CHAPTER 4

Means, standard deviations and standard errors

4.1 INTRODUCTION

A frequency distribution (see Section 3.2) gives a general picture of the distribution of a variable. It is often convenient, however, to summarize a numerical variable still further by giving just two measurements, one indicating the average value and the other the spread of the values.

4.2 MEAN, MEDIAN AND MODE

The average value is usually represented by the arithmetic mean, customarily just called the **mean**. This is simply the sum of the values divided by the number of values.

$$\text{Mean,}\ \bar{x} = \frac{\Sigma x}{n}$$

where x denotes the values of the variable, Σ (the Greek capital letter sigma) means 'the sum of' and n is the number of observations. The mean is denoted by $\bar{x}$ (spoken 'x bar').

Other measures of the average value are the **median** and the **mode**. The median was defined in Section 3.3 as the value that divides the distribution in half. If the observations are arranged in increasing order, the median is the middle observation.

$$\text{Median} = \frac{(n+1)}{2}\text{th value of ordered observations}$$

If there is an even number of observations, there is no middle one and the average of the two ‘middle’ ones is taken. The **mode** is the value which occurs most often.

Example 4.1

The following are the plasma volumes of eight healthy adult males:

2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12 litres

(a) $n = 8$

$\Sigma x = 2.75 + 2.86 + 3.37 + 2.76 + 2.62 + 3.49 + 3.05 + 3.12 = 24.02$ litres

Mean, $\bar{x} = \Sigma x/n = 24.02/8 = 3.00$ litres

(b) Rearranging the measurements in increasing order gives:

2.62, 2.75, 2.76, 2.86, 3.05, 3.12, 3.37, 3.49 litres

Median $= (n+1)/2 = 9/2 = 4½$th value

$=$ average of 4th and 5th values

$= (2.86 + 3.05)/2 = 2.96$ litres

(c) There is no estimate of the mode, since all the values are different.

The mean is usually the preferred measure since it takes into account each individual observation and is most amenable to statistical analysis. The median is a useful descriptive measure if there are one or two extremely high or low values, which would make the mean unrepresentative of the majority of the data. The mode is seldom used. If the sample is small, either it may not be possible to estimate the mode (as in Example 4.1c), or the estimate obtained may be misleading. The mean, median and mode are, *on average*, equal when the distribution is symmetrical and unimodal. When the distribution is positively skewed, a **geometric mean** may be more appropriate than the arithmetic mean. This is discussed in Chapter 13.

4.3 MEASURES OF VARIATION

Range and interquartile range

Two measures of the amount of variation in a data set, the range and the interquartile range, were introduced in Section 3.3. The **range** is the simplest measure, and is the difference between the largest and smallest values. Its disadvantage is that it is based on only two of the observations and gives no idea of how the other observations are arranged between these two. Also, it tends to be larger, the larger the size of the sample. The **interquartile range** indicates the spread of the middle 50% of the distribution, and together with the median is a useful adjunct to the range. It is less sensitive to the size of the sample, providing that this is not too small; the lower and upper quartiles tend to be more stable than the extreme values that determine the range. These two ranges form the basis of the **box and whiskers plot**, described in Sections 3.3 and 3.4.

Range = highest value – lowest value

Interquartile range = upper quartile – lower quartile

Variance

For most statistical analyses the preferred measure of variation is the **variance** (or the **standard deviation**, which is derived from the variance, see below). This uses all the observations, and is defined in terms of the *deviations* (x–$\bar{x}$) of the observations from the mean, since the variation is small if the observations are bunched closely about their mean, and large if they are scattered over considerable distances. It is not possible simply to average the deviations, as this average will always be zero; the positive deviations corresponding to values above the mean will balance out the negative deviations from values below the mean. An obvious way of overcoming this difficulty would be simply to average the sizes of the deviations, ignoring their sign. However, this measure is not mathematically very tractable, and so instead we average the *squares* of the deviations, since the square of a number is always positive.

$$\text{Variance, } s^2 = \frac{\Sigma(x - \bar{x})^2}{(n - 1)}$$

Degrees of freedom

Note that the sum of squared deviations is divided by $(n - 1)$ rather than n, because it can be shown mathematically that this gives a better estimate of the variance of the underlying population. The denominator $(n - 1)$ is called the number of **degrees of freedom** of the variance. This number is $(n - 1)$ rather than n, since only $(n - 1)$ of the deviations $(x - \bar{x})$ are independent from each other. The last one can always be calculated from the others because all n of them must add up to zero.

