moderate, and heavy); in Example 1.2, the last two tables were combined and presented together for simplicity. Assuming that the confounder, smoking, is not an effect modifier (i.e., smoking does not alter the relationship between lung cancer and shipbuilding); however, 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 is 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 a particular level of the confounder, we have

	Disease Classification		
Exposure	Yes (+)	No (−)	Total
Yes (+)	a	b	$a + b$
No (−)	c	d	$c + d$
Total	$a + c$	$b + d$	n

Because 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 (some kind of weighted average of level-specific odds ratios):

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

The summations, in both numerator and denominator, are across all levels of the confounder.

Example 1.15

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 as follows:

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

There are three 2×2 tables, one for each level of smoking.

(1) We begin with the 2×2 table for *nonsmokers*:

Smoking	Shipbuilding	Cases	Controls	Total
No	Yes	11 (a)	35 (b)	46
	No	50 (c)	203 (d)	253
	Total	51	238	299 (n)

we have, for the nonsmokers,

$$\begin{aligned}\frac{ad}{n} &= \frac{(11)(203)}{299} \\ &= 7.47\end{aligned}$$

$$\begin{aligned}\frac{bc}{n} &= \frac{(35)(50)}{299} \\ &= 5.85\end{aligned}$$

The process is repeated for each of the other two smoking levels.

(2) For moderate smokers

Smoking	Shipbuilding	Cases	Controls	Total
Moderate	Yes	70 (a)	42 (b)	112
	No	217 (c)	220 (d)	437
	Total	287	262	549 (n)

$$\begin{aligned}\frac{ad}{n} &= \frac{(70)(220)}{549} \\ &= 28.05\end{aligned}$$

$$\begin{aligned}\frac{bc}{n} &= \frac{(42)(217)}{549} \\ &= 16.60\end{aligned}$$

(3) And for heavy smokers

Smoking	Shipbuilding	Cases	Controls	Total
Heavy	Yes	14 (a)	3 (b)	17
	No	96 (c)	50 (d)	146
	Total	110	53	163 (n)

$$\frac{ad}{n} = \frac{(14)(50)}{163} = 4.29$$

$$\frac{bc}{n} = \frac{(3)(96)}{163} = 1.77$$

The above results from the three levels of the confounder (smoking) are combined to estimate the common odds ratio:

$$\mathrm{OR}_{\mathrm{MH}} = \frac{7.47 + 28.05 + 4.28}{5.85 + 16.60 + 1.77} = 1.64$$

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 another similar example aiming at the possible effects of oral contraceptive (OC) use on myocardial infarction (MI).

Example 1.16

A case–control study was conducted to investigate the relationship between MI and OC use. The data, stratified by cigarette smoking, were

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

There are two 2×2 tables, one for each level of smoking.

(1) We begin with the 2 × 2 table for *nonsmokers*:

Smoking	OC Use	Cases	Controls	Total
No	Yes	4 (*a*)	52 (*b*)	56
	No	34 (*c*)	754 (*d*)	788
	Total	38	806	844

$$\begin{aligned}\frac{ad}{n} &= \frac{(4)(754)}{844}\\ &= 3.57\end{aligned}$$

$$\begin{aligned}\frac{bc}{n} &= \frac{(52)(34)}{845}\\ &= 2.09\end{aligned}$$

(2) And for smokers,

Smoking	OC Use	Cases	Controls	Total
Yes	Yes	25 (*a*)	83 (*b*)	108
	No	171 (*c*)	853 (*d*)	1024
	Total	196	936	1132

$$\begin{aligned}\frac{ad}{n} &= \frac{(25)(853)}{1132}\\ &= 18.84\end{aligned}$$

$$\begin{aligned}\frac{bc}{n} &= \frac{(83)(171)}{1132}\\ &= 12.54\end{aligned}$$

The above results from the two levels of the confounder (smoking) are combined to estimate the common odds ratio:

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

This combined estimation of the odds ratio, 1.53, represents an approximate increase of 53% in myocardial infarction risk for OC users.

1.4. COMPUTATIONAL AND VISUAL AIDS

Much of this book is concerned with arithmetic and algebraic procedures for data analysis. In many biomedical investigations, particularly those involving large quantities of data, most analyses—for example, regression analysis of Chapter 8—give rise quickly to difficulties in computational implementation. In these investigations it will be necessary to use statistical software specially designed to do these jobs. All calculations described in this book, or any other introductory books can be readily carried out using statistical packages, and any student or practitioner of data analysis will find the use of such packages essential.

