

TOPIC 6

MEASURES OF VARIATION

If people's eyes tend to blank out tables of figures, you can be darn sure that they blank out the small writing that goes around them.

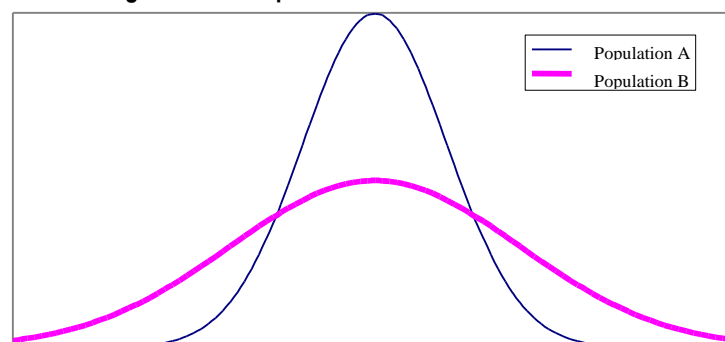
Alan Graham, 1994

The concept of variation

sometimes averages aren't enough

A measure of the average value can provide a lot of useful information about a set of observations, but in many cases it is not sufficient to tell us everything about the variable. Consider, for example, Figure 6.1 below:

Figure 6.1 Comparison of Two Distributions



the distributions are different yet give the same averages

While the two distributions shown have the same average values, whether measured as a mean, a median, or a mode, we could not say that the distributions were the same. To describe and compare them we need additional information; we need alternative ways of describing the distributions. After the average value, the next most important property of the distribution that we need to measure is the variability of the distribution. From Figure 6.1 we can see that distribution 'B' is much more variable (or spread out) than distribution 'A'. In this section we shall look at different ways of measuring variability.

actual level of variability

We want to measure variability for two main reasons. Firstly we may be interested in the actual level of variability and in comparing this with another distribution. If we are looking at income distributions, for example, then the government may be interested not only in the average income level, but also in the variability of income level between people and also between different regions of a country. Many policies are designed to help redistribute income from the richest to the poorest (thereby reducing the

variability of income levels), and so we would need to measure variability to see if it changes over time.

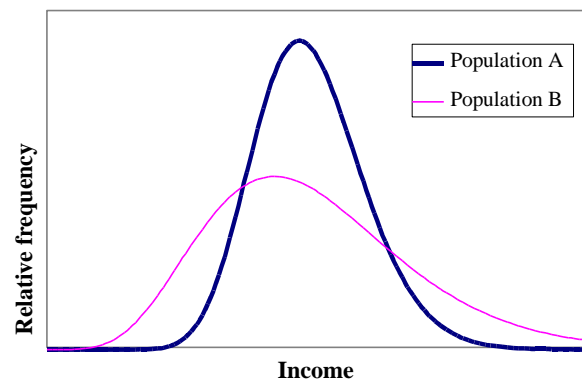
variability due to sample variation

The second reason for wanting to measure variability is when we use sampling to compare population values. We then need to take variability into account. We want to be able to distinguish between differences that might have just happened by chance (that is, in the selection of the samples) and those that indicate some real change.

example

Let us look at an example where we are comparing two population distributions.

Figure 6.2 Comparison of Income Levels of Two Populations



variability is not necessarily reflected in averages

Population 'A' represents the distribution of annual income per household in one region and Population 'B' represents the distribution of annual income of households in another region. Both have the same mean level of income of \$1,800 per year, but we cannot say that the two distributions are the same. The distribution of income of Population 'B' is far more spread out than Population 'A'. It also has, therefore, a greater degree of variability.

two different measures of variability

It is clear that we should not only compare measures of location when looking at populations, but also measures of variability. In this topic we shall consider two different measures of variability which are basically of two types:

- a. measures of the distance between representative values of the population; and
- b. measures of the distance of every unit of the population from some specified central value.

range and standard deviation for ungrouped data

As examples of these measures of variability we shall look in this topic at the range and the standard deviation (or variance) for ungrouped data. More complicated techniques (such as finding the standard deviation when the observations are grouped in a frequency distribution) are covered in advanced training.

