

## TOPIC 5

# MEASURES OF LOCATION

*The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill.*

*A. Einstein*

### The concept of location

#### **want to reduce the information in a frequency distribution further**

In Topic 3 we saw how observations of a variable could be summarised by forming a frequency distribution. This distribution contains a lot of information about the variable. It shows how many high and low values there are and by looking at some of the graphical presentations we get a visual impression of the distribution of this variable. In many situations this is sufficient, but we often need to reduce the information in a frequency distribution even further.

#### **comparisons between distributions**

For example, if we want to compare two distributions it can be difficult and confusing to look at all information. This is particularly true for people who do not have a good understanding of statistics. As statisticians we will often need to assist people who are not familiar with frequency distributions, such as administrators and policy makers. Under such circumstances, you will have to think of some other measures that will be easily understood by such people. In this topic we shall see how to calculate some values which can be considered to represent some feature or property of the distribution of a population under study. We can then use these values to make comparisons and to form the basis of more complex decisions.

#### **example**

Let us consider a very simple example. Suppose that a friend wishes to know how well or poorly you are doing in college. You might choose to collect all of your college grade reports and compile a simple frequency distribution such as the following:

<u>Grade</u>	<u>f (frequency)</u>
A	4
B	9
C	6
D	1
E	0

#### **efficient and effective reporting**

This would readily indicate that you had received four grades of A, nine grades of B, six grades of C and one grade of D. However, all this detail is probably not necessary to answer your friend's question and it would probably be tiresome for both you and your friend. In addition, presenting the data in this form makes it awkward for your friend to compare your performance to other college students. A better method would be to select one or two summary measures of your college grades so that they could be reported quickly and conveniently.

**location or comparative rank**

One summary measure that you would probably want to convey would be the general location of the distribution of your grades. You could simply state verbally that your college work was slightly below the B level; if you wished to be more precise, you could even report your numerical grade (point average), a single number that describes the general location of this set of scores. In either case you would summarise your performance by referring to a central point of the distribution that would be representative of your grades. It would clearly be misleading to describe your overall performance as being at the A or D grade, even though you did receive such grades. This is just one of many situations which benefit from the use of a **measure of location** or **central tendency**, that is, a single measure that attempts to describe the location of a set of scores.

**example**

Let us consider a situation which we, as people who are involved in statistical work, are often confronted. Consider the following table which shows two frequency distributions of annual household cash income in different regions of a country:

**Table 5.1 Comparison of two frequency distributions**

Annual Household Cash Income			
Region A		Region B	
Income (\$)	Frequency (No. of Households)	Income (\$)	Frequency (No. of Households)
Less than 500	137	Less than 1,000	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	88
20,000 & over	14		
<b>Total</b>	<b>1,751</b>		<b>614</b>

Source: Illustrative data only

**average and variability**

Suppose the government wants to compare household income in region A with that in region B. This kind of analysis will be very important for deciding policies for each region. However, presenting the data in the above table makes comparisons difficult. We have the same variable in each case but a different number of observations and different income classes. What we need to do is to look at the distributions for the two regions and to find some way of describing certain characteristics of each one, which can then be compared easily. There are several different characteristics we could choose, but in practice we tend to concentrate on just two: an **average** value of the variable and the **variability** (or spread) of the distribution. We choose these because they have an obvious meaning and they are all we usually need to describe the whole distribution effectively. These two measures form the basis of almost all statistical analysis and we shall deal with averages (or measures of location) in this topic.

**TIP**

**average** and **variability** form the basis of almost all statistical analysis.

## Definitions and types of averages

### statistical terms

In this topic we shall be concerned with observations of variables and we shall only be dealing with one variable at a time. The observations will be in their original state (for the case where the observations are grouped into a frequency distribution, refer to Topic 3, “More on measures of location”). In order to be able to make general statements that will be true about any set of data, we shall need to use some special statistical notations (or symbols) and definitions. We can use certain letters and symbols to stand for some items. There are, however, one or two new ideas that we must mention before we can go on to look at averages in detail.