There are many specialized packages for statistical analyses; some are very well-known, such as SAS, SPSS, and BMDP. However, students and investigators contemplating to use of one of these commercial programs should read the specifications of each program—may be with help from course instructor—before choosing options needed or suitable for any particular arithmetic or algebraic procedure. But our book is very different; all calculations described in this book can be readily carried out using *Microsoft Excel*, very user-friendly and very popular software available in every personal computer. Notes on the use of Excel are included in separate sections at the end of each chapter.

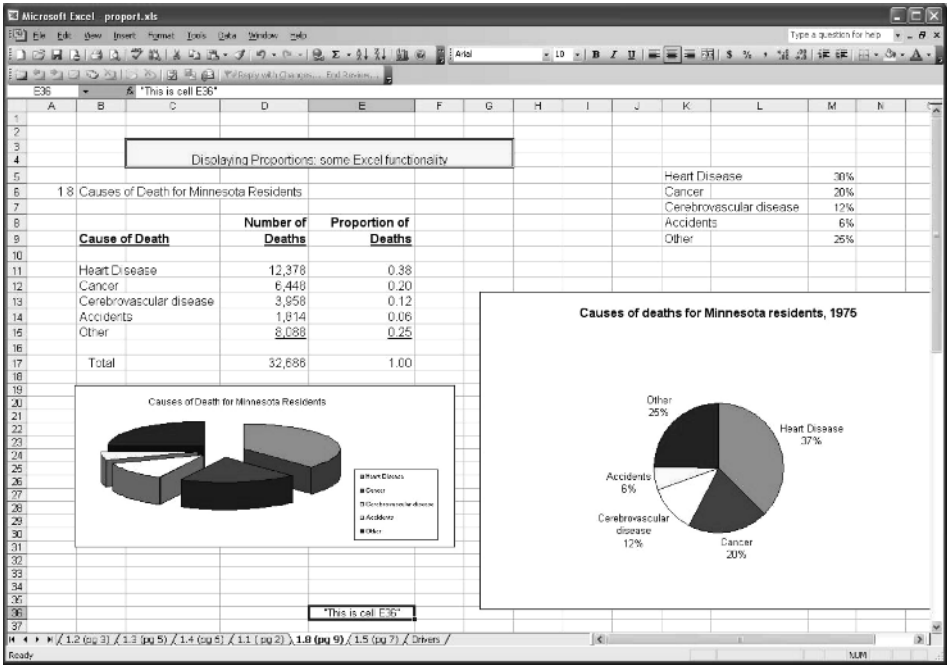
A *worksheet* or *spreadsheet* is a (blank) sheet where you do your work. An Excel *file* holds a stack of worksheets in a *workbook*. You can *name* a sheet, put data on it and *save*; later, *open* and use it. You can *move* or size your windows by dragging the borders. You can also scroll up and down, 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* (A, B, C, AA, BB, . . .) and rows are *numbered* (1, 2, . . .). The intersection of each column and row is a box called a *cell* (a square). Every cell has an *address*, also called *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 column A and B.

You can click a cell to make it *active* (for use, for example, to type a number in it); 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 left most cell and dragging the mouse to the lower right most cell. To move around inside a selected range, press Tab or Enter to move forward one cell at a time.

Excel is software 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 questionnaire and receive returns from 75 people; your data require a file with 75 rows and 10 columns—not counting labels or other identifications (a column for subjects' ID and a row for factors' names). If you made an error, it can fixed; for example, to hit the *Del key* that wipes out the cell contents. You can change your mind again, deleting the delete by clicking the *Undo button* (or *backspace*, reversed curved arrow). Remember, you can widen your columns by double-clicking their right borders or dragging a right border further to the right.