The range

largest – smallest

The simplest way to measure the variation or spread, given a set of observations, is to calculate the **range**. The range of a set of observations is defined to be the difference between the smallest and the largest values in the set. This is very simple to understand and easy to calculate and so has an obvious appeal. It is used in practice, but is only really useful when the variable under consideration has a fairly even spread of values over the range. It has some obvious drawbacks which tend to restrict its use in

practice; some of the more important disadvantages are:

disadvantages

- a. because the range is the difference between the largest and the smallest value, it is very sensitive to very large or very small observations. The inclusion of just one freak (that is, rare or unusual) value will greatly affect the range;
- b. the range is dependent on the number of observations. Increasing the number of observations can only increase the range; it can never make it less. This means that it is difficult to compare ranges for two distributions with different numbers of observations;
- c. while the range is very easy to calculate, it has the disadvantage that it ignores all the data in between the highest and the lowest values. If, for example, we consider the following three sets of data:

Set 1	3, 5, 7, 9, 11, 13, 15, 17, 17, 17, 17
Set 2	3, 5, 5, 5, 17, 17, 17, 17, 17, 17, 17
Set 3	3, 6, 7, 8, 10, 11, 14, 14, 15, 16, 17

we see that the range for all three sets are the same ($17-3 = 14$), but the degree of variation is by no means the same;

- d. it is difficult to calculate the range for data grouped in a frequency distribution. All we can really do is take the difference between the lower limit of the first class and the upper limit of the last class. This would obviously depend on the definitions of the classes, and is impossible if you have an open-ended class. However, some judgments can be made depending on the knowledge of the subject matter under observation. For practical purposes the open-ended classes are usually closed by guessing a value for the open-end.

example

Let us consider another example, the values of imports in various Pacific island countries in 1995.

Table 6.1 Total imports by country, 1995 (in thousand AUD)

Country	Value of imports (A\$' 000)
Cook Islands	65,363
Fiji	1,172,052
Kiribati	47,547
Marshall Islands	100,073
Papua New Guinea	1,741,935
Samoa	126,689
Solomon Islands	224,254
Tonga	98,047
Tuvalu	12,535
Vanuatu	124,521

Source SPESS 14, 1998, Pacific Community, Noumea

method

The range of the import values is the difference between the largest and the smallest value and in this case the range is:

$$\text{Range} = \$ (1,741,935,000 - 12,535,000) = \$1,729 \text{ million}$$

do not usually calculate range for grouped data

The range from a grouped frequency distribution is not usually calculated because of the reasons given in the section on disadvantages of the range. However, it can be obtained approximately by taking the difference between the upper limit of the last class and the lower limit of the first class. We must note that it can sometimes be very difficult and at times meaningless if either or both of these classes are open-ended. Let us consider once again the example of annual household cash income in two regions of a country, which are given in the following frequency distributions:

Table 6.2 Comparison of the range of income of two regions

Annual Household Cash Income			
Region A		Region B	
Income (\$)	Frequency (No. of Households)	Income (\$)	Frequency (No. of Households)
Less than 500 (200*)	137	Less than 1,000 (500*)	86
500 – 999	278	1,000 - 1,999	137
1,000 - 1,499	406	2,000 - 2,999	64
1,500 - 1,999	331	3,000 - 3,999	47
2,000 - 4,999	188	4,000 - 6,999	130
5,000 - 9,999	259	7,000 - 9,999	62
10,000 - 19,999	138	10,000 & over (20,000*)	88
20,000 & over (30,000*)	14		
Total	1,751		614

Source: Table 5.1 (illustrative data only)

* = Assumed limits

open-ended class intervals

Obviously we cannot calculate the range of income in such cases because of the presence of open-ended class intervals at both ends. However, if we do have to calculate income ranges for the two populations, we will be forced to make some assumptions. These assumptions may be well-founded or ill-founded, but nevertheless, if a decision has to be made, we will have to put some values in the open-ended classes. In the example above, the assumed values are:

<u>Assumed Values (\$)</u>	<u>Region</u>	
	'A'	'B'
Lower limit (first class)	200	500
Upper limit (last class)	30,000	20,000

range

The income ranges for the distributions may now be calculated as follows:

<u>Region 'A'</u>	<u>Region 'B'</u>
Income Range = $\$(30,000-200) = \$29,800$	Income Range = $\$(20,000-500) = \$19,500$

clearly state assumptions

In the above example, although we have derived the income ranges as \$29,800 for region 'A' and \$19,500 for region 'B' the ranges could be meaningless if it was later realised that the assumed values were incorrect. However, statisticians and planners are often confronted with such problems in their

everyday work and decisions such as those taken in the case of the ranges of income in regions 'A' and 'B' are the types of decisions which they have to live with. The important thing is that the assumptions applied to generate a result are clearly stated.

