CHAPTER 1: INTRODUCTION, BASICS, DESCRIPTIVE STATISTICS

MOTIVATION

In all walks of life, especially health care, we are under constant threat of being overwhelmed by data. But unless we are going to practice on the basis of hunches, we need these data to make good decisions and so optimize our treatment or policy plans. So, our first objective must be to understand a little about data.

Data do not arise by magic; data collection is determined by a human, with all his or her prejudices and proneness to error. We rely on instruments, such as questionnaires, mercury barometers and spectrophotometers, to measure and record variables of interest, such as people’s ages or blood pressures, or amounts of chlorofluorocarbons in the atmosphere. But all instruments have limitations on their accuracy.

Humans tend to accept a body of raw data at face value; a computer print-out of numbers can appear to be an unalterable truth. It may, in fact, be junk. Subsequent statistical processing will not recycle junk into truth (“garbage in, garbage out” principle).

The next problem is: how do we extract meaningful information from masses of raw data? (How do we see the forest, when the trees keep getting in the way?) Descriptive statistics provides systematic and powerful tools to help us out.

The final objective is to interpret the data intelligently. Often, this means interpreting the results of someone else’s analysis of the data. This demands a knowledge of why and how the data were collected, to what statistical manipulations the data were subjected, and perhaps most importantly, a good background knowledge of the field of inquiry.
§ 1.1 INTRODUCTION

This chapter is concerned with the nature of data, how they can be presented and how to summarise information. This is the domain of descriptive statistics. Later on in the course we shall explore the other major arm of statistics: inferential statistics. The latter is concerned with using information from a sample to speculate on what might be true of a population from which the sample was taken.

Statistics like most other fields has a language. If you learn some of the basic vocabulary, things go a lot easier. Like learning any language, it is one thing to know the words and another to get a real feeling for what they mean. I’ll try and give examples frequently – you should stop and think a little on the way. Rote learning of the vocabulary is insufficient.

§ 1.2 SOME BASICS

§ 1.2.1 Data and Variables

A variable is a characteristic of a population which can take different values. The population might be a human population, for example, the aboriginal population as recorded in the 1991 census; variables of interest in this population might be age, income, number of children and education level. One could envisage a population of teaching hospitals and perhaps be interested in such variables as annual budget, number of medical students and number of X-rays taken per year. The Health Commission might be interested in a population of small chemical processing plants and wish to measure variables related to the health of employees – for example, fresh air recirculation times, or number of emergency dousing showers per factory. Whatever the specifics, a population embraces all the elements that we might wish to measure – even if actually doing so might be logistically impossible.

The variables we will deal with in statistics are called random variables because the values they take are generated by chance events. Sometimes the randomness is obvious, as when the data are generated by rolling a die and observing the variable we call “face value” which can take values 1, 2, 3, 4, 5 or 6. But as long as there is a random component to the measurement we can make use of the variable in a statistical sense. So a researcher might be interested in the rise in blood pressure (the variable) in patients who are fed a controlled high salt diet. The blood pressure response in patients can be thought of as the result of two components: a constant non-random response in all patients due to the preselected salt dosage, and a random unpredictable response component due to each individual’s physiological makeup, personality and perhaps scores of other unknown factors.

Data are measurements collected on a variable as a result of taking observations. Often, data will have associated units of measurement, for example, data collected by
observing a patient’s blood pressure will have units of millimetres of mercury (mmHg). Most often, due to time and resource constraints, we are dealing with data collected on a subset or sample of a population. (A counter-example to this is the census carried out by the Australian Bureau of Statistics every five years: the ABS aims to collect information on the entire population of Australia.)

How a sample is selected from the wider population is of crucial importance, as we shall see later.

Data may be classified as being discrete if the variable can take only a finite number of values, for example, number of pregnancies, number of heads appearing on 35 tosses of a coin; or continuous if the variable can (at least within a certain range) take any value along the number line, for example, height, plasma cholesterol level, blood pressure.

§ 1.2.2 Measurement Scales

Depending on the nature of the variable, we have different measurement scales.