Standard deviation

A disadvantage of the variance is that it is measured in the square of the units used for the observations. For example, if the observations are weights in grams, the variance is in grams squared. For many purposes it is more convenient to express the variation in the original units by taking the *square root* of the variance. This is called the **standard deviation** (s.d.).

$$\text{s.d., } s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{(n - 1)}}$$

or equivalently

$$s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{(n - 1)}}$$

When using a calculator, the second formula is more convenient for calculation, since the mean does not have to be calculated first and then subtracted from each of the observations. The equivalence of the two formulae is demonstrated in Example 4.2. (Note: Many calculators have built-in functions for the mean and standard deviation. The keys are commonly labelled $\bar{x}$ and σ_{n-1}, respectively, where σ is the lower case Greek letter sigma.)

Example 4.2

Table 4.1 shows the steps for the calculation of the standard deviation of the eight plasma volume measurements of Example 4.1.

$$\Sigma x^2 - (\Sigma x)^2/n = 72.7980 - (24.02)^2/8 = 0.6780$$

gives the same answer as $\Sigma(x - \bar{x})^2$, and

$$s = \surd(0.6780/7) = 0.31 \text{ litres}$$

Table 4.1 Calculation of the standard deviation of the plasma volumes (in litres) of eight healthy adult males (same data as in Example 4.1). Mean, $\bar{x} = 3.00$ litres.

	Plasma volume x	Deviation from the mean $x - \bar{x}$	Squared deviation $(x - \bar{x})^2$	Squared observation x^2
	2.75	−0.25	0.0625	7.5625
	2.86	−0.14	0.0196	8.1796
	3.37	0.37	0.1369	11.3569
	2.76	−0.24	0.0576	7.6176
	2.62	−0.38	0.1444	6.8644
	3.49	0.49	0.2401	12.1801
	3.05	0.05	0.0025	9.3025
	3.12	0.12	0.0144	9.7344
Totals	24.02	0.00	0.6780	72.7980

Interpretation of the standard deviation

Usually about 70% of the observations lie within one standard deviation of their mean, and about 95% lie within two standard deviations. These figures are based on a theoretical frequency distribution, called the normal distribution, which is described in Chapter 5. They may be used to derive reference ranges for the distribution of values in the population (see Chapter 5).

Change of units

Adding or subtracting a constant from the observations alters the mean by the same amount but leaves the standard deviation unaffected. Multiplying or dividing by a constant changes both the mean and the standard deviation in the same way.

For example, suppose a set of temperatures is converted from Fahrenheit to centigrade. This is done by subtracting 32, multiplying by 5, and dividing by 9. The new mean may be calculated from the old one in exactly the same way, that is by subtracting 32, multiplying by 5, and dividing by 9. The new standard deviation, however, is simply the old one multiplied by 5 and divided by 9, since the subtraction does not affect it.

Coefficient of variation

$$\mathrm{cv} = \frac{s}{\bar{x}} \times 100\%$$

The **coefficient of variation** expresses the standard deviation as a percentage of the sample mean. This is useful when interest is in the size of the variation relative to the size of the observation, and it has the advantage that the coefficient of variation is independent of the units of observation. For example, the value of the standard deviation of a set of weights will be different depending on whether they are measured in kilograms or pounds. The coefficient of variation, however, will be the same in both cases as it does not depend on the unit of measurement.

4.4 CALCULATING THE MEAN AND STANDARD DEVIATION FROM A FREQUENCY DISTRIBUTION

Table 4.2 shows the distribution of the number of previous pregnancies of a group of women attending an antenatal clinic. Eighteen of the 100 women had no previous pregnancies, 27 had one, 31 had two, 19 had three, and five had four previous pregnancies. As, for example, adding 2 thirty-one times is

Table 4.2 Distribution of the number of previous pregnancies of a group of women aged 30–34 attending an antenatal clinic.

	No. of previous pregnancies					
	0	1	2	3	4	Total
No. of women	18	27	31	19	5	100

equivalent to adding the product (2 × 31), the total number of previous pregnancies is calculated by:

$$\begin{aligned}\Sigma x &= (0 \times 18) + (1 \times 27) + (2 \times 31) + (3 \times 19) + (4 \times 5)\\ &= 0 + 27 + 62 + 57 + 20 = 166\end{aligned}$$

The average number of previous pregnancies is, therefore:

$$\bar{x} = 166/100 = 1.66$$

In the same way:

$$\begin{aligned}\Sigma x^2 &= (0^2 \times 18) + (1^2 \times 27) + (2^2 \times 31) + (3^2 \times 19) + (4^2 \times 5)\\ &= 0 + 27 + 124 + 171 + 80 = 402\end{aligned}$$

The standard deviation is, therefore:

$$s = \sqrt{\frac{(402 - 166^2/100)}{99}} = \sqrt{\frac{126.44}{99}} = 1.13$$

If a variable has been grouped when constructing a frequency distribution, its mean and standard deviation should be calculated using the original values, not the frequency distribution. There are occasions, however, when only the frequency distribution is available. In such a case, approximate values for the mean and standard deviation can be calculated by using the values of the mid-points of the groups and proceeding as above.