use points other than the highest and lowest

We can get around most of the problems of the range as a measure of the variation by using other points in the distribution rather than the two extreme points. Another choice would be to measure what we call the **quartile deviation** or the 'semi inter-quartile range' (that is, to measure the mean average difference between the **upper** and **lower quartiles**). For a discussion of upper and lower quartiles refer to Topic 5, "More on measures of location". The quartile deviation is not included in these notes, but covered in the advanced analysis course.

use percentiles

Another alternative is to use the difference between, say, the **10th** and the **90th percentile** (that is, those values for which 10 per cent and 90 per cent of the observed values are below). As measures of variation, both of these are quite useful. They are not affected by any one or two extreme or rare observations, they are less dependent on the number of observations, and they will tend to differentiate between different sets of observations. In the case of ungrouped frequency distributions, we can nearly always calculate these values. In the case of grouped frequency distributions, a problem occurs when one of the percentile or quartile values falls in an open-ended class.

Standard deviation

standard deviation as a measure of spread

Although the range is a simple measure of variation or spread, it has many disadvantages. We therefore need a measure which will overcome these disadvantages while still providing a good measure of variation. One method is the **mean deviation** where we measure the distance of observations from the mean. However, the mean deviation incorporates absolute values and these are difficult to deal with mathematically. The **standard deviation** is based on the same principles as the mean deviation, but in this case we eliminate the signs of the deviations from the mean by squaring them.

method

How does the standard deviations work? Like the mean, the standard deviation takes all the observed values into account. If there were no dispersion at all in a distribution, all the observed values would be the same. The mean would also be the same as this repeated value. So if everyone had the same height of 180cm, the mean would be 180cm. No observed value would deviate or differ from the mean. But, with dispersion, the observed values do deviate from the mean, some by a lot, some by only a little. Quoting the standard deviation of a distribution is a way of indicating a kind of "average" amount by which all the values deviate from the mean. The greater the dispersion, the bigger the deviations and the bigger the standard deviation.

principle of standard deviation

The standard deviation is found by adding the squares of the deviations of the individual values from the mean of the distribution, dividing this sum by the number of items in the distribution, and then finding the square root of this number.

Lets now explain the procedure for calculating the standard deviation in more detail.

In terms of a population consisting of N values $x_1, x_2, x_3 \dots x_N$ with a mean μ (pronounced mu or mew) the **standard deviation of a population** is defined as:

Formula

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

Definition

To describe the formula we will work through the steps to calculate the standard deviation.

First we calculate μ :

μ is calculated the same way as \bar{x} in the previous chapter (i.e. we add up all the numbers and divide by how many numbers there were). We call it μ when we are dealing with a population, rather than \bar{x} when it is a sample.

We subtract μ from each x value: $(x_i - \mu)$

Square each of these values: $(x_i - \mu)^2$

Sum these values to get the total: $\sum (x - \mu)^2$

Divide by the number of units in the population (N): $\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$

Take the square root of everything: $\sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$

The standard deviation of a population is denoted by σ (the Greek letter for small sigma).

variance

The square of the standard deviation is called the **variance** and is denoted by σ^2 . When we square the result of a formula which has a square root, the square root sign is cancelled out and disappears. We then have:

formula

$$\text{Variance } (\sigma^2) = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

sample variance

If we are dealing with a sample and wish to calculate the **sample variance** (or **sample standard deviation**) in order to estimate the value for the population, the formula is changed slightly. In this case s^2 stands for the sample variance, \bar{x} the sample mean, and n the sample size. The formula for the sample variance is then:

$$\text{Sample Variance } (s^2) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)}$$

and the **sample standard deviation** is given by:

$$\text{Sample Standard Deviation (s)} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

sample = (n-1)

These formulae for samples are effectively the same as those for populations, except that we have used the division $(n - 1)$ instead of N . The important thing to remember is that when calculating the variance or standard deviation of a sample, divide by $(n - 1)$. When calculating the variance or standard deviation of a population, divide by N .

large standard deviation = large spread

Note that the more the values of individual items differ from the mean, the greater will be the square of these differences and therefore the greater the sum of squares. Therefore, the greater the sum of squares, the larger will s (the standard deviation) be. Hence, the greater the dispersion, the larger the standard deviation will be.

example

We will now go through the calculation of the standard deviation using the following data.