### $x_n$

If we have a sample taken from a population we always use the letter  $n$  to denote the number of observations in the sample. The values of the particular variables observed are denoted by  $x_1, x_2, \dots, x_n$  (the symbol ‘...’ means ‘and so on’). Thus  $x_3$  means the third value of variable  $x$  in the sample. For example, suppose that you go fishing everyday for a week and the numbers of fish you catch are as follows:

Table 5.2 Number of fish caught per day

Observation	Day	Number of Fish
$x_1$	1	11
$x_2$	2	5
$x_3$	3	3
$x_4$	4	17
$x_5$	5	12
$x_6$	6	9
$x_7$	7	6
<b>Total</b>		<b>63</b>

Source: Illustrative data only

### notation

So  $x_3$  is the third value of the variable ‘number of fish caught’, so  $x_3 = 3$ .

### two types of population

The population from which we select the sample may be one of two types. Firstly it may have a definite size, that is, we can count all the individuals in the population. In this case the population is called ‘finite’ and the size of the population is denoted by  $N$ . Examples of finite populations are the people living in a country, the business enterprises operating in an island, and all the farms in a particular area. The second type of population has no limit on its size and we cannot count the number of individuals, such a population is known as ‘infinite’. Examples of infinite populations are all the taro plants that might ever be grown on an island, all the fish that might ever be caught in a area of the sea, and all the pigs that might ever be kept in a country.

### differences between populations and samples

We will want to distinguish between populations and samples, because there are some important differences. When we are dealing with data that are from a sample of a population, the notation will be different from when we are looking at data from a complete population. When we are dealing with the whole population we use letters from the Greek alphabet to denote the values we calculate; in particular, we shall be using the letters  $\mu$  (mu) and  $\sigma$  (sigma). The values calculated from the population are called parameters. For a sample, on the other hand, we use ordinary English letters to represent calculated values and we call these values estimates.

**estimate the population parameters**

Very often we do not have information about a population, rather we have a series of observations from a sample. What we do is to estimate the population parameters by calculating sample estimates. We will come back to this point when we talk about different types of averages.

**what does 'average' mean?**

The word 'average' is used a lot in everyday English. For example, people often refer to an average man, an above average performance and below average temperature. 'Average' is used in the sense of 'typical', 'usual' or 'normal'. We also use the word average a lot in statistics, although its meaning is not quite the same.

**example**

Most people think of the average of a group of numbers as the result of adding them all up and then dividing by however many numbers there are. For example, the average of 6.7, 9.6, 12.8, 13.0 and 15.9 would be:

$$\frac{6.7+9.6+12.8+13.0+15.9}{5} = \frac{58.0}{5} = 11.6$$

**arithmetic mean**

In fact there are several types of averages that we use in statistics, the one described above is known more accurately as the **arithmetic mean**.

**comparisons between different populations**

Very often in statistics we wish to make comparisons between different populations, in fact a large part of statistical theory is concerned with this problem. For example, we may want to compare household incomes in different areas, the incidence of tooth decay between different groups of school children or the weights of fish caught at different times of the year.