1.4.1. Computer Screen

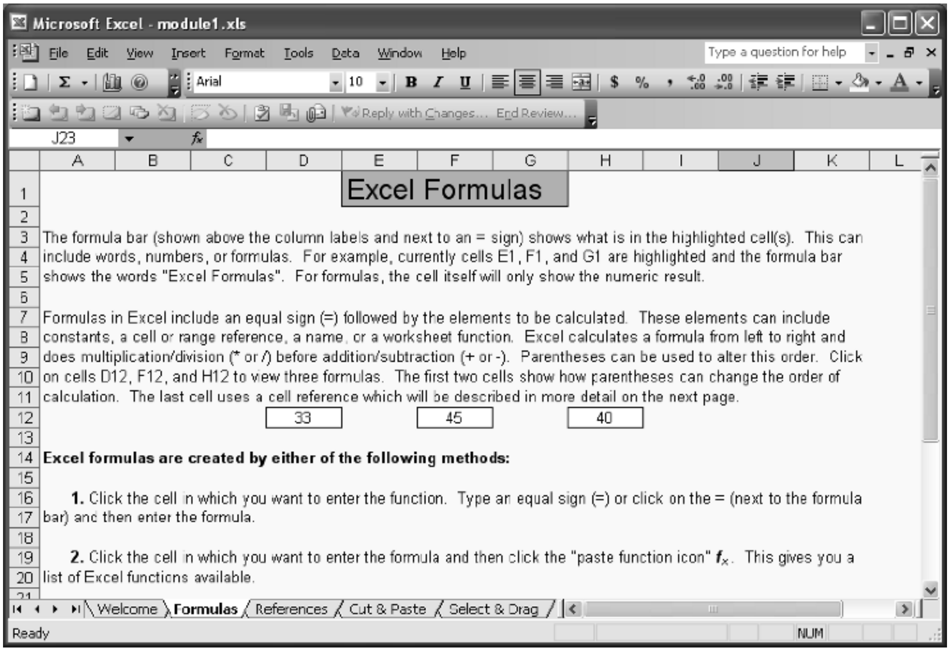


1.4.2. Formula Bar

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 and should be used to change the order of calculations. To use formulas (e.g., for data transformation), do it in one of two ways: (i) 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 (ii) click the cell you want to fill, then click the paste function icon, marked as "f*", which will give you—in a box—a list Excel functions available for your use.

1.4.3. Cut and Paste

The *cut and paste* procedure greatly simplifies the typing necessary to form a chart or table or 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 or cells to which the information is to be placed, and clicking on the paste button (clipboard and page icon).

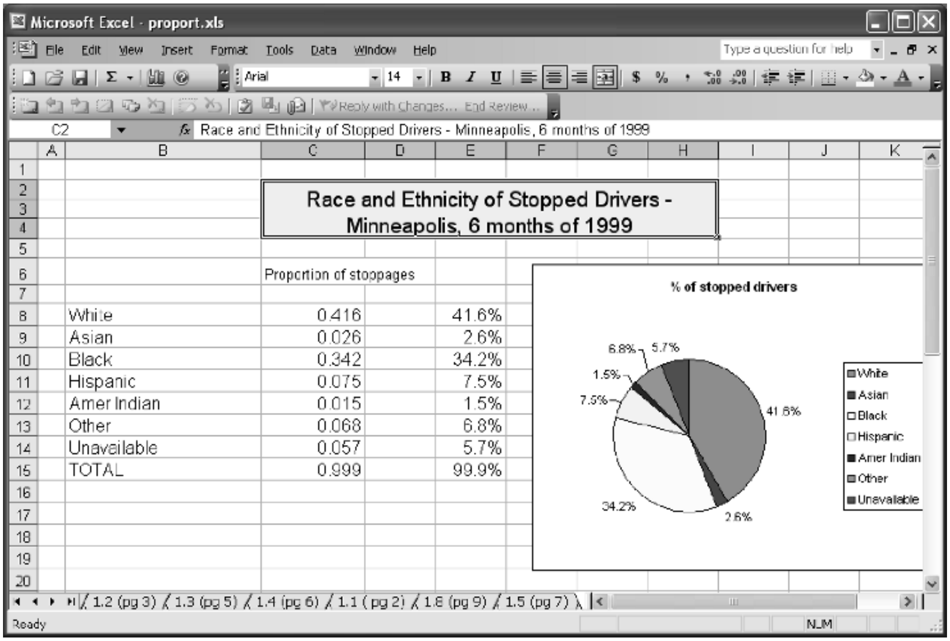


1.4.4. Select and Drag

The *select and drag* is very efficient for *data transformation.* Suppose you have the weight and height of 15 men, the weights are in C6:C20 and the 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 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 and once you release the button, the cells will be filled.

1.4.5. Bar and Pie Charts

Forming a *bar chart* or pie to display proportions is a very simple task. Just click any blank cell to start (you have to do this so the computer knows where to put your resulting figure—even tough you can move it elsewhere later), you will 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 chard, and line chart; the list is on the left side. Choose your chart type and follow instructions. There are many choices, including three-dimensional (3D) ones. You can put data and charts side by side for impressive presentation.



