Chapter 3
Describing and presenting data

All epidemiological studies involve the collection of data on the exposures and outcomes of interest. In a well planned study, the raw observations that constitute the data contain the information that satisfies the objectives of the study. The aim of data analysis is to extract the pertinent information from these raw observations in a concise way.

The first stage in data analysis is the preparation of an appropriate form in which the relevant data can be collected and coded in a format suitable for entry into a computer; this stage is referred to as data processing. The second stage is to review the recorded data, checking for accuracy, consistency and completeness; this process is often referred to as data editing. Next, the investigator summarizes the data in a concise form to allow subsequent analysis—this is generally done by presenting the distribution of the observations according to key characteristics in tables, graphs and summary measures. This stage is known as data reduction. Only after data processing, editing and reduction should more elaborate statistical manipulation of the data be pursued.

3.1 Data processing

All the various steps of data processing should be planned when the study is designed, before any data are collected. All forms used for recording data should be carefully designed and tested to ensure that the data can easily be extracted for processing. This general principle applies to all data-collection forms: questionnaires, as well as forms for recording results from laboratory assays, data extraction forms from hospital notes, etc.

Most epidemiological studies involve the collection of large amounts of data that are not easy to process by hand. Fortunately, microcomputers are now available at reasonable prices (see Chapter 18). Before data collected on a form can be entered into a computer file, they must first be coded. For instance, sex may be coded as 1 for male, or 2 for female; only the number 1 or 2 will be entered in the computer file. Numerical data (e.g., number of children) do not require coding, as the exact number can be entered. Most data-collection forms are designed so that every possible answer is assigned a code. However, even when such 'pre-coded' forms are used, further data coding will still be required for the answers to 'open-ended' questions or to the 'other' category of 'closed-ended' questions. For a more detailed discussion of these issues, see Appendix 2.1 and Section 18.3.6.

For each type of data-collection form, a computer file will be created to enter the data. Data-entry programs can be designed so that the computer screen resembles the layout of the data-collection form. This helps to minimize errors made during data entry.

3.2 Data editing

Basic editing of the data involves checking each variable for illogical or unusual values. For example, sex may be coded 1 for male or 2 for female. Another code, perhaps 9, is used to designate an unknown value. It is preferable to assign specific codes to unknown or missing information than to leave these values blank, as it is impossible to tell whether a blank field corresponds to data that are truly missing or to data that have been omitted in error. A code of zero should, however, be avoided, because missing information may be interpreted by some computers or programs as zero. The range and distribution of each variable should be examined, and any inadmissible values should be checked against the original data forms.

In addition to checking for incorrect or unusual values, the distribution of each variable should be examined to see if it appears reasonable. Such an evaluation may reveal important problems that might not otherwise come to light. It is also necessary to cross-check the consistency of codes for related variables. For instance, males should not have been admitted to hospital for hysterectomy or females for prostatectomy. Careful editing will involve many such consistency checks and is best accomplished by computer programs designed to flag such errors.

Most data-entry programs can check and edit data interactively as they are entered. The computer can be programmed to display an error message on the screen and give an audible warning, allowing inadmissible values to be rejected and corrected immediately. A sophisticated data-entry program can also check for consistency between variables and can eliminate some potential inconsistencies by providing appropriate codes automatically. For example, if a subject is male, the program can automatically supply the correct code for 'Have you ever been pregnant?'. Nevertheless, even with the most sophisticated editing during data entry, it remains essential to edit the data before analysis, to check on their completeness, and to examine the distribution of each variable, as data-entry programs cannot perform these functions (see Section 18.3.6).

To minimize error during the handling of data, three basic precautions are recommended. First, avoid any unnecessary copying of data from one form to another. Second, use a verification procedure during data entry. Data should always be entered twice, preferably by two people; the two data-sets can then be compared and any inconsistencies resolved. Third, check all calculations carefully, either by repeating them or, for example, by checking that subtotals add to the correct overall total. All computer procedures should be tested initially on a small subset of the data and the results checked by hand.