**TIP**

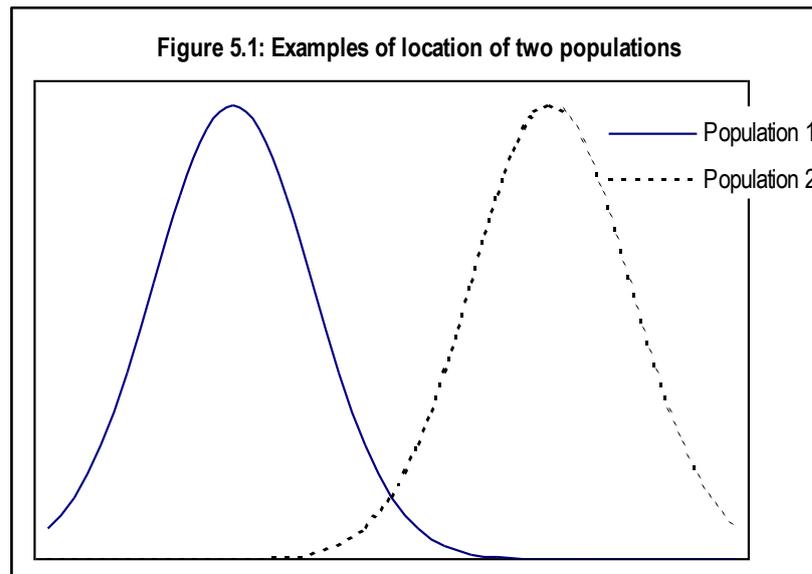
It is much easier when comparing populations to be able to determine if there are real differences between the **arithmetic means** of the populations than between some other types of averages.

**make sure the distributions are similar**

If the two populations we want to compare have very different types of distribution, then this comparison can be very difficult. However, in many situations we find that the shapes of the two distributions are quite similar, and comparisons are easier to make. In this case what we would like to do is to find one or two ways of describing the population, of summarising the distribution by certain characteristics.

**one summary value is a measure of location**

By a measure of the location of a distribution we mean finding one value which summarises in some way the size of all the different values in the distribution. Other terms are used to describe the same thing and some writers talk about averages being measures of central tendency or measures of central location; these are just different ways of expressing the same idea. We shall use the term **measure of location**. We can see in the diagram below that in general the values in Population 2 are larger than the values in Population 1.



### types of averages

In this topic we shall consider the following types of average as measures of location:

- a. arithmetic mean;
- b. median; and
- c. mode.

### ungrouped data

In discussing these averages we will consider observations in their original state. For a discussion of these averages when the observations are grouped in a frequency distribution, refer to the end of this topic, "More on measures of location". Two other important measures of location, quartiles and the geometric mean, are also discussed there.

## Arithmetic mean

### quantitative data only

The arithmetic mean involves performing calculations on the observed values of the variable. Therefore, the arithmetic mean is only applicable to quantitative variables (that is, variables which take numerical values).

If we observe a sample of  $n$  values of a particular variable, we can list these observations as  $x_1, x_2, \dots, x_n$ .

Then the arithmetic mean of this sample is written as  $\bar{x}$  (pronounced x bar) and is defined to be:

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

$\Sigma$

The Greek letter  $\Sigma$  (capital sigma) stands for 'the sum of', and the ' $i=1$ ' and ' $n$ ' above and below the  $\Sigma$  sign tells us that the sum is from  $x_1$  to  $x_n$ . The expression for  $\bar{x}$  is just a shorthand way of writing:

*'The mean of a set of  $n$  numbers is the sum of all the numbers, divided by  $n$ '.*

**example**

For the fish example in Table 5.2, the calculation would be:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{63}{7} = 9$$

Therefore, the arithmetic mean of the number of fish caught per day is 9.

**apply the mean back to the data**

In the above example the answer to the question ‘what is the arithmetic mean of the number of fish caught per day?’ was a number that could actually occur in practice. We see that on the sixth day the number caught was equal to this ‘average’ value.

**discrete variables**

However, for **discrete variables** such as ‘number of fish caught’, when we calculate the arithmetic mean and use the answer as our average result, there is no guarantee that we will get a number that can actually occur. Consider the following example:

**Table 5.3** Number of people in each household in a small village

Household Number	Number of people
1	5
2	9
3	1
4	3
5	4
6	5
7	2
8	7
<b>Total</b>	<b>36</b>

Source: Illustrative data only

In this case, the arithmetic mean of the observed values is  $^{36}/_8 = 4.5$  people per household.