Firstly, note that not all data are numbers, though they are often represented as numbers. For example, when recording the sex of a subject, we may choose to code males as “1” and females as “2”, especially when entering the data into a computer. In this example, the numbers 1 and 2 have none of the usual properties of numbers: it doesn’t make sense to add them or say that because females are assigned the number 2 then females possess more sex than males who were assigned the number 1. In fact, we could just as easily have assigned the “numbers” the other way around. The only mathematical relation that makes sense is that of equivalence or non-equivalence. Two observations are equal if both are female (or both male) and non-equal otherwise. Such a scale is called nominal or categorical, and because of this lack of mathematical properties it is called the “weakest” scale. Other examples of categorical variables are “country of birth”, “colour of eyes” and “A-B-O blood type”.

A slightly stronger scale, one that uses the mathematical notion of ordering is the ordinal scale. An example of a variable measured on an ordinal scale might be the position of an examination candidate in the results order of merit. The number of the ranked position tells us nothing about how much difference exists between successive positions on the scale. That is, the highest ranked candidate (he or she would score 1 on our ordinal scale) may have achieved an exam mark of 100%. The person coming second in the order of merit may have scored 95%. There is no way of knowing what percentage result the person ranked third would have achieved. It might be 94% or it might be 20%. All we know is that he or she achieved a score less than the person ranked 2 and greater than the person ranked 4. A clinical example of an ordinal scaled variable is the APACHE II score (Acute Physiology And Chronic Health score) assigned to emergency hospital admissions.

The next scale is the interval scale. It possesses strong mathematical properties due to the fact that in this scale equal differences between points represent equal
differences in the measured quantity. For example, the difference between 12 metres and 11 metres is the same as the difference between 4 metres and 3 metres.

The *ratio* scale is considered a refinement of the interval scale. In this scale, the order and size of interval are important, but the ratio between two measures also has meaning. This occurs when there is a true zero point associated with the scale. For example, temperature in degrees Celsius is an interval measure which has an arbitrary zero point; 0°C Celsius is not a true absence of heat, just a convenient reference point and it does not make sense to say that 30°C is twice as hot as 15°C. In contrast, height has a natural zero point (the ground) and we can say that a 2 metre person is twice as high as a 1 metre person who is twice as high as a 0.5 metre child. Another example of a variable which is measured on a ratio scale is temperature in degrees Kelvin. This has a true zero point, absolute zero (0 K = −273°C), at which point there remains nothing of the quantity we are measuring (kinetic energy of the molecules).

Note that the stronger measures can always be “collapsed” down to form a weaker measure but not vice versa. For example, we may arbitrarily choose a cutoff value of 1.5 metres to separate people into short or tall persons. As long as we have the original height data, recorded on a ratio scale, we can collapse it down to a nominal (tall or short) scale.

This last example also illustrates a general principle of data analysis: when data are simplified or summarised, information is always lost, understanding may be gained. Here, if we measure heights on enough people, we may be overwhelmed by the number of individual measurements, but being able to say that x% are tall and (100 - x)% are short may give a useful insight into this aspect of the population. Of course, using this nominal scale of measuring height we are no longer able to say what an individual’s actual height is.

The reason we stress scales of measure is that the information content of data depends on the scale, and different descriptive techniques and different statistical tests are appropriate to different scales.

§ 1.3 DESCRIPTIVE STATISTICS

There is one important point I would like to make regarding descriptive statistics. There is often a tendency to neglect a thorough examination of the data by descriptive means and go straight on to apply a battery of statistical tests to the data. Nothing could be more ill-advised. You are urged always to take the time to examine a set of data descriptively from a number of different perspectives to get a feel for it. Much sense and nonsense will be discriminated if you follow this advice.

Descriptive statistics includes methods for presenting and summarising data. These allow us to digest and understand large quantities of data, and to effectively communicate to others important aspects of our research.
§ 1.3.1 Frequency Distributions and Data Presentation

If you arrange your raw data so that the scores on a variable of interest are in order of magnitude, that is, you rank the data, and then indicate by means of a table or graph how often a score occurs, then you will have constructed a frequency distribution – a tally of the scores.

Example 1.1

Suppose we wish to present information on the drinking habits of Australian males. Let us say that 1000 males were selected at random from the electoral roll and their drinking habits were ascertained. For convenience we might categorise the variable “drinking habits” into 6 classes of “grams of alcohol consumed per day”: 0 to 9 g/day, 10 to 19 g/day, 20 to 29 g/day, 30 to 39 g/day, 40-59 g/day and 60-99 g/day. The remainder of §1.3.1 and §1.3.2 discuss the presentation of these data.