3.3 Data reduction

After the data are edited, they should be examined by means of simple tabulations, graphs and basic summary measures. Different types of data must be presented and summarized in different ways: the correct choice of methods used therefore depends on the type of data collected. The remainder of this chapter describes ways of presenting and summarizing two main types of data: quantitative and categorical (or qualitative) variables.

3.3.1 Quantitative data

Quantitative variables can either have a numerical value along a continuous scale (e.g., age, weight, height), or can be whole numbers representing counts of a particular event (e.g., number of children, number of sexual partners).

Presentation of quantitative data

The frequencies with which the different possible values of a variable occur in a group of subjects is called the frequency distribution of the variable in the group. For example, we may wish to present the distribution of height (in cm) of a sample of 1250 women who were examined in a certain breast-screening clinic. As height is measured on a continuous scale, it can have a large number of distinct values; it is therefore more useful to group the values before presenting the distribution (Table 3.1).

Height (cm)	Number of women	Percentage
145–149	75	6.0
150–154	153	12.2
155–159	261	20.9
160–164	323	25.8
165–169	201	16.1
170–174	144	11.5
175–179	91	7.3
180–184	2	0.2
Total	**1250**	**100.0**

Table 3.1.
Distribution of height in a sample of 1250 women attending a certain breast-screening clinic: hypothetical data.

The percentage frequency distribution shown in the final column allows one to make comparison with distributions in other groups of women. There is no need to calculate the percentages precisely: for example, 56 out of 1250 can often be expressed 4.5%, rather than 4.48%. When percentage frequency distributions are reported on their own, the total number of subjects on which the distribution is based should always be given (in this example, 1250). For instance, it might be misleading to report that 20% of women were between 155 and 159 cm tall if only five women had been measured in total.

There are no universal rules on how the data should be grouped. As a rough guide, the number of groups should be 5–20, depending on the number of observations involved. If the interval chosen is wide, too much detail will be lost; if it is narrow, the table may be difficult to interpret. All intervals should have the same width, although the categories at either extreme may be open-ended (e.g., ≥180 cm). There should be no gaps between the groups. The table should be labelled to show clearly how observations that fall on the boundaries are classified.

A frequency distribution can be displayed graphically as a *histogram*, as shown in Figure 3.1. In this type of graph, the number (or percentage) of observations is plotted for different values, or groups of values, of the variable being studied (in this example, height). In a histogram, it is the *area* of the rectangle, not its height, that represents the frequency—the vertical scale is measured in frequency per unit of value and the horizontal scale in units of value. The larger the sample measured, the narrower the grouping interval that can be chosen, so that the histogram becomes smoother and more closely resembles the distribution of the total population. At the limit, when the width of the intervals is so narrow that they practically correspond to a single unit, the resulting diagram would be a smooth curve.

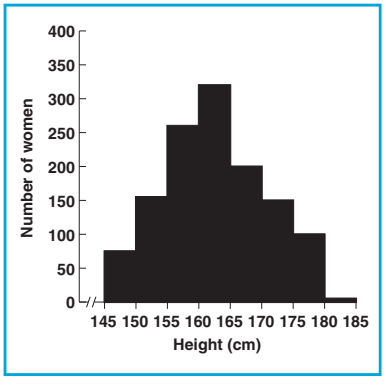


Figure 3.1.
Histogram showing the distribution of height in 1250 women who attended a breast-screening clinic (data shown in Table 3.1).

Summarizing quantitative data

Tabular and graphical methods are an important way of examining the data. But it is also useful to summarize the data numerically. The two most important features of a distribution are usually the central value and the spread about this value.

Measures of central value

(1) The most commonly used measure of the central value of a distribution is the *arithmetic mean*. This is the sum of the observations divided by n, the number of observations. For example, if the weights (in kg) of eight healthy women are

45.3, 49.8, 50.5, 60.7, 65.2, 69.4, 73.2, 75.9,

the arithmetic mean of the weights of these women is

$(45.3 + 49.8 + 50.5 + 60.7 + 65.2 + 69.4 + 73.2 + 75.9)/8 = 490/8 = 61.25 = 61.3$ kg

(2) The *median* is another commonly used measure of central value. It is the value that divides the distribution in half when the observations are ranked in order. Thus, the median is the middle observation. For an even number of observations, the median is calculated as the mean of the two middle values. A general expression for finding the median is:

Median = $(n+1)/2$ th value of the ordered observations,

where n is the total number of observations.