Table 6.3: 2000 Secondary School Enrolment by Province, PNG

Province	Enrolments	Deviation from mean	Deviations squared
Western	961	-2,470	6,100,900
Gulf	1,523	-1,908	3,640,464
NCD	4,854	1,423	2,024,929
Central	3,344	-87	7,569
Oro	3,134	-297	88,209
SHP	1,682	-1,749	3,059,001
EHP	5,768	2,337	5,461,569
Simbu	6,182	2,751	7,568,001
	Mean = 3,431	0	27,950,642

Source: Illustrative data only

first find the mean

To calculate the standard deviation we first calculate \bar{x} .

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

$$= (961 + 1,523 + 4,854 + 3,344 + 3,134 + 1,682 + 5,768 + 6,182)/8$$

$$= 3,431$$

In Column 3 we subtract the mean value from the values for each year.

In Column 4 we square the deviations and sum these squared deviations, giving a total of 27,950,642.

data from a population

If the above data are considered to be from a population, then to derive the standard deviation we divide the sum of the squared deviations by the number of the observations ($N = 8$) and take the square root. In this case we have:

$$\text{Population Standard Deviation } (\sigma) = \sqrt{\frac{27,950,642}{8}} = \sqrt{3,493,830.25} = 1,869.18$$

data from a sample

However, if the data are considered to be a sample from a population, then to derive the standard deviation we divide the sum of the squared deviations by one less than the number of the observations or industries ($n-1 = 7$) and take the square root. In this case we have:

$$\text{Sample Standard Deviation } (s) = \sqrt{\frac{27,950,642}{7}} = \sqrt{3,992,948.86} = 1,998.24$$

In this example we would probably consider the data to be sample data, so would divide by 7.

awkward with a large set of numbers

Although this is a fairly simple procedure to calculate the standard deviation of a small set of numbers, it is quite a cumbersome procedure for a large set of numbers. First of all we have to determine the mean of the set, then calculate the deviations of each observation from the mean, square these and add them up. Even with the aid of a calculator the operations take quite a lot of time. It is best to use a computer to perform the calculations.

rearrange the formula

We can, however, make the calculation much easier by rearranging the formula for the variance. Thus, for a sample, we have:

Sample formula

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - \left(\left[\sum_{i=1}^n x_i \right]^2 / n \right)}{n-1}$$

population formula

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^N x_i^2 - \left(\left[\sum_{i=1}^N x_i \right]^2 / N \right)}{N}$$

steps for sample variance

Lets run through this formula for the sample variance.

For the sample variance we first square each individual x value: x_i^2

We then calculate the sum of those squared numbers: $\sum_{i=1}^n x_i^2$ Call this total A.

We also calculate the total of the individual x values: $\sum_{i=1}^n x_i$

We square this total: $\left[\sum_{i=1}^n x_i \right]^2$

and divide by n (the number in the sample): $\left(\left[\sum_{i=1}^n x_i \right]^2 / n \right)$ Call this total B

We then take A - B and divide by $(n-1)$:
$$\frac{\sum_{i=1}^n x_i^2 - \left(\left[\sum_{i=1}^n x_i \right]^2 / n \right)}{n-1}$$

not as complicated as it looks

Although the second formula looks more complicated, it is in fact much easier to use when we are using a calculator. For example, let us consider the following sample values which are the same observations that we had considered in Table 6.3.

example

961 1,523 4,854 3,344 3,134 1,682 5,768 6,182

total and the mean of the observations

- a. Calculating the variance of the sample the first way would entail firstly obtaining the total and the mean of the observations. We have:

$$\sum x_i = 27,448, \quad n = 8 \quad \bar{x} = 3,431$$

The deviations from the mean are:

-2,470 -1,908 1,423 -87 -297 -1,749 2,337 2,751

The sum of the squares of the deviation is 27,950,642. Thus, the variance is:

$$s^2 = 27,950,642 / 7 = 3,992,948.86$$

second method

- b. Calculating the variance using the second method or formula we need:

$$\sum x_i = 27,448 \quad \text{and} \quad \sum x_i^2 = 122,124,730$$

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^n x_i^2 - \left(\left[\sum_{i=1}^n x_i \right]^2 / n \right)}{n-1} \\ &= [122,124,730 - \{(27,448)^2 / 8\}] / 7 \\ &= (122,124,730 - 94,174,088) / 7 \\ s^2 &= 3,992,948.86 \end{aligned}$$

second method is easier and faster

Thus we see that if we use the memory function in a calculator, the second calculation can be done without having to write any intermediate results. You will also note that the variance derived using either of the two methods is the same (3,992,948.86) except that the second method is easier and faster.