You might note that the classes chosen in Example 1.1 don’t overlap; the class limits are not shared between two classes as they would be if, for example, we erroneously defined the second and third classes as 10-20 and 20-30 g/day. In this latter case, if a person drank 20 g/day, how would we know to which class he should be assigned? Such problems are avoided by ensuring that classes are mutually exclusive. The limits of the intervals, e.g. 40 and 59 in the fifth interval, are called the observed class limits. Since we are presumably prepared to measure the consumption to the nearest gram, the true class limits are, for this interval, 39.5g to 59.5g, and similarly for the other intervals. You might present the distribution of the observed drinking scores as a table:

<table>
<thead>
<tr>
<th>grams/day</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>125</td>
<td>125/1000 = 0.125</td>
</tr>
<tr>
<td>10-19</td>
<td>250</td>
<td>250/1000 = 0.250</td>
</tr>
<tr>
<td>20-29</td>
<td>400</td>
<td>400/1000 = 0.400</td>
</tr>
<tr>
<td>30-39</td>
<td>150</td>
<td>150/1000 = 0.150</td>
</tr>
<tr>
<td>40-59</td>
<td>50</td>
<td>50/1000 = 0.050</td>
</tr>
<tr>
<td>60-99</td>
<td>25</td>
<td>25/1000 = 0.025</td>
</tr>
</tbody>
</table>

TABLE 1.1 Frequency and relative frequency distributions of alcohol use in 1000 Australian males

For example, 250 men were recorded as consuming at least 10, but not more than 19, grams of alcohol per day. Note that the sum of the frequencies adds to the total number in the sample (that is, 1000). The first two columns are all you need to present the frequency distribution. If you were a little more artistic, you could present the same information by means of a bar graph (see Fig 1.1). The height of each bar is proportional to the frequency of occurrence of its respective score.
§ 1.3.2  **Relative Frequencies**

The third column of Table 1.1 defines a related distribution – the distribution of *relative frequencies*. Note that the fractions add to 1 (the percentages would add to 100%). A *pie graph* might do the same job: each slice of the pie would have an area proportional to the relative frequency of the score it was representing. For example, a consumption score of 20-29 grams per day would account for two-fifths (40%) of the pie. (See Fig 1.2.)

![Bar graph](image)

**Fig 1.1**  Bar graph

![Pie graph](image)

**Fig 1.2**  Pie graph

Beware pie charts that show slices at acute angles to the plane of the page, or “exploded” pie charts (slices cut away from the rest of the pie), or those giving a 3-dimensional effect. Such charts can deceive the eye and lead to misinterpretation. Usually, the simpler the graph, the better.
§ 1.3.3 The Histogram

There is another, much more informative way of presenting relative frequencies. This is the histogram. A histogram looks a bit like a bar graph, but:

- it is used in connection with interval- or ratio-scaled variables, whereas bar graphs can be used with a categorical or an interval/ratio variable (after the latter has been suitably categorised);

- histograms represent relative frequencies (how “dense” are the data in each particular sub-interval); bar graphs are used for frequencies and misused for relative frequencies; and

- histograms give you an idea of the shape of the relative frequency distribution. Bar charts are just tallies and can’t tell you about distribution shapes.

Figure 1.3 is a histogram representing the relative frequencies (and shape) of a distribution of birthweights from a cohort of 489 infants born in an Adelaide teaching hospital during 1988. Note the negative skew of the distribution (long tail on the left) due to several atypical low birthweights.

If you were faced with raw data (and no computer to help!), you might use a “rule-of-thumb” to decide the number of mutually exclusive intervals. One recommendation is that of Sturges who prescribes the formula given by Expression E1.1 (best suited to symmetrical data distributions, with sample size less than 500):

\[ N = 1 + \log_2(n) = 1 + [3.3 \times \log_{10}(n)] \]\n
E1.1
where N is the number of intervals and n is the number of observations. With our n of 489, this rule would yield N = 10 intervals.

The histogram is constructed of N blocks: each block is centred over the midpoint of its interval and starts and finishes at its true lower and upper class limit respectively. (You might note that the computer generated histogram of Fig 1.3 has centred its blocks over the interval midpoints, but has labelled the x-axis in birthweight increments of 500 grams, rather than showing the interval midpoints.) The important thing to know is that, unlike the bar graph, it is the _area_ of each block, not the height, that is proportional to the relative frequency. If, when constructing your histogram, you make some intervals of longer length than others (perhaps to accommodate sparse data at the ends of the distribution), then the base of that block will be longer, but you must ensure that the height is correspondingly decreased so as to keep the _area_ proportional to the relative frequency of data represented by the block.