**the mean can be confusing with discrete data**

This kind of result sometimes confuses people. After all you cannot have 0.5 of a person, so how can we say that the ‘average’ number of people per household is 4.5? The important thing to remember is that the arithmetic mean is an artificial concept. We use it because it is fairly easy to calculate and understand, and it is mathematically convenient if we want to do more advanced calculations.

This problem will only occur with discrete data, since with continuous data, by definition, any value within the possible range of values can occur in practice. Thus it is not difficult to understand what is meant by the statement ‘the arithmetic mean height of a group of men is 1.75 m’.

So we have to be careful when talking about the arithmetic mean of a set of discrete data. We must realise that it is a type of average, a measure of the location of the population from which the data came. It does not mean that it is the most likely value to occur in practice, or even that it can occur at all.

## The median

### *median splits a set of values into two equal parts*

The median is another type of average, or measure of location of a set of numbers, and basically it is a very simple concept. The **median** is that value which splits a set of values into two equal parts. It is the middle number of the set when arranged in order of size.

Suppose, for example, we had the following set of observations:

14	9	16	3	1	7	5
----	---	----	---	---	---	---

To find the median you have to arrange the observations in order of size as follows:

### *order the observations*

1	3	5	7	9	14	16
---	---	---	---	---	----	----

### *the middle value*

The median is the middle value, which in this case is 7. As we can see there are as many numbers less than 7 (1, 3 and 5) as there are numbers greater than 7 (9, 14 and 16).

You will also see that before the set of observations are re-arranged in order of size, the middle value is 3. This is **not** the median. In order to find the median or the middle value, you must first re-arrange the original set of observations in order of size. Thus, if we have  $n$  observations, the median will be the value of the  $\frac{n+1}{2}$ th observation in the ordered list.

### *even number of observations*

A problem occurs if we wish to determine the median of a set of observations when there is no middle value, that is, when the number of observations in the set is even. We adopt a convention to deal with this as follows. The median of the set:

1	3	5	7	9	14	16	21
---	---	---	---	---	----	----	----

is defined as the arithmetic mean of the two middle values, which in this case are 7 and 9. The median therefore is  $(7 + 9) / 2 = 8$ .

### *unaffected by outliers*

It is clear that if the median depends upon the value of the middle value or values in a series, it is unaffected by extreme high or low values (**outliers**). For example, consider the following information on the size of coconut plantations on a particular island, the values being in hectares:

1.3	1.3	1.5	1.7	2.0	2.1	2.3	2.7	2.8	3.0	3.7	5.0	5.5	7.0	120.1
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-------

### *the large value has no affect ...*

The last value represents the area of a commercial plantation, while the other figures are of small holdings. The arithmetic mean area of the plantations is 10.8 ha, the median is 2.7 ha. If the commercial plantation is replaced by a plantation with an area of 3.1 ha, the arithmetic mean then becomes 3.0 ha, while the median remains unchanged.

**... unlike the arithmetic mean**

The median therefore is unaffected by the value of very large or very small observations, while the arithmetic mean is. If there is doubt about the accuracy of observations at either extremes of the scale of measurements, the median is a better 'average' than the arithmetic mean.

**sample notation**

From a sample of observations we usually denote the median by  $\tilde{x}$  (pronounced x tilde). There is no universally accepted symbol to denote the median of a population.

**The mode****peak in the frequency**

If a set of observations has a peak in its frequency at a certain point then there is said to be a **mode** at that point.

As you will see in the examples given at the end of this topic, the distribution of a population can be **unimodal** (having only one mode); **bimodal** (having two modes); or **multimodal** (having more than two modes); or it can have no mode at all.

**the mode is a type of average**

Like the arithmetic mean and the median, the mode is a type of average. It is a measure of the location of the distribution.

**discrete data**

When dealing with sample observations, the concept of the mode is most useful in connection with frequency distributions. For a **discrete distribution** the mode is that value that occurs most often. For example, the number of occupants (people) per household (Table 5.3); the modal value is 5 people per household. This occurs more often than any other value.