Properties of the standard deviation

remember

When using the standard deviation it is important to remember the following points:

- ☆ the standard deviation is used only to measure the spread about the mean;
- ☆ the standard deviation is never negative;
- ☆ the standard deviation is sensitive to extreme values (called outliers). A single outlier can raise the standard deviation a great deal, distorting the picture of spread; and
- ☆ the greater the spread, the greater the standard deviation.

Coefficient of variation

the mean adds meaning to the standard deviation

The standard deviation by itself is not very meaningful unless it is considered along with the arithmetic mean. For example, a standard deviation of \$100 when the mean income is \$10,000 implies a much greater relative variation than a standard deviation of \$100 for a mean GDP figure of \$10,000,000. Also, comparing the variability of two populations with different units of measurement (for example, income levels in Papua New Guinea (Kina) and Vanuatu (Vatu) can be very difficult.

interested in variation from the mean

Hence, the variability in a set of observations can usefully be measured relative to a central measure such as the arithmetic mean. Such a measure is provided by the **coefficient of variation**, which is the ratio of the standard deviation to the arithmetic mean, usually expressed as a percentage, and is given by the formula:

formula

$$\text{Coefficient of Variation (C.V.)} = (\sigma / \bar{x}) \times 100$$

(The $\times 100$ converts the number to a percentage.)

can compare data

To compare the variability of two sets of figures would therefore involve comparing their respective coefficients of variation. The coefficient of variation allows for comparisons when:

- the means of the distributions being compared are far apart, or
- the data are in different units.

percentage

The units are converted to a common denominator (a percent).

example

If we look at the data in Table 6.3, we can calculate the coefficient of variation as:

$$\begin{aligned} \text{C.V.} &= (\sigma / \bar{x}) * 100 \\ &= 1,869.18 / 3,431 * 100 \\ &= 54.48\% \end{aligned}$$

illustrative example

Let's use some made up data to illustrate the coefficient of variation.

The mean income of homeowners in Australia is \$40,000 with a standard deviation of \$4,000. In

Kiribati, the mean income of home owners is \$12,000 with a standard deviation of \$1,200. (Note that the means are far apart and the standard deviations are different. Compare and interpret the relative dispersion in the two groups on incomes.

solution

The first impulse is to say that there is more dispersion in the incomes in Australia because the standard deviation is greater. However, when we convert the two measurements to relative terms using the coefficient of variation, we find that the relative dispersion is the same.

Australia

$$\begin{aligned} CV(\text{Australia}) &= (\sigma / \bar{x}) * 100 \\ &= \$4,000/\$40,000 * 100 \\ &= 10\% \end{aligned}$$

Kiribati

$$\begin{aligned} CV(\text{Kiribati}) &= (\sigma / \bar{x}) * 100 \\ &= \$1,200/\$12,000 * 100 \\ &= 10\% \end{aligned}$$

similar CV

In summary the income for both Australia and Kiribati have similar amount of variation.

example

We could also compare two different types of data – incomes and age of homeowners. We could compare the spread of incomes of homeowners in Australia with say the spread of the age of homeowners. The mean age of homeowners is 40 years with a standard deviation of 10 years.

age

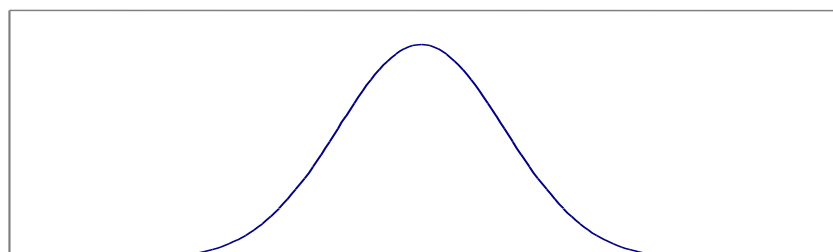
$$\begin{aligned} CV(\text{age}) &= (\sigma / \bar{x}) * 100 \\ &= (10/40) * 100 \\ &= 25\% \\ CV(\text{income}) &= 10\% \end{aligned}$$

We can see that there is greater relative dispersion in the ages of the homeowners than in their incomes.