4.5 SAMPLING VARIATION AND STANDARD ERROR

As discussed in Chapter 2, the sample is of interest not in its own right, but for what it tells the investigator about the population which it represents. The sample mean, $\bar{x}$, and standard deviation, s, are used to estimate the mean and standard deviation of the population, denoted by the Greek letters μ (mu) and σ (sigma) respectively.

The sample mean is unlikely to be exactly equal to the population mean. A different sample would give a different estimate, the difference being due to

sampling variation. Imagine collecting many independent samples of the same size from the same population, and calculating the sample mean of each of them. A frequency distribution of these means (called the **sampling distribution**) could then be formed. It can be shown that:

1 the mean of this frequency distribution would be the population mean, and

2 the standard deviation would equal $\sigma/\sqrt{n}$. This is called the **standard error of the sample mean**, and it measures how precisely the population mean is estimated by the sample mean. The size of the standard error depends both on how much variation there is in the population and on the size of the sample. The larger the sample size n, the smaller is the standard error.

We seldom know the population standard deviation, σ, however, and so we use the sample standard deviation, s, in its place to estimate the standard error.

$$\text{s.e.} = \frac{s}{\sqrt{n}}$$

Example 4.3

The mean of the eight plasma volumes shown in Table 4.1 is 3.00 litres (Example 4.1) and the standard deviation is 0.31 litres (Example 4.2). The standard error of the mean is therefore estimated as:

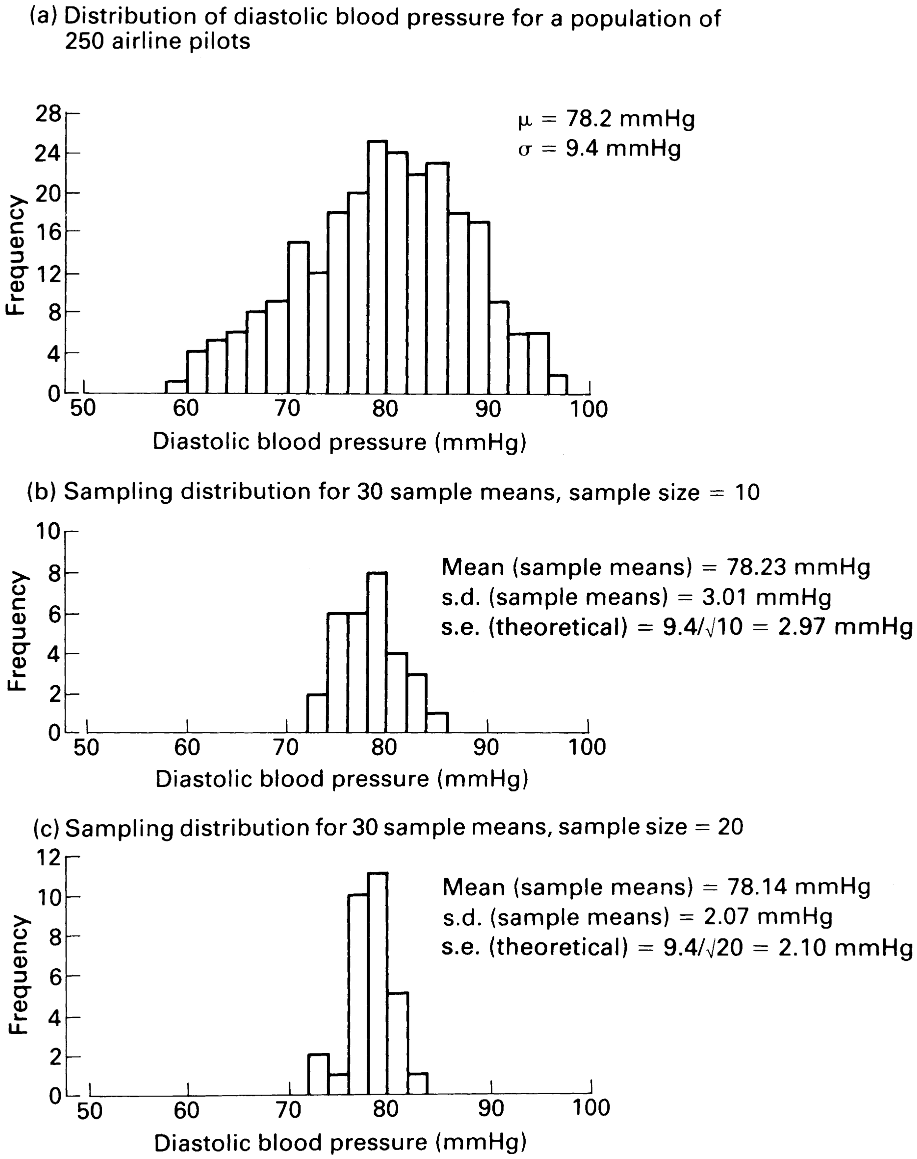
$$s/\sqrt{n} = 0.31/\sqrt{8} = 0.11 \text{ litres}$$