By the way, this relative frequency distribution is called an _empirical relative frequency distribution_; “empirical” (check out the Greek derivation) means you have gone out and collected the data without resort to any theories about how the data may have been generated. We’ll be looking at _theoretical_ relative frequency distributions (called _probability distributions_) later.

§ 1.3.4  **Other Graphical Methods**

There are scores of other ways of presenting information graphically. Whole texts are devoted to the exposition of statistical graphics and the pitfalls associated with constructing and interpreting the various types. Each type of graph is best suited for a particular task. For example, we have seen that the _histogram_ is a good way of showing the relative frequencies of an interval-scaled variable. A _scatterplot_ (we’ll meet it later) shows the distribution of two interval-scaled variables simultaneously. The curiously named _box and whisker plot_ (see Fig 1.5) is used to show the distribution of either ordinal or interval-scaled variables. A _stem and leaf_ plot is an alternative to the histogram but packs in even more information for the skilled observer. The list seems endless; statistical graphics is an ongoing area of research and development by statisticians and computer programmers.
§ 1.4 SUMMARY MEASURES

It is time to turn our attention to measures of central tendency and measures of dispersion (or variability). Such summary measures, calculated from a sample of data, are called sample statistics, or just statistics. (Do not confuse this term with the subject ‘Statistics’. A sample statistic is just a number.) Sample statistics provide a shorthand way of describing a data distribution.

§ 1.4.1 Measures of Central Tendency

If we are confronted with a set of data, we may ask: is there one value which is typical of the set of data as a whole? (An abused, ambiguous and to-be-avoided term is the “average”.) This value might convey a useful summary of the data at the unavoidable expense of the loss of information inherent in the unsummarised raw data set. We will consider three measures of central tendency: the mean, the median and the mode.

§ 1.4.1.1 The Mean

The commonest measure of central tendency is the arithmetic mean. The formula is:

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

Expression E1.2 directs us to add up the raw data (\(\Sigma\) is the summation operator) and divide by the number of data points. Note that the symbol for the sample mean is a modified Roman letter, \(\bar{x}\). Most sample statistics, such as the mean, are represented mathematically as Roman letters. The arithmetic mean of the four blood pressure observations: 120 mmHg, 130 mmHg, 150 mmHg and 170 mmHg is 142.5 mmHg. Notice that the mean has the same units as the raw data.

§ 1.4.1.2 The Mean: Uses and Abuses

The chief advantage of the arithmetic mean is that it is part of an entire mathematical system in statistics and crops up in probability theory, many standard data distributions and throughout inferential statistics. Another extremely important property of the mean is that it is a very stable or reliable measure of central tendency. If we were to take repeated samples of say, 100 persons from the general population, and ascertain the mean blood sugar in each group of 100, we would find that the means would vary much less from sample to sample than the other measures of central tendency discussed below.

[As we shall see, the sample mean is often used to estimate the population mean. This is an example of a sample statistic being used to estimate a population parameter. For example, The National Heart Foundation, in its third Risk Factor]
Prevalence Study (1989), chose a sample of 3000 adults in the Adelaide metropolitan area, calculated the mean serum cholesterol level of the 3000 people, and used this sample mean as an estimate of the mean serum cholesterol in the entire adult population of Adelaide.]

However, the mean may mislead when the raw data contain extreme values (leading to skew, that is, asymmetric, frequency distributions) since it is very sensitive to such extremes.

Example 1.2
In a study of employees’ respiratory illness, it would be unwise to use the mean as the sole measure of central tendency of the number of workdays missed due to colds. This is because the relatively few people who took many days off would cause the mean to rise disproportionately and not reflect the main sense of the information we wish to convey. A little data set pertaining to this example will be used later in this section.

§ 1.4.1.3 The Median
In the case of Example 1.2, it might be better to use a summary measure of central tendency which is less sensitive to extreme values. Such a measure is the median. 50% of the data are less than the median and 50% are greater. In other words, the median represents the 50th percentile of the distribution. The median is not as amenable to mathematical manipulation as the mean, although certain statistical techniques utilise it.