In this example, the value of the median is the 4.5th value, i.e., the average of the fourth and fifth values, (60.7 + 65.2) / 2 = 63.0 kg.

The choice of measure used will depend on the nature of the data and the purpose of the analysis. The mean is often the preferred measure of central value because it takes into account every observation and it is easy to use in the most common types of statistical analysis. Its major disadvantage is that it can be affected by *outliers*—single observations that are extreme in comparison with most observations and whose inclusion or exclusion changes the mean markedly.

The median is a useful descriptive measure when outliers make the mean unrepresentative of the majority of the data. It is also particularly useful when certain observations are not recorded precisely because they are above or below a certain level; in these circumstances, the mean cannot be calculated, but the median can determined so long as definite values are known for more than one half of all subjects. For instance, the mean survival time of a group of cancer patients can be calculated only when all the patients have died; however, the median survival time can be calculated while almost half the patients are alive. The main disadvantage of this measure is that it ranks the data, but does not make use of their individual values.

When the shape of a distribution is roughly symmetric about a central axis, the mean and the median are approximately equal (as is the case for the data on weight given in the example above).

Measures of variation

In addition to a measure of the central value of a distribution, it is also useful to have an idea of the variation or spread of values around this central value. Several measures of variation are used:

(1) The *range* is the interval between the largest and smallest values. The major advantage of this measure is that it is very simple to calculate. The main disadvantage is that it is based only on two extreme observations and gives no indication of how other observations are distributed in between.

(2) The three values that divide a distribution into quarters are called the *quartiles*. Of the total number of observations, 25% lie below the lower quartile, 50% below the middle quartile and 75% below the upper quartile. The middle quartile is the median. The distance between the lower quartile and the upper quartile is called the *inter-quartile range* and is sometimes used to describe variability. The inter-quartile range contains 50% of the observations.

Similarly, *percentiles* are values that divide the distribution into percentages. The 50th percentile corresponds to the median, while the 25th and the 75th percentiles correspond to the lower and upper quartiles, respectively.

A simple but useful semi-graphical way of summarizing data using percentiles is the box-and-whisker plot. Figure 3.2 shows a box-and-whisker

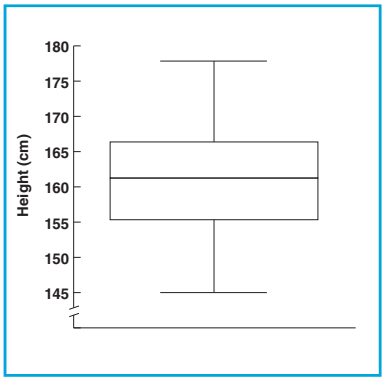


Figure 3.2.
Box-and-whisker plot of the heights of a sample of 1250 women (data shown in Table 3.1).

plot for the data given in Table 3.1. The box indicates the lower and upper quartiles, while the central line is the median. The points at the ends of the 'whiskers' are the 2.5th and the 97.5th percentile values.

(3) The *standard deviation* is the most commonly used measure of the average spread of values about the mean. If the values of a variable do not vary greatly within a population, observations will lie closely around the mean, whereas if there is substantial variation, the observations will be scattered widely about the mean.

This variability or *variance* can be measured in terms of how much, on average, observations differ from the mean: in other words how far, on average, each observation *deviates* from the mean. Figure 3.3 illustrates this for the weights of the eight healthy women given in the example above. The deviations from the mean are shown by the lines d_1, d_2, ... d_8. First, the sum of the these deviations is calculated; however, the sum of the deviations from the arithmetic mean is, by definition, zero, since negative deviations cancel out positive deviations. In calculating the dispersion of values around the arithmetic mean, it is irrelevant whether the deviations are positive or negative: only their absolute numerical magnitude is of interest. Hence, to avoid getting zero when the deviations are added together, the individual deviations are first squared, eliminating negative values. The average of these squared deviations is called the *variance*.