**continuous data**

For a **continuous distribution** the determination of the mode is rather complicated, and so for our purposes we shall only be concerned with the modal class. This is the class with the highest frequency. So in Table 5.4 for example, 3.0 to 3.9 kg is the modal class for the fish data of Table 3.2.

**Table 5.4 Distribution of fish weights**

Class (kg)	Frequency	Cumulative Frequency
2.0–2.9	7	7
3.0–3.9	19	26
4.0–4.9	16	42
5.0–5.9	12	54
6.0–6.9	6	60
7.0–7.9	3	63
<b>Total</b>	<b>63</b>	

Source: Table 3.2

**example**

Determining the mode or modal class for a frequency distribution sometimes produces more meaningful results than calculating the arithmetic mean. For example, consider the following data. In a survey of school children, the number of teeth that needed filling were counted for each child; the data were as follows:

Table 5.5 Number of teeth requiring filling

Number of Teeth Requiring Filling	Number of Children
0	113
1	156
2	37
3	21
4	11
5	7
6	3
7 or more	2
<b>Total</b>	<b>350</b>

Source: Illustrative data only.

### **mean vs. mode**

The arithmetic mean number of teeth to be filled per child is 1.16; to a non-statistician this is rather a meaningless statement. It is probably more useful to say that the mode was one filling per child.

### **continuous data can cause problems**

Determining the modal class for continuous frequency distributions often produces problems. The modal class will vary depending on how the classes are defined, and for data with a fairly even distribution between classes a change in the definition of the classes can change the modal class. For this reason the mode is of limited value and should be used with care.

## **Summary of the different types of averages and their uses**

### **different averages**

To complete this topic we shall look at the different types of averages we have discussed and summarise their advantages and disadvantages, and in which circumstances they should be used.

### **Arithmetic mean**

#### **the 'mean'**

Often just referred to as the mean and is what most people refer to when they talk about averages.

#### **ADVANTAGES**

- ☺ understood by almost everybody;
- ☺ easily calculated and can be determined in all cases for quantitative data;
- ☺ takes account of all the values in a sample;
- ☺ important mathematically, it is easy to manipulate algebraically;
- ☺ particularly useful when comparing different populations, or estimating population parameters from sample estimates.

#### **DISADVANTAGES**

- ☹ affected by extreme high or low values;
- ☹ can produce an average which cannot occur in practice;
- ☹ can produce an average which does not really reflect the nature of the distribution;
- ☹ can only be used for quantitative data.

**MEDIAN****ADVANTAGES**

- ☺ easy to calculate;
- ☺ not affected by extreme values;
- ☺ can be determined in all situations except for nominal data.

**DISADVANTAGES**

- ☹ it does not use all the values in the sample;
- ☹ it can produce an average which cannot occur in practice;
- ☹ it is difficult to deal with algebraically.

***skewed distributions***

The median can also be used in many situations. It is best used, however, when dealing with observations from a skewed distribution (see Topic 6 for a discussion of skewed distributions), since in these cases it is probably more informative than the mean.

**Mode****ADVANTAGES**

- ☺ easy to calculate for some data;
- ☺ has an obvious meaning in many situations;
- ☺ unaffected by extreme values;
- ☺ can be used for nominal data.

**DISADVANTAGES**

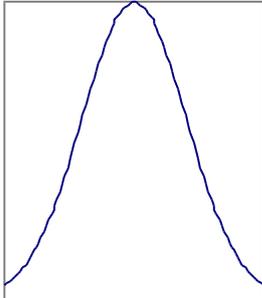
- ☹ difficult to calculate for continuous frequency distributions;
- ☹ sometimes it may not exist;
- ☹ for a frequency distribution it depends on the class intervals chosen;
- ☹ not well suited to mathematical treatment;
- ☹ does not use all the data in a sample.