To illustrate the above, let us say we surveyed 5 workers and asked them how many days-off-work they took during the preceding 12 months. The results were:

\[
2 \quad 4 \quad 1 \quad 3 \quad 2
\]

To make life easier, we will rank these results:

\[
1 \quad 2 \quad 2 \quad 3 \quad 4
\]

The mean is 2.4 days and the median (the middle score) is 2 days. But now we will add one more worker’s results to the raw data. This worker was atypical: he took 30 days off – this new extreme datum will affect the shape of the distribution of sick days: it will be skewed to the right (a positive skew). The new data set is:

\[
1 \quad 2 \quad 2 \quad 3 \quad 4 \quad 30
\]

The median is now 2.5 days. (By convention, if there is an even number of observations we take the mean of the n/2th and [(n/2) + 1]th scores. Here, n = 6, so we get the mean of the third and fourth scores). This is not much different from the original median calculated without the extreme value (2 days). The median appears to be robust (relatively unaffected by atypical data points). However, the mean has
been dragged upwards to 7 days, which is well away from where the majority of the data lie.

Example 1.3
The Commonwealth Statistician reports the median, rather than the mean, household income, since the earnings of the few very wealthy Australians would lead the mean to distort the general picture. Similarly, each month the local newspaper lists the median house prices in each suburb: the mean would be too sensitive to the occasional mansion that was sold.

§ 1.4.1.4 The Mode

The mode is the value of the data which occurs most frequently. A set of data may have no, one, two or more modes. The mode of the little data set of Example 1.2 is 2, since it occurs more often (twice) than any other raw datum. Since there is only one mode, the distribution is termed unimodal. Example 1.4 shows a bimodal frequency distribution of scores on a Biostatistics exam.

Example 1.4 bimodal distribution:

<table>
<thead>
<tr>
<th>score</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>13</td>
</tr>
<tr>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>32</td>
<td>19</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>28</td>
</tr>
<tr>
<td>60</td>
<td>11</td>
</tr>
</tbody>
</table>

The two modes are the scores 29 and 50, since they occur equally most often. When graphed, bimodal frequency distributions look like two-humped camels.

§ 1.4.2 Variability of Distributions

It seems obvious that a summary measure of central tendency, such as the mean, inadequately describes a set of data. We would also like to summarise the variability or dispersion of the data.

Indeed, two distributions may have identical means but the spread of values around the mean may differ considerably. Fig 1.4 shows three relative frequency distributions. For ease of drawing, they have been smoothed to give nice continuous curves. (In fact, each distribution is an example of a theoretical Normal probability curve – see Chapter 2.) We would have much more confidence in using the mean as a summary descriptor of the distribution with the least spread (distribution #1) than the one whose values don’t cluster as tightly to the mean (distribution #3). Distribution
#2 has a different mean but the same variability as distribution #1. This situation is often termed a “shift of location”.

![Diagram showing variability of distributions](image)

**Fig 1.4** Variability of distributions

§ 1.4.2.1 **Measures of Variability**

Once again, we have a choice of measures of variability. The simplest is the range (highest value minus lowest value). The interquartile range is the difference between the value cutting off the upper 25% of scores (the value at the 75th percentile) and the value cutting off the lower 25% of scores (25th percentile). As you might expect, the interquartile range as a measure of variability is often used with the median (the 50th percentile) as a measure of central tendency to give an idea of the distribution.

A graphical representation of this is the box and whisker plot. The ends of the box (called the hinges) are the 25th and 75th percentiles, the median is the line within the box and the whiskers stretch out, below and above, to observed data no more than 1.5 times the interquartile range. Data outside the ends of the whiskers (called fences) are often termed outliers, and may be represented by asterisks or by open circles. Fig 1.5 shows the distribution of birthweights of babies, overall and by sex, in the Adelaide birth cohort. Note that the distributions are roughly similar between the sexes, and that the distributions are reasonably symmetrical. There is one outlier in particular that commands our attention: one male baby’s birthweight was only 509 grams. We would investigate this further, firstly by checking that the recorded measurement was valid.
However, the most commonly used measure of variability is the variance or its square root, the standard deviation. The variance is a kind of “average” of the squared differences of each of the n observations from their mean. The formulae are simple:

\[
\text{Sample Variance} = S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \quad \text{E1.3}
\]

The advantages of the variance and standard deviation are that they are easy to manipulate and, like the mean, they are part of a general statistical system.

A useful trick is that you can roughly estimate the standard deviation by dividing the range by 4. This only works (i) for symmetric data distributions; (ii) when the number of data points is reasonably large; and, in any case, it is only very approximate.