$$\text{Variance} = (d_1^2 + d_2^2 + \ldots + d_8^2)/n$$

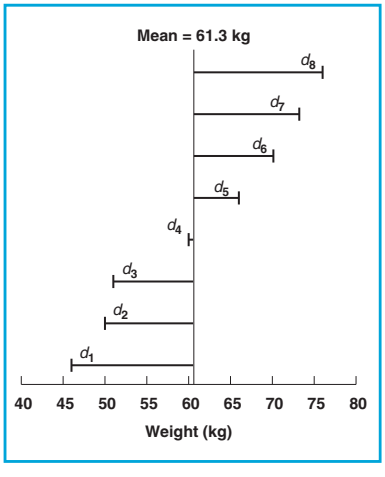


Figure 3.3.
Deviations from the mean of the weights of eight women: hypothetical data.

The variance is a very useful mathematical measure of the average spread of values around the mean, but is difficult to interpret because it is expressed as the square of the units measured. In our example, the variance of weight will be expressed as kg^2. As it is usually more convenient to express the variation in terms of the original, unsquared units, the *square root of the variance* is usually used. This is known as the *standard deviation* (SD). In this example, the SD is equal to 10.8 kg.

A small standard deviation indicates that most values are very close to the mean, whereas a large one indicates that many lie far from the mean: i.e., the more the values in a population vary, the bigger the standard deviation.

As a general rule, provided a distribution is roughly symmetrical and has a bell-like shape, with a dome of central values and two tails at the extremes (characteristic of what statisticians call a 'normal distribution'), the mean is the central value and the standard deviation is a measure of spread such that one standard deviation either side of the mean includes roughly 70% of the observations, and two standard deviations include roughly 95% (Figure 3.4).

3.3.2 Categorical data

The values of categorical (also called qualitative) variables represent attributes, rather than actual measurements of some quantity. The following are examples of categorical variables: sex (male/female), marital status (single/married/divorced/widowed), oral contraceptive use (ever-users/never-users), country of birth (Colombia, Spain, etc.).

There are various types of categorical variable. If the variable can only have two different values, the categorical variable is called binary (e.g., sex). Sometimes the different categories of a variable can be ordered on some scale (e.g., the severity of pain could be categorized as mild, moderate or severe). In this case, the variable is called an ordered categorical variable.

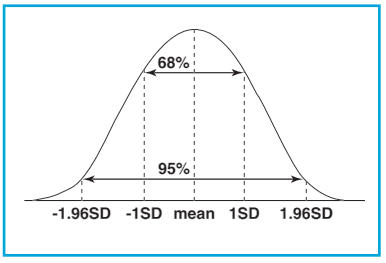


Figure 3.4.
A normal distribution. SD = standard deviation.

Presentation of categorical data

As with quantitative variables, we can present the frequency distribution of a categorical variable as the number of times each value (category) occurs in the group of subjects being studied. Consider Example 3.1:

Example 3.1. *A study was conducted in Spain and Colombia to assess the relationship between cervical cancer and exposure to human papillomavirus (HPV), selected aspects of sexual and reproductive behaviour, use of oral contraceptives, screening practices and smoking. The study included 436 cases of histologically confirmed squamous-cell carcinoma and 387 controls randomly selected from the general population that generated the cases. Each participant responded to a structured questionnaire which collected data on a large number of variables: age, number of sexual partners, educational level, smoking, etc. (Bosch* et al.*, 1992). Table 3.2 shows the distribution of the cervical cancer cases by educational status and oral contraceptive use.*

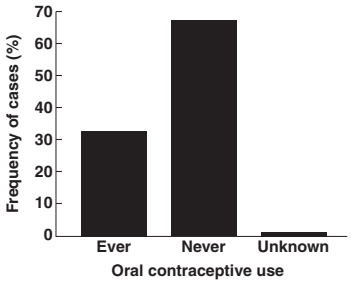
Table 3.2.
Distribution of cervical cancer cases by educational status and oral contraceptive use.[a]

Baseline characteristics	Number of cases	Percentage
Education (schooling)		
Ever	317	72.7
Never	119	27.3
Total	**436**	**100.0**
Oral contraceptive use		
Ever	141	32.4
Never	291	66.7
Unknown	4	0.9
Total	**436**	**100.0**

[a] Data from Bosch *et al.* (1992).