1.4.6. Rate Standardization

This is a good problem to practice with Excel: use it as a calculator. Recall this example:

(i) Florida's rate = 1085.3
(ii) Alaska's rate = 396.8

If Florida has Alaska's population:

		Florida	
Age Group	Alaska's	Age-Specific Rate	Expected Deaths
0–4	40,000	375.3	150
5–9	128,000	60.3	77
20–44	172,000	190.5	328
45–64	58,000	1101.5	639

(continued)

(Continued)

		Florida	
65 +	9,000	4397.9	396
Total	407,000		1590
Age adjusted rate for Florida =			390.6

You proceed with these steps:

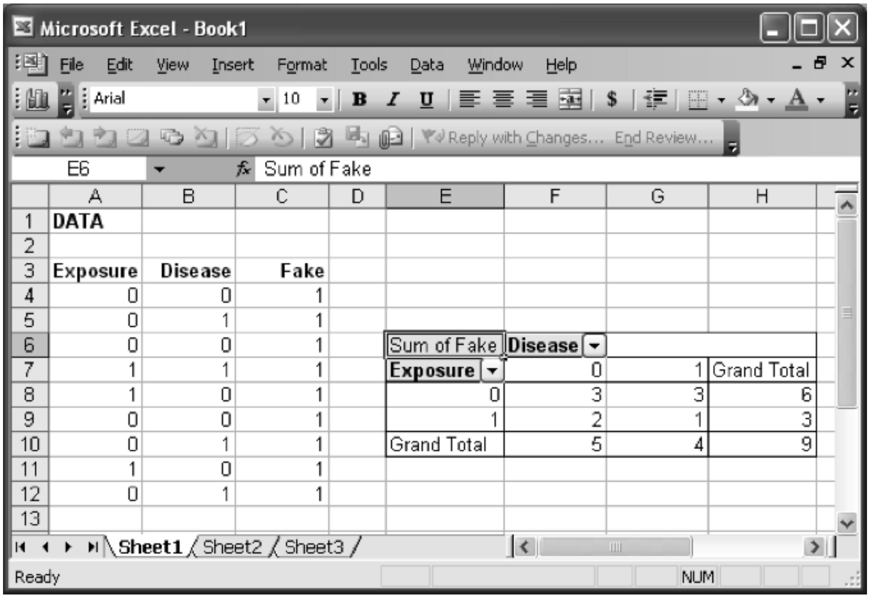
(1) Use *formula* to calculate the first expected number.
(2) Use *drag and fill* to obtain other expected numbers.
(3) Select last column, then click *Autosum icon* (Σ) to add the numbers of expected deaths.

1.4.7. Forming 2 × 2 Tables

Recall that in a *data file* you have one row for each subject and one column for each variable. Suppose two of those variables are *categorical*, say binary (with two categories coded as 1 for "yes" and 0 for "no"), and you want to form a 2 × 2 table (a table for 2 rows for Factor A, and 2 columns for Factor B) so you can study their relationship—for example, to calculate an odds ratio to measure the strength of the association between factors A and B. To achieve that, you can simply follow the following steps; as noted later, step 0 is only optional:

(0) Create a (dummy factor), call it—say- *fake* or any other name, and fill up that column with "1" (you can enter "1" for the first subject, then select and drag).
(1) 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 *PivotTable Report*. Click "next" (to indicate that data are here, in Excel, then *highlight* the area containing your data (including variable names on first row—use the mouse; or you could identify the *range of cells*—say, C5:E28, as a response to 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*.
(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, besides exposure and disease, available in the data set; even a column for names or ID; then there is no need to create the dummy factor "fake." 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 and then choose "count" (to replace "sum").



EXERCISES

1.1. Self-reported injuries among left-handed and right-handed people were compared in a survey of 1896 college students in British Columbia, Canada. Ninety-three of the 180 left-handed students 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 into 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 in order to evaluate the hypothesis that tea consumption and premenstrual syndrome are associated. One hundred eighty-eight 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 into 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:

Condom Use	Cases	Controls
Never	176	488
Ever	51	186

Compute the proportion of subjects in each group who never used condoms.

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

(a) consisting of redness, tearing, and pain in one or both eyes for more than 3 days duration,

(b) having diagnosed as EKC by an ophthalmologist.

In late October 1977, one of the two ophthalmologists (Physician A) 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 individuals came to Physician A for diagnosis and treatment). The following table provides exposure history of 22 persons with EKC between October 27, 1977 and January 13, 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.