You notice that, in E1.3, to get the average of the n squared deviations, we divide the sum of the squared deviations, \( \sum_{i=1}^{n} (x_i - \bar{x})^2 \), by (n – 1) and not by n. This probably seems a bit unreasonable, but it can be shown that dividing by (n – 1) gives a sample standard deviation which is a better estimate of the population standard deviation than if n is the divisor. “Better”, in this context, means unbiased, that is, the sample statistic is not systematically different from the population parameter. This is an important consideration when we use inferential statistics.
§ 1.4.2.2 Degrees of Freedom

The value of \((n – 1)\) is called the number of degrees of freedom associated with our sample estimate of the population variance. To get the variance, we need to calculate the \(n\) deviations around the mean. But there is one implicit restriction imposed here: the sum of these deviations around the mean must sum to zero (try it on a small data set and see!; or prove it using elementary algebra). Put another way, we have \(n\) observations but, if we are to obey the constraint, we can only choose \((n – 1)\) freely; the value of the last one is predetermined by the mean. So, we wind up with \((n – 1)\) “degrees of freedom”.

Degrees of freedom pop up all over statistics – we will meet them again in this course – but usually there are some simple rules to enable their easy calculation in a particular context.

Example 1.2 (continued)

If you calculate the variance and standard deviation of our sample of sick days:

\[
1 \quad 2 \quad 2 \quad 3 \quad 4 \quad 30
\]

you should find that the variance is 128 days\(^2\) and the standard deviation is 11.31 days.

Note that the units of the standard deviation are the same as that of the raw data, that is, days, whilst the variance has those units squared.

If we think of this sample variance as being an estimate of the variance of sick days in the population of workers, then the number of degrees of freedom associated with this variance estimate is \((n – 1) = 5\).

§ 1.4.2.3 A Nice Property of Symmetric Distributions

For unimodal, symmetric distributions, roughly two-thirds of the data (or \(2/3\) of the area of a relative frequency distribution) is enclosed within \(±1\) standard deviation from the mean, and \(95\%\) of the data (\(95\%\) of the area) is enclosed within \(±2\) standard deviations from the mean. You should remember this. In many health-related situations this property allows you to make a judgement on how unusual an individual might be within the defined population. Fig 1.6 shows these proportions in an idealised, smoothed distribution.
Example 1.5a

In a large scale survey investigating the prevalence of risk factors for cardiovascular disease, systolic blood pressure was measured twice on each subject, so that the reliability of measures could be ascertained. The difference between the two measures was calculated for each of the 10,000 subjects and the relative frequency distribution of these within-subject differences was plotted (see Fig 1.6). The mean difference was 0 mmHg (demonstrating an absence of bias in the repeat measure), the standard deviation was 2 mmHg. Since the distribution is clearly symmetrical, we estimate that 95% of differences lies within 4 mmHg of the mean, 0. If, say, one subject’s repeat measures differed by more than 4 mmHg, then we know that (i) this is possible, but (ii) it is unusual, and maybe we would investigate the circumstances of this subject more closely.

Example 1.5b

Birthweight is approximately symmetrically distributed in a population. If the mean weight of newborns in a population is 3500 grams, and the standard deviation is 750 grams, what proportion of babies will weigh less than 2000 grams? In the absence of other information, we might estimate that 0.025 or 2.5% of newborns would weigh less than 2000 grams, as this is 2 standard deviations below the mean.
§ 1.5 Reporting Measurements

If we are in the business of collecting or interpreting data, and drawing conclusions which may affect the well-being of others, then we must be concerned with how well we measure characteristics of interest.

There are many ways to assess this; what follows is intended to give a general background to the particular techniques you may come across.

§ 1.5.1 A Cardinal Principle

It seems to be common practice to quote:

- *measurements* as though they were *error-free*;
- summary measures of *central tendency* without accompanying measures of *variability*;

and, as we shall explore later,

- *sample estimates* of population parameters as though they were not subject to *variability*.