Bar charts are often used to present the distribution of categorical variables (Figure 3.5). In this type of graph, the value of the quantity of interest is represented by the length of the bar.

Figure 3.5.
Distribution of cervical cancer cases by oral contraceptive use (data from Bosch *et al.*, 1992).



Summarizing categorical data

For categorical data, percentages are the only summary measures that can be calculated (as in Table 3.2). They are particularly useful when making comparisons between different groups.

3.3.3 Two variables—quantitative or categorical

So far, we have considered the frequency distribution of a variable within a single group of subjects. Often, we need to compare frequency distributions between two or more groups defined by another variable; for example, the distribution of oral contraceptive use among Colombian and Spanish cases of cervical cancer. Thus, we wish to examine the association between two categorical variables: oral contraceptive use and country of residence. One way to do this is to tabulate the data as in Table 3.3.

Table 3.3.
Distribution of cervical cancer cases by oral contraceptive use (rows) according to country of residence (columns).[a]

Oral contraceptive use	Country of residence (number (%))	
	Colombia	Spain
Ever	77 (41.4)	64 (25.6)
Never	109 (58.6)	182 (72.8)
Unknown	0 (0)	4 (1.6)
Total	**186 (100.0)**	**250 (100.0)**

[a] Data from Bosch *et al.* (1992)

When considering two variables simultaneously, it is useful to classify them according to their purpose in the investigation, as either *explanatory* or *response* variables. Explanatory variables are characteristics of the subjects (exposures) that will be used to explain some of the variability in the response, which is the outcome of interest.

Table 3.3 shows the distribution of the response variable (oral contraceptive use) according to each category of the explanatory variable (country of residence). In this example, it is appropriate to calculate column percentages to show the distribution of the response variable.

These data may also be presented as in Table 3.4. In this case, it is appropriate to calculate row percentages.

The distribution of oral contraceptive use among cervical cancer cases in the two countries can be illustrated in a two-bar chart arranged as in Figure 3.6.

Table 3.4.
Distribution of cervical cancer cases by oral contraceptive use (columns) according to country of residence (rows).[a]

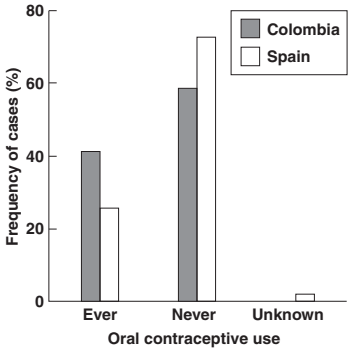
Country of residence	Oral contraceptive use (number (%))			
	Ever	Never	Unknown	Total
Colombia	77 (41.4)	109 (58.6)	0 (0)	**186 (100.0)**
Spain	64 (25.6)	182 (72.8)	4 (1.6)	**250 (100.0)**

[a] Data from Bosch *et al.* (1992)

Quantitative variables (such as age) can also be tabulated by grouping the values of the variable. Table 3.5 shows the age distribution of cervical cancer cases at the time of their enrollment into the study in each country. The distribution of the response variable (age) is shown according to each category of the explanatory variable (country).

Tabulations are useful for showing the association between categorical variables; they are not suitable for illustrating the association between quantitative variables (unless they are grouped into categorical variables as in Table 3.5). This is discussed in more detail in Section 11.2.1; here, only graphical representation of the association between two quantitative variables is considered. As an example, we can plot data on vitamin C intake and levels in the plasma in a sample of 25 individuals on a *scattergram* (Figure 3.7). The values of each variable are represented on an axis. A symbol (often a dot or a cross) is used to represent each individual.

Figure 3.6.
Distribution of cervical cancer cases by oral contraceptive use according to country of residence (data from Bosch *et al.*, 1992).



Age group (years)	Country of residence (number (%))	
	Colombia	Spain
<30	15 (8.1)	7 (2.8)
30–39	48 (25.8)	41 (16.4)
40–44	27 (14.5)	30 (12.0)
45–54	50 (26.9)	61 (24.4)
≥55	46 (24.7)	111 (44.4)
Total	**186 (100.0)**	**250 (100.0)**

[a] Data from Bosch *et al.* (1992).