***information on most 'popular'***

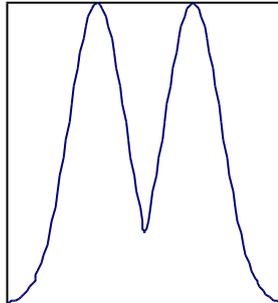
The mode is useful in the case of frequency distributions when information on the most "popular" class or the one which occurs with the greatest frequency is required.

## Frequently occurring shapes of population distributions

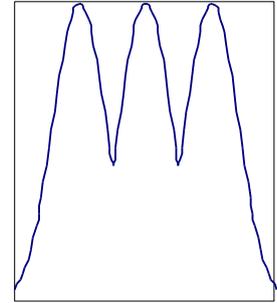
**A Unimodal Distribution**



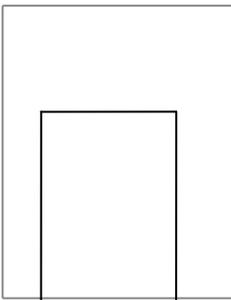
**A Bimodal Distribution**



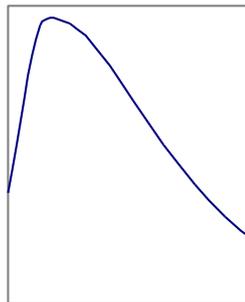
**A Multimodal Distribution**



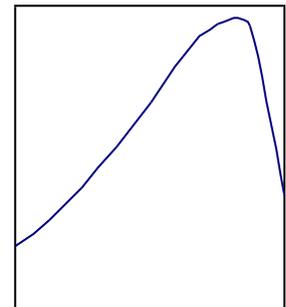
**Distribution With No Modes**



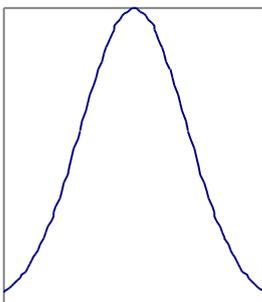
**Skewed Right Distribution**



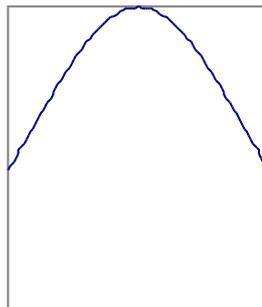
**Skewed Left Distribution**



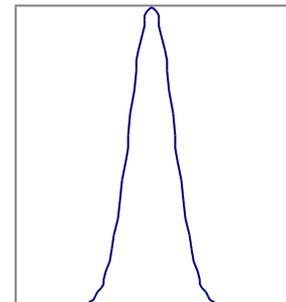
**Normal Distribution**



**Large Variance Distribution**



**Small Variance Distribution**





## ... Exercises ...

1. The number of fish caught on each day of a week are:

4    2    3    6    10    1    8

What is the value of  $x_2$  and  $x_6$ ?

$x_2 =$  \_\_\_\_\_  $x_6 =$  \_\_\_\_\_

2. The local shipping company employs 9 people. The length of service, in completed years, for each employee is as follows:

4    1    4    12    10    6    8    14    4

Calculate:

- (a) the mean
- (b) the median; and
- (c) the mode.

---

---

---

---

---

---

---

---

3. Police files reveal the ages of persons arrested for shop lifting:

**16, 41, 25, 21, 30, 17, 29, 50, 30 and 39.**

- (a) What is the mean age?
- (b) What is the median age?
- (c) What is the modal age?

---

---

---



## ... Self-Review ...

1. The following data represent the amount spent (in dollars) by a random sample of 14 households on basic food items for one month:

57	34	27	41	25	18	39
33	37	39	38	47	31	42

Calculate:

- (a) the mean
- (b) the median; and
- (c) the mode.

---

---

---

---

---

---

---

---

---

---

2. If one of the numbers had been incorrectly typed as 57 and should have been 250, what does this do to the mean, median and mode. Interpret this result.