However, this is an inadequate approach in the physical, medical or social sciences. Our cardinal principle is:

*Since one (usually) cannot:*

- measure without error;
- summarise without loss of information; or
- estimate without imprecision,

*a statement of measurement should be accompanied by some statement of error, and a summary measure or an estimate by some indication of its variability.*

**Example 1.6**

A single measurement of a continuous variable, for example an individual’s blood pressure, will often be accompanied by an error dependent on the sensitivity of the measurement scale. If a mercury sphygmomanometer is graduated in 2 mmHg intervals, a reported systolic blood pressure of 120 mmHg may be inferred to be not less than 119 mmHg and not greater than 121 mmHg; the measurement should be quoted as 120 ± 1 mmHg. The error is half the width of the interval of the scale.
**Example 1.7**

In the report of a survey of unemployed persons in a low socioeconomic area, the number of consecutive months of unemployment was summarised as 11 ± 3 months (mean ± standard deviation), giving readers of the report some idea of the average time and the spread of measures.

We will explore statements of variability concerning sample estimates of population parameters later in the course.

§ 1.5.2 **Accuracy, Precision and Bias**

An *accurate* measurement is one which is close to the true (probably unknown) value. A measurement is *precise* if repeated measurements yield close to the same value. *Bias* arises from any *systematic* error in measuring (or collecting, analysing and interpreting) data. An accurate measure is a precise, unbiased measure. Example 1.8 illustrates these concepts.

**Example 1.8**

Let us consider measures of blood sugar (glucose) in a patient with diabetes mellitus. This is a disease in which the pancreas produces insufficient amounts of the hormone insulin to cope with the blood sugar load. One method of treatment is to inject small amounts of exogenous insulin at set times during the day; the dosage may depend on the blood sugar at the time.

The patient will take a drop of blood, smear it on a glucose-sensitive reagent strip which changes colour according to the blood glucose level, and insert the strip in a glucometer which measures the blood glucose by means of light reflection off the coloured strip. A high reading, indicating a high blood glucose, will lead to a larger insulin dose being injected. Glucometers are calibrated against glucose solutions of known concentration, supplied by the manufacturer.

On this occasion, the patient’s *true* blood glucose is 12 mmol/L and a perfectly precise, totally unbiased glucometer would accurately register this.

Now consider what might happen if the glucometer becomes relatively insensitive; say the internal light collecting apparatus becomes dirty. Measurements will become *imprecise*. On this occasion, the glucometer might read 12 mmol/L just by chance, but on repeated measures, the values will fluctuate, so that the patient can have no confidence in the test (see Fig 1.7a).

Even if there were no bias in the system, so that the fluctuation was distributed symmetrically around the true value, the situation is pretty unsatisfactory.
Example 1.8 (continued)

Now let’s suppose that the problem with the imprecision is cleared up, but now the
 glucometer is calibrated against a glucose solution of 7 mmol/L which is incorrectly
 labelled as being 5 mmol/L. No matter how precise the apparatus is now, the
 patient’s reading will always be biased by 2 mmol/L, since a systematic error has
 been introduced (see Fig 1.7b).

Fig 1.7a Imprecise, but unbiased measures

Fig 1.7b Precise, but biased measures

One can readily think of other examples where these concepts of accuracy come into
 play. Here is one from the field of epidemiology – the study of the distribution,
 causes, and effects of disease in populations.

Example 1.9

The National Heart Foundation wishes to estimate the mean blood pressure in the
 Australian adult population. The aim of the survey is to give the Foundation the
 information it needs to set national guidelines for the control of hypertension, a major
 risk factor for cardiac, cerebro-vascular and renal disease. The study design chosen
 is a cross-sectional survey – it gives a “snap-shot” picture of people’s blood pressure
 at a single point in time. A random sample of adults is chosen from the Australian
 Commonwealth Electoral Roll and the blood pressure of each selected person is
 measured. (A random sample is one where each person in the population has a
 known – usually equal – chance of being selected.) The mean blood pressure of the
 sample then gives an estimate of the mean blood pressure in the population. But we
 know that certain groups are relatively under-represented on the Electoral Roll, for
 example, new immigrants, Aboriginals and the unemployed. So, these people will
 also be under-represented in the sample. If the blood pressures of these groups differ
 from that of the general population, the sample mean blood pressure will be a biased
 estimate of the true mean of the total population.
§ 1.6 SUMMARY

This chapter has introduced several important concepts. They are:

- Descriptive Statistics is useful for summarising information and communicating it effectively;

- Since information is lost in the process of summarising, great care must be taken in the preparation and interpretation of descriptive statistics;

- Biological systems are inherently variable; also, measuring characteristics in biological systems necessarily introduces error. These errors and variability must be quantified and presented.