Understanding standard deviations and standard errors

Example 4.4

Figure 4.1 shows the results of a game played with a class of 30 students to illustrate the concepts of sampling variation, the sampling distribution, and standard error. Blood pressure measurements for 250 airline pilots were used, and served as the population in the game. The distribution of these measurements is shown in Figure 4.1(a). The population mean, μ, was 78.2 mmHg, and the population standard deviation, σ, was 9.4 mmHg. Each value was written on a small disc and the 250 discs put into a bag.

Each student was asked to shake the bag, select ten discs, write down the ten diastolic blood pressures, work out their mean, $\bar{x}$, and return the discs to the bag. In this way 30 different samples were obtained, with 30 different sample means, each estimating the same population mean. The mean of these sample means was 78.23 mmHg, close to the population mean. Their distribution is shown in Figure 4.1(b). The standard deviation of the sample means was 3.01 mmHg, which agreed well with the theoretical value, $\sigma/\sqrt{n} = 9.4/\sqrt{10} = 2.97$ mmHg, for the standard error of the mean of a sample of size ten.



(a) Distribution of diastolic blood pressure for a population of 250 airline pilots

μ = 78.2 mmHg
σ = 9.4 mmHg

(b) Sampling distribution for 30 sample means, sample size = 10

Mean (sample means) = 78.23 mmHg
s.d. (sample means) = 3.01 mmHg
s.e. (theoretical) = 9.4/√10 = 2.97 mmHg

(c) Sampling distribution for 30 sample means, sample size = 20

Mean (sample means) = 78.14 mmHg
s.d. (sample means) = 2.07 mmHg
s.e. (theoretical) = 9.4/√20 = 2.10 mmHg

Fig. 4.1 Results of a game played to illustrate the concepts of sampling variation, the sampling distribution, and the standard error.

The exercise was repeated taking samples of size 20. The results are shown in Figure 4.1(c). The reduced variation in the sample means resulting from increasing the sample size from 10 to 20 can be clearly seen. The mean of the sample means was 78.14 mmHg, again close to the population mean. The standard deviation was 2.07 mmHg, again in good agreement with the theoretical value, $9.4/\sqrt{20} = 2.10$ mmHg, for the standard error of the mean of a sample of size 20.

In this game, we had the luxury of results from several different samples, and could draw the sampling distribution. Usually we are not in this position: we have just one sample that we wish to use to estimate the mean of a larger population, which it represents. We can draw the frequency distribution of the values in our sample (see, for example, Figure 3.3 of the histogram of haemoglobin levels of 70 women). Providing the sample size is not too small, this frequency distribution will be similar in appearance to the frequency distribution of the underlying population, with a similar spread of values. In particular, the sample standard deviation will be a fairly accurate estimate of the population standard deviation. As stated in Section 4.2, approximately, 95% of the sample values will lie within two standard deviations of the sample mean. Similarly, approximately 95% of all the values in the population will lie within this same amount of the population mean.

The sample mean will not be exactly equal to the population mean. The theoretical distribution called the **sampling distribution** gives us the spread of values we would get if we took a large number of additional samples; this spread depends on the amount of variation in the underlying population and on our sample size. The standard deviation of the sampling distribution is called the **standard error** and is equal to the standard deviation of the population, divided by the square root of n. This means that approximately 95% of the values in this theoretical sampling distribution of sample means lie within two standard errors of the population mean. This fact can be used to construct a range of likely values for the (unknown) population mean, based on the observed sample mean and its standard error. Such a range is called a **confidence interval**. Its method of construction is not described until Chapter 6 since it depends on using the normal distribution, described in Chapter 5. In summary:

- The standard deviation measures the amount of variability in the population.
- The standard error ($=$ standard deviation $/\sqrt{n}$) measures the amount of variability in the sample mean; it indicates how closely the population mean is likely to be estimated by the sample mean.
- Because standard deviations and standard errors are often confused it is very important that they are clearly labelled when presented in tables of results.