Normal distribution

used extensively

A particular distribution that is used extensively in statistical theory is the **normal distribution**:



properties

The normal distribution has several key properties.

- it is symmetrical;
- it is bell shaped;
- mean of the distribution is the peak; and
- the area under the curve is always 1.

always have the four

Normal distributions can have different means and standard deviations, but they always have these four key properties.

everyday examples

Many phenomena in every day life can be described by the Normal curve, for example people's height. A small number of people in the population are very short, a small number are very tall, and the majority of the population fall in some middle range. Many other phenomena are also normally distributed, for example test scores and weights of people.

We could discuss the normal distribution extensively, but for now that is all you need to know.

Reference Ranges for a Standard Deviation**analysis of data normally distributed**

When analysing normally distributed data, the standard deviation is used with the mean to calculate where the data lie within certain reference ranges. The most important thing to understand about reference ranges is that for any set of normally distributed data:

reference ranges

☆ about 68% of the data lie in the interval $\bar{x} - s < \bar{x} < \bar{x} + s$

(That is, 68% of the data lie in the range from the mean minus the standard deviation to the mean plus the standard deviation)

☆ about 95% of the data lie in the interval $\bar{x} - 2s < \bar{x} < \bar{x} + 2s$

☆ about 99% of the data lie in the interval $\bar{x} - 3s < \bar{x} < \bar{x} + 3s$

where \bar{x} = the mean; and

s = the standard deviation

68% reference range

If we look at the data in Table 6.3, we can calculate the 68% reference range for the data as:

68% Reference range: $(\bar{x} - s, \bar{x} + s)$
 $(3431 - 1998.24, 3431 + 1998.24)$
 $(1432.76, 5429.24)$

That is, 68 % of the data lies in the range 1,432.76 to 5,429.24.

95% reference range

We can calculate the 95% reference range as:

95% Reference range: $(\bar{x} - 2s, \bar{x} + 2s)$
 $(3431 - 2(1998.24), 3431 + 2(1998.24))$

$$(3431 - 3996.48 , 3431 + 3996.48)$$

$$(-565.48 , 7427.48)$$

That is, 95 % of the data lies in the range -565.48 to $7,427.48$.

Summary of the measures of variability

RANGE

- ☺ is easily calculated, except for frequency distributions, and is well understood;
- ☹ is based on the two extreme observations and is thus very unstable;
- ☹ is difficult to manipulate mathematically;
- ☹ provides no information about the general behaviour of the distribution;
- ★ should only be used as a rough guide to the level of variability.

VARIANCE/STANDARD DEVIATION

- ☺ is a measure of variability using information from every observation;
- ☺ with some manipulation, the calculations are reasonably straight-forward;
- ☺ has a central role in mathematical and statistical theory and is very widely used;
- ☹ can be affected by extreme values;
- ★ is the most commonly used measure of variability.

COEFFICIENT OF VARIATION

- ☺ is independent of the units of observations. Therefore, it is useful in comparing distributions where the units of observations are different;
- ☹ a disadvantage of the coefficient of variation is that it is unstable when the arithmetic mean is close to zero.

One final characteristic of a distribution

understand the underlying structure

The objective of summarising a set of data is to make it possible to comprehend the underlying structure and pattern of the distribution of the values of the variable under consideration. The attempt in summarising the data is to reduce them to a few measures which would give us an indication of the central values, variation of the values, concentration of the frequencies and shape of the distribution. The frequency distribution describes the population we are considering, and the measures of location and variation help us to characterise the distribution by simple measures.

skewed distributions – asymmetrical

Another way of characterising a distribution is to study its **skewness** (that is, whether the distribution is not symmetrical and, if not, whether the observations are concentrated in the low or high values). Examples of skewed distributions are income, land holding size and household size. For such distributions, one is interested in finding out the type of skewness, whether there are more units with low values than units with high values, or whether there are more units with high values than units with low values.

'right' tail

A distribution is said to be **positively skewed** if large frequency values are concentrated to the left of the distribution and the distribution has small frequency values to the right of the distribution (that is, the distribution has a 'right tail' and has more low values than high values).