Exposed Cohort	Number Exposed	Number Positive
Nursing Home A	64	16
Nursing Home B	238	6

Compute and compare the proportions of cases from the two nursing homes; what would be your conclusion?

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.

The following table gives the rate of positive tuberculin reactions, determined for various groups of students according to degree of exposure to the index case.

Exposure Level	Number Tested	Number Positive
High	129	63
Low	325	36

(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 (a)

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 caesarean section deliveries. Of the 5824 infants included in the study, 2850 were electronically monitored and 2974 were not. The outcomes are as follows:

Caesarean Delivery	EFM Exposure: Yes	EFM Exposure: No	Total
Yes	358	229	587
No	2492	2745	5237
Total	2850	2974	5824

(a) Compute and compare the proportions of caesarean 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 (a)

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 individuals who were involved in bicycle accidents during a 1-year period.

Head Injury	Wearing Helmet: Yes	Wearing Helmet: No	Total
Yes	17	218	237
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 (a)

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

(1) Data for Men

	Myocardial Infarction		Coronary Death	
Drink in the Last 24 h	Controls	Cases	Controls	Cases
No	197	142	135	103
Yes	201	136	159	69

(2) Data for Women

	Myocardial Infarction		Coronary Death	
Drink in the Last 24 h	Controls	Cases	Controls	Cases
No	144	41	89	12
Yes	122	19	76	4

(a) Refer to myocardial infarction (tables on the left); calculate the odds ratio associated with drinking, separately for men and women.
(b) Compare the two odds ratios in (a); when the difference is properly confirmed, we have an effect modification.
(c) Refer to coronary deaths (tables on the right); calculate the odds ratio associated with drinking, separately for men and women.
(d) Compare the two odds ratios in (c); when the difference is properly confirmed, we have an effect modification.

1.9. Data taken from a study to investigate the effects of smoking on cervical cancer are stratified by the number of sexual partners. Results are as follows:

		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

(a) Calculate the odds ratio associated with smoking, separately for the two groups, those with zero or one partners and those with two or more partners.
(b) Compare the two odds ratios in (a); when the difference is properly confirmed, we have an effect modification.
(c) Assuming that the odds ratios for the two groups, those with zero or one partners 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.

1.10. The following table provides the proportions of currently married women having an unplanned pregnancy; data are tabulated for several different methods of contraception.

Method of Contraception	Proportion with Unplanned Pregnancy
None	.431
Diaphragm	.149
Condom	.106
IUD	.071
Pill	.037

Display these proportions in a bar chart.

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

Smoking Status	CHD Deaths/1000 years	Lung Cancer Deaths/1000 years
Never-smoker	2.22	0
Ex-smoker	2.44	.43
Smoker 1–19 cigarettes/day	2.56	.22
Smoker 20–29 cigarettes/day	4.45	1.29
Smoker 40 + cigarettes/day	3.08	1.45

For each cause of death, display the rates in a bar chart.

1.12. The following table 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.

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

Display the proportions of the three response categories for each group, Blacks and Whites, in separate pie charts.

1.13. The following frequency distribution provides the number of cases of pediatric AIDS between 1983 and 1989.

Year	Number of Cases
1983	122
1984	250
1985	455
1986	848
1987	1412
1988	2811
1989	3098

Display the trend of numbers of cases using a line graph.

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 below.

Age	Race	Percent with Breast Exam	
		1973	1985
Total	Black	61.7	74.8
	White	65.9	69.0
20–39 years	Black	77.0	83.9
	White	77.6	77.0
40–59 years	Black	54.8	67.9
	White	62.9	67.5
60–79 years	Black	39.1	64.5
	White	44.7	55.4

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. Display the same data using a line graph.

1.15. Consider the following data:

	Tuberculosis		
X-Ray	No	Yes	Total
Negative	1739	8	1747
Positive	51	22	73
Total	1790	30	1820

Calculate the sensitivity and specificity of X-ray as a screening test for tuberculosis.

1.16. Sera from a T-lymphotropic virus type (HTLV-I) risk group (prostitute women) were tested with two commercial “research” enzyme-linked immuno-absorbent assays (EIA) for HTLV-I antibodies. These results were compared with a gold standard, and outcomes are shown below.

	Dupont’s EIA		Cellular Product’s EIA	
Truth	Positive	Negative	Positive	Negative
Positive	15	1	16	0
Negative	2	164	7	179

Calculate and compare the sensitivity and specificity of these two EIAs.