Table 3.5.
Age distribution of cervical cancer cases according to country of residence.[a]

3.4 Final remarks

Data analyses should always begin by using basic tables, graphical techniques and summary statistics to explore the data. Tables, graphs and summaries should, however, be presented in a sensible way, and must not be misleading.

General rules for designing tables

(1) A table should have a title that gives a clear indication of its contents. The reader should be able to determine without difficulty precisely what is tabulated.
(2) Column and row headings should be brief but self-explanatory.
(3) The source of the data should be included, so that original sources can be checked.
(4) Units of measurement should be clearly indicated.
(5) Totals should be given. These are helpful in indicating the quantity of data in the table and to allow comparison with data presented elsewhere.

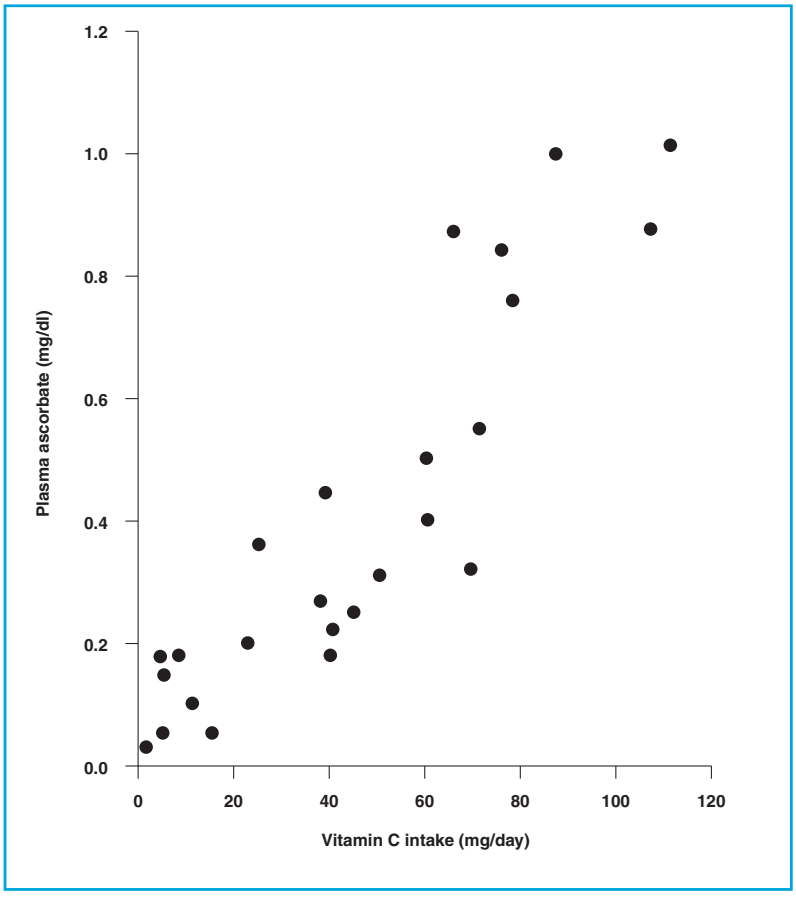


Figure 3.7.
Plasma ascorbate levels in 25 study subjects in relation to their vitamin C intake: hypothetical data.

(6) If percentages are given, the base for the percentage should be clearly indicated. In a complex table displaying percentages without indication of their base, the reader is uncertain how or in which way the percentages total 100%. Such a table can easily be misinterpreted. A glance at the location of the 100% value almost always informs the reader immediately how the percentages in the table were derived.
(7) Approximations and omissions can be explained in footnotes. If some observations are excluded from the table, their number should be indicated clearly.
(8) Tables are used to present results in a more concise and clear way than would be possible in the text. Often, a certain degree of complexity is necessary to save space. However, avoid compressing too much information into a single table. Two or three simple tables may be better than a single large and complex one.