'left' tail

A distribution is said to be **negatively skewed** if large frequency values are concentrated to the right of the distribution and the distribution has small frequency values to the left of the distribution (that is, the distribution has a 'left tail' and has more high values than low values).

three main features

A distribution can be considered to have three main features which are of interest in studying a population. These features are:

- 1 its central values;
- 2 its variation from the central values;
- 3 whether the distribution is symmetric about the central values; and if not symmetric, whether it is leaning to the left or right.

3. The local market reported the following number of people buying vegetables for the past 9 days:

81 65 58 47 30 51 92 85 42

- (a) Calculate the range.
- (b) Calculate the standard deviation (assume the values are sample values).
- (c) Calculate the coefficient of variation.
- (d) Calculate the reference range that contains approximately 95% of observations.



Excel – functions 2

More statistical functions

In Topic 5, you were shown how to use the functions related to Measures of Location. In this section, those relevant to Measures of Variation are illustrated. You don't have to use the functions – instead you can set up a worksheet with the three columns (observation, deviation from the mean and deviations squared). See the computer notes for Topic 7 to set up the worksheet to calculate the variance, standard deviation and standard error from sample data. You have to be careful because the way your sample was selected determines how the standard error is calculated. If you have any doubts about the correct formula to use, contact the SPC Statistics Programme for help.

When calculating the variance or standard deviation, it might be more useful to use the 'worksheet' method rather than the Excel function. If you have the columns set up in your worksheet you can see the different components of the equation ($\sum x^2$ etc), and it would be easier to find out why you had a larger or smaller than expected deviation in your data. You also have to be aware that Excel uses its average function which includes 0 values in the count of observations (n) which might not be appropriate in all circumstances.

The range

You don't really need to use a function to calculate the range – use the sort buttons on the Standard toolbar. You can sort from smallest to largest with the  button, and from largest to smallest with the  button. Be careful when you sort data – either select ALL your data, or click with the mouse in the column you want to sort by: it is very easy to corrupt your data with the sort buttons (you don't get a warning like you do with the sort option on the Data menu).

Population variance

Excel calculates the variance for a POPULATION using the formula:

$$\frac{n \sum x^2 - (\sum x)^2}{n^2}$$

which is a different way of writing the one used in your notes.

Format: = **varp**(cell range)

Example =varp(A1:A2333) will calculate the variance for the POPULATION in cells A1 to cell A2333.

Sample variance

Excel calculates the variance for a SAMPLE using the formula:

$$\frac{n \sum x^2 - (\sum x)^2}{n(n-1)}$$

which again is a different way of writing the one used in your notes.

Format: = **var**(cell range)

Example =var(A1:A2333) will calculate the variance for the SAMPLE in cells A1 to cell A2333.

Population standard deviation

Excel calculates the standard deviation for a POPULATION using the formula:

$$\sqrt{\frac{n \sum x^2 - (\sum x)^2}{n^2}}$$

which is a different way of writing the one used in your notes.

Format: **= stdevp(cell range)**

Example = stdevp(A1:A2333) will calculate the standard deviation for the POPULATION in cells A1 to cell A2333.

Sample standard deviation

Excel calculates the standard deviation for a SAMPLE using the formula:

$$\sqrt{\frac{n \sum x^2 - (\sum x)^2}{n(n-1)}}$$

which again is a different way of writing the one used in your notes.

Format: **= stdev(cell range)**

Example =stdev(A1:A2333) will calculate the standard deviation for the SAMPLE in cells A1 to cell A2333.

Confidence interval

You can use Excel to calculate the confidence interval for a mean. You have to type in the standard deviation so the function is not that 'user friendly'.

Format: **= confidence(alpha,standard_dev,size)**

Where

alpha is the significance level used to compute the confidence level. The confidence level equals 100*(1 - alpha)%, or in other words, an alpha of 0.05 indicates a 95 percent confidence level.

Standard_dev is the population standard deviation for the data range and is assumed to be known.

Size is the sample size.

Example Suppose we observe that, in a sample of 50 commuters, the average length of travel to work is 30 minutes with a population standard deviation of 2.5. We can calculate with 95% confidence that the population mean is in the interval:

=CONFIDENCE(0.05,2.5,50) equals 0.692951. In other words, the average length of travel to work equals 30 ± 0.692951 minutes, or 29.3 to 30.7 minutes.