1.17. The following table provides the number of deaths for several leading causes among Minnesota residents for the year 1991; only the death rate for heart disease is given (number of deaths per 100,000 population).

Cause of Death	Number of Deaths	Deaths/100,000
Heart disease	10,382	294.5
Cancer	8,299	?
Cerebrovascular disease	2,830	?
Accidents	1,381	?
Other causes	11,476	?
Total	34,368	?

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

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

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

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 Example 1.6.

1.19. A case–control study was conducted relating to the epidemiology of breast cancer and the possible involvement of dietary fats, along with other vitamins and nutrients. It included 2024 breast cancer cases that were admitted to Roswell Park Memorial Institute, Erie County, New York, from 1958 to 1965. A control group of 1463 was chosen from the 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). The following are data for 1500 women over 54 years of age.

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

Calculate the odds ratio associated with a decrease in ingestion of foods containing vitamin A.

1.20. Refer to the data set in Example 1.2,

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

(a) calculate the odds ratio associated with employment in shipyards for nonsmokers,

(b) calculate the same odds ratio for smokers. Compare the results with those of (a); when the difference is properly confirmed, we have an "effect modification," where smoking alters the effect of employment in shipyards as a risk for lung cancer, and

(c) assuming that the odds ratios for the two groups, nonsmokers and smokers, are equal, 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 U.S. population) were used to examine the relationship between Pap testing and some socioeconomic factors. The following table provides the percentages (%) of women who reported *never having had* a Pap test (these are from metropolitan areas):

Age	Income	Whites	Blacks
25–44	Poor	13	14.2
	Nonpoor	5.9	6.3
45–64	Poor	30.2	33.3
	Nonpoor	13.2	23.3
65 and over	Poor	47.4	51.5
	Nonpoor	36.9	47.4

(a) Calculate the odds ratios associated with race (Black versus White) among
(i) 25–44 nonpoor,
(ii) 45–64 nonpoor, and
(iii) 65 + nonpoor.
Briefly discuss a possible effect modification if any.

(b) Calculate the odds ratios associated with income (poor versus nonpoor) among
(iv) 25–44 Black,
(v) 45–64 Black, and
(vi) 65 + Black.
Briefly discuss a possible effect modification if any.

(c) Calculate the odds ratios associated with race (Black versus White) among
(vii) 65 + Poor and
(viii) 65 + Nonpoor
and briefly discuss a possible effect modification.

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

		Daily Alcohol Consumption	
Age	Sample	80 + g	0–79 g
25–44	Cases	5	5
	Controls	35	270
45–64	Cases	67	55
	Controls	56	277
65 and over	Cases	24	44
	Controls	18	129

(a) Calculate the odds ratio associated with high alcohol consumption, separately for the three age groups.
(b) Compare the three odds ratios in (a); when the difference is properly confirmed, we have an effect modification.

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.

1.23. Postmenopausal women who develop endometrial cancer are, on the average, 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 post-menopausal 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 between weight, replacement estrogen therapy, and endometrial cancer; results are as follows.

		Estrogen Replacement	
Weight (kg)	Sample	Yes	No
<57	Cases	20	12
	Controls	61	183
[0,1-4]			
57–75	Cases	37	45
	Controls	113	378
[0,1-4]			
>75	Cases	9	42
	Controls	23	140

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

(b) Compare the three odds ratios in (a); when the difference is properly confirmed, 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.

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. The following are summarized data for age at menopause and age at first live birth.

	Cases	Controls
Age at first Live Birth		
<22	621	898
22–24	795	909
25–27	791	769
>27	1043	775
Age at Menopause		
<45	459	543
45–49	749	803
>49	1378	1167

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

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. The following are parts of the data:

		No. of Men Surveyed	
Factor	Level	Total	No. with Gallstone
Smoking	Never	621	11
	Past	776	17
	Current	1342	33
Alcohol	Never	447	11
	Past	113	3
	Current	2179	47
BMI (kg/m^2)	<22.5	719	13
	22.5–24.9	1301	30
	>24.9	719	18

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

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

1.26. Data were collected from 2197 white ovarian cancer patients and 8893 white controls in 12 different U.S. 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. The following are parts of the data related to unprotected intercourse and to history of infertility.