General rules for designing graphs

(1) A graph should have a self-explanatory legend.
(2) A graph should help the reader to understand the data. It should not be cluttered with too much detail.
(3) Axes should be clearly labelled and units of measurement indicated. It is important for the reader to be able to tell precisely what is being illustrated and in which units.
(4) Scales are extremely important. Whenever feasible, they should start at zero; otherwise, this should be clearly indicated by a break in the axis.
(5) Graphs are generally more easily interpreted if the explanatory (exposure) variable is displayed along the horizontal axis and the response (outcome) variable along the vertical axis.
(6) Avoid graphs that give a three-dimensional impression, as they may be misleading (people visualize less easily in three dimensions).
(7) Choose the type of graph that uses as little ink as possible without loss of information. Figure 3.8 presents the same data in two different ways. In graph (*a*) the data are difficult to interpret because of the three-dimensional columns, multiple outcome scales, grid-lines and hatching. Most ink was used to present features that can be

omitted without any loss of information. The same data are shown in a much simpler and clearer manner in graph (*b*).

General rules for reporting summary measures

(1) Always report the number of observations on which the summary is based. For binary responses (A or B), report the percentage of A or B, but not both.

(2) If the median is used as a measure of the central value of a quantitative distribution, give the lower and upper quartiles (or the range) as well.

(3) If the mean is used as a measure of the central value of a quantitative distribution, give the standard deviation as well.

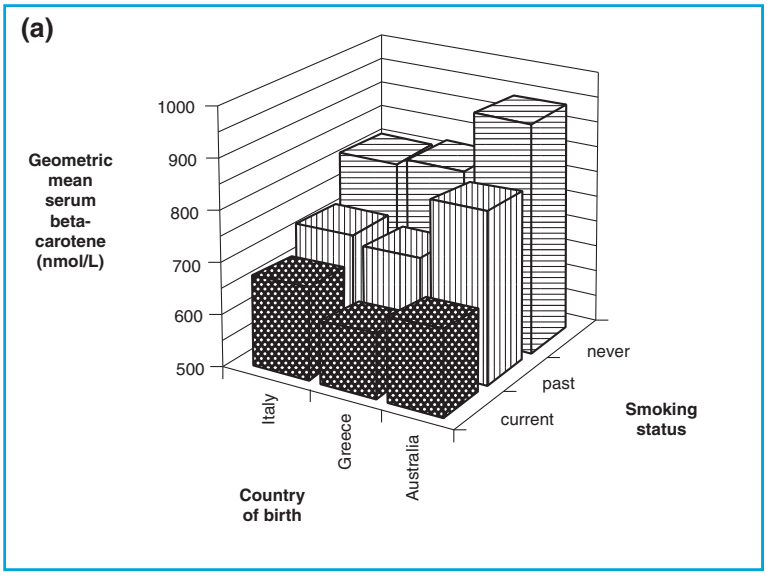


Figure 3.8.
Two graphical presentations of data from 750 subjects on their country of birth, smoking status and serum concentrations of beta-carotene: (*a*) 'business' graph; (*b*) 'scientific' graph (reproduced, with permission, from Jolley, 1993. © by The Lancet Ltd, 1993).

Further reading

* Further details of ways of presenting and summarizing data are given by Altman (1991) and Bland (1987).

* An excellent book on graphical methods is that by Tufte (1983).

* A critical view of the use (and misuse) of graphs in epidemiology is given by Jolley (1993).

Box 3.1. Key issues

- After the data have been edited, they should be examined using simple tabulations, graphs and summary measures.
- The choice of the correct methods to present and summarize the data depends on the type of data collected: *quantitative* or *categorical*.
- The frequency distribution of a *quantitative* variable can be presented in tables, or graphically, in histograms. It can be summarized by reporting a measure of central value and a measure of spread of the distribution (i.e., arithmetic mean with standard deviation, or median with inter-quartile range).
- The frequency distribution of a *categorical* variable can be presented in tables, or graphically, in bar charts. Percentages of individual values in each category are the only summary measures that can be calculated for this type of variable.
- Tables and bar charts can be used to examine the relationship between two categorical variables, and scattergrams to examine the association between two quantitative variables. It is important, however, to first decide which variable is the explanatory variable and which is the response variable.
- Tables, graphs and summary measures should be intelligently designed so as to ensure accurate representation of the data.