	Cases	Controls
Duration of Unprotected Intercourse (years)		
<2	237	477
2–9	166	354
10–14	47	91
15 and over	133	174
History of Infertility		
No	526	966
Yes, no drug use	76	124
Yes, drug use	20	11

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), choose the lowest level as the baseline and calculate the odds ratio associated with each other level.

1.27. Post-neonatal 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 thoroughly explored. 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 and the following data are concerned with wheezing lower respiratory tract illnesses (wheezing LRI: No/Yes).

	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
Over 30	110	20	116	25

For each of the two groups, boys and girls, choose the lowest age group as the baseline and calculate the odds ratio associated with each other age group.

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 following data,

	Test Result		Total
Field Loss	Positive	Negative	
Yes	13	7	20
No	413	4567	4980

calculate the sensitivity and specificity of this screening test.

1.29. Recall the news report of Example 1.11: “A total of 35,238 new AIDS cases were 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.”

From the above information, 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 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 had the following follow-up data (in addition to those in Example 1.12):

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
Over 70	17	846

Compute the follow-up death rate for each age group, then choose "21–30" as the baseline and calculate the relative risk associated with each other age group.

1.31. Given the following mortality data for the State of Georgia for the year 1977:

Age Group	No. Deaths	Persons
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

(a) From the above mortality table, calculate the crude death rate for the State of Georgia.

(b) From the above mortality table and the mortality data for Alaska and Florida for the year 1977 (same data as given in Example 1.13 but shown again below),

	Alaska			Florida		
Age Group	No. Deaths	Persons	Deaths/100,000	No. Deaths	Persons	Deaths/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	1101.5
65 +	444	9,000	4933.3	63,505	1,444,000	4397.9
Total	1613	407,000	396.3	91,750	8,455,000	1085.2

calculate the age-adjusted death rate for Georgia and compare them to those for Alaska and Florida, the U.S. population given in Example 1.13, reproduced below, being used as standard.

Age Group	Persons
U.S. Standard Population	
0–4	84,416
5–19	294,353
20–44	316,744
45–64	205,745
65 +	98,742
Total	1,000,000

(c) Calculate, again, 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 the above Exercise 1.30. 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 the state of Georgia?

1.33. 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, at least 5 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. The following table relates diabetes incidence rate (new cases/1000 person-years) to body mass index (BMI, a measure of obesity defined as weight/(height)^2).

Body Mass Index	Incidence Rate
<20	.8
20–25	10.9
25–30	17.3
30–35	32.6
35–40	48.5
Over 40	72.2

Display these rates by means of a bar chart.

1.34. In the course of selecting controls for a study to evaluate 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.

	Cups of Coffee Per Day			
Admission by	0	1–4	5 or more	Total
Acute condition	340	457	183	980
Chronic condition	2440	2527	868	5835

(a) Each of the above 6815 subject 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.

1.35. 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. The following table provides data for hospital workers with uniform socioeconomic status at an urban teaching hospital in Boston, Massachusetts.

Personnel	Exposure	No. Tested	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 (as compared to infrequent contacts); do this separately for physicians and nurses.

(c) Compare the two ratios obtained in (b); a large difference would indicate a “three-term interaction” or “effect modification,” where frequent 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.

1.36. The results of the Third National Cancer Survey have shown substantial variation in lung cancer incidence rates for White males within Allegheny County, Pennsylvania, which may be due to different smoking rates. The following table gives the percentages (%) of current smokers by age for two study areas.

	Lawrenceville		South Hills	
Age (years)	No. Surveyed	Smoking Rate	No. Surveyed	Smoking Rate
35–44	71	54.9	135	37.0
45–54	79	53.2	193	28.5
55–64	119	43.7	138	21.7
Over 65	109	30.3	141	18.4

(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.

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

1.37. 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

(i) "low birth weight" to describe the second type

(ii) "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. The following tables give the distributions of age (first table) and socioeconomic status (second table).

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

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

(a) Refer to the age data (first table) and choose the "30 or over" group as baseline; calculate the odds ratio associated with each other age group. Is this true, in general, that the younger the mother the higher the risk?

(b) Refer to the socioeconomic data (second table) and choose the "lower" group as the baseline, calculate the odds ratio associated with each other group. Is this true, in general, that the poorer the mother the higher the risk?

1.38. Adult male residents of 13 counties of western Washington State, 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 following data are tabulated by religious background.

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

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.

1.39. 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 following are summarized data for age at menopause and age at first live birth.

	Cases	Controls
Age at First Live Birth		
<22	621	898
22–24	795	909
25–27	791	769

(continued)

(Continued)

	Cases	Controls
>27	1043	775
Age at Menopause		
<45	459	543
45–49	749	803
>49	1378	1167

Find a way (or ways) to further summarize the data 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.

1.40. It has been hypothesized that dietary fiber decreases the risk of colon cancer, while 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 for males and females as follows ("low" means below median):

	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

For each group (males and females), using "low fat, high fiber" as the baseline, calculate the odds ratio associated with each other dietary combination. Any evidence of an effect modification (interaction between consumption of fat and consumption of fiber)?

1.41. In 1979, the U.S. 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 recent public controversy surrounding the issue of the health effects of possible exposure to Agent Orange. The following are data relating Vietnam service to eight post-traumatic stress disorder symptoms among the 1787 veterans who entered the military service between 1965 and 1975.

		Service in Vietnam	
Symptom	Level	Yes	No
Nightmares	Yes	197	85
	No	577	925
Sleep problems	Yes	173	160

(continued)

(Continued)

Symptom	Level	Service in Vietnam Yes	No
	No	599	851
Troubled memories	Yes	220	105
	No	549	906
Depression	Yes	306	315
	No	465	699
Temper control problem	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

Calculate the odds ratio for each symptom.

1.42. The following table compiles data 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. The results are:

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

Find a graphical way to display the downward trend of accuracy over time.

1.43. 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 give because another physician suggested it. The same question was asked of a sample of 71 residents. The data were as follows:

Type of Physician	Frequency of Unnecessary Transfusion: Very Frequent (1/week)	Frequent (1/2 weeks)	Occasionally (1/month)	Rarely (1/2 months)	Never
Attending	1	1	3	31	13
Resident	2	13	28	23	5

Choose "never" as the baseline and calculate the odds ratio associated with each other frequency and "residency."

1.44. When a patient is diagnosed as having cancer of the prostate, an important question in deciding on treatment strategy for the patient 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. The following table 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 (multiplied by 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 one signifies a positive or more serious state and a zero 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, and Stage); the effects of the two continuous factors (age and acid phosphatase) will be studied in an exercise in the next chapter.

(a) Arrange the data on Notes 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 Notes and Grade into a 2×2 table, calculate the odds ratio associated with Grade and give your interpretation.

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

(This is a very long data file; its electronic version is available from the author upon request. Two-by-two tables can be easily formed using the method for Excel as explained in Section 1.4)Data

Case	X-Ray	Grade	Stage	Age	Acid	Nodes
1	0	1	1	64	40	0
2	0	0	1	63	40	0
3	1	0	0	65	46	0
4	0	1	0	67	47	0
5	0	0	0	66	48	0
6	0	1	1	65	48	0
7	0	0	0	60	49	0
8	0	0	0	51	49	0
8	0	0	0	66	50	0

(continued)

(Continued)

Case	X-Ray	Grade	Stage	Age	Acid	Nodes
10	0	0	0	58	50	0
11	0	1	0	56	50	0
12	0	0	1	61	50	0
13	0	1	1	64	50	0
14	0	0	0	56	52	0
15	0	0	0	67	52	0
16	1	0	0	49	55	0
17	0	1	1	52	55	0
18	0	0	0	68	56	0
19	0	1	1	66	59	0
20	1	0	0	60	62	0
21	0	0	0	61	62	0
22	1	1	1	59	63	0
23	0	0	0	51	65	0
24	0	1	1	53	66	0
25	0	0	0	58	71	0
26	0	0	0	63	75	0
27	0	0	1	53	76	0
28	0	0	0	60	78	0
29	0	0	0	52	83	0
30	0	0	1	67	95	0
31	0	0	0	56	98	0
32	0	0	1	61	102	0
33	0	0	0	64	187	0
34	1	0	1	58	48	1
35	0	0	1	65	49	1
36	1	1	1	57	51	1
37	0	1	0	50	56	1
38	1	1	0	67	67	1
39	0	0	1	67	67	1
40	0	1	1	57	67	1
41	0	1	1	45	70	1
42	0	0	1	46	70	1
43	1	0	1	51	72	1
44	1	1	1	60	76	1
45	1	1	1	56	78	1
46	1	1	1	50	81	1
47	0	0	0	56	82	1
48	0	0	1	63	82	1
49	1	1	1	65	84	1
50	1	0	1	64	89	1
51	0	1	0	59	99	1
52	1	1	1	68	126	1
53	1	0	0	61	136	1