250	34	27	41	25	18	39
33	37	39	38	47	31	42

---

---

---

---

---

---

---

---



## More on Measures of Location

### Introduction

#### *grouped data*

In Topic 5 we looked at the arithmetic mean and median when the observations are in their original state (that is, not grouped). In this section will consider the case where the observations are grouped in a frequency distribution. We will also consider two other useful measures of location; quartiles and the geometric mean.

### Calculating the arithmetic mean of a frequency distribution

#### *discrete data*

First of all let us consider the simple case of a **discrete frequency distribution**. The data given in Table 5.6 were for the number of destinations per holiday in the Cook Islands in 1991.

**Table 5.6** Number of Destinations on Holiday by Number of Persons (Cook Islands Visitor Survey 1991)

Number of Destinations	Persons
1	1,811
2	683
3	342
4	273
5	137
6 or more	171
<b>Total</b>	<b>3,417</b>

Source: Cook Islands Visitor Survey 1991, Survey Report No. 13, TCSP, Table 23, p. 31.

#### *calculating the mean*

If we want to calculate the mean number of rooms per number of destinations it is not sufficient just to add up the number of destinations and divide by the number of groups. We have many more single trips (one destination) than with a tour of 5 destinations, and we have to take this into account. In addition we have to deal with the last group '6 or more'. Since the frequency of this group is small, little error will be introduced if we assume an average size of 7 destinations per trip for all units in this last class. The calculation then is as follows:

$$\begin{aligned} \text{Arithmetic Mean } \bar{x} &= \frac{(1 \times 1,811) + (2 \times 683) + (3 \times 342) + (4 \times 273) + (5 \times 137) + (7 \times 171)}{3,417} = \frac{7,177}{3,417} \\ &= 2.1 \quad \text{destinations per trip} \end{aligned}$$

**formula for frequency distribution**

Just as we had a mathematical formula for the arithmetic mean of a set of ungrouped numbers, so we have a similar formula for use with a frequency distribution. In this case we call the number of classes 'k', the value for each class will be denoted by  $x$  (that is, the value for class  $i$  will be denoted by  $x_i$ ) and the frequency of each class by  $f$  (that is, the frequency for class  $i$  will be denoted by  $f_i$ ).

The formula for the mean  $\bar{x}$  is then given by:

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

**method**

In other words, the following steps need to be taken to calculate the mean of the distribution:

- for each class, calculate the product of the frequency for the class and the value for the class ( $f_i x_i$ );
- sum the products calculated in (a) over all  $k$  classes ( $\sum f_i x_i$ );
- calculate the total frequency by summing the frequencies for each class over all  $k$  classes ( $\sum f_i$ ); and
- divide the result calculated in (b) by the result calculated in (c).

**example**

When dealing with a continuous distribution we use the class midpoint as our value  $x_i$  as in the following example using the fish data of Table 3.2.

**Table 5.7 Weight of 63 fish (kg)**

Class (kg)	Class mark ( $x_i$ )	Frequency ( $f_i$ )	Frequency $\times$ Class mark ( $f_i x_i$ )
2.0–2.9	2.45	7	17.15
3.0–3.9	3.45	19	65.55
4.0–4.9	4.45	16	71.20
5.0–5.9	5.45	12	65.40
6.0–6.9	6.45	6	38.70
7.0–7.9	7.45	3	22.35
<b>Total</b>		<b>63</b>	<b>280.35</b>

Source: Table 3.2

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} = \frac{280.35}{63} = 4.45 \text{ kg}$$

**true class limits**

Note that the mid point of the range 2.0–2.9 is shown as 2.45. This is because the true range is 1.95 to 2.95 as we assume the weights of fish have been rounded to the nearest first decimal. Obviously, if we have original data (as we do in the case of the 63 fish) it is better to calculate the mean direct from the data (the result is similar in this case, 4.42 from the original data, 4.45 from the frequency distribution).

**example**

Let us look at some more examples of calculating the arithmetic mean from frequency distributions. Let us look at the distribution of percentage scores obtained by students in a mathematics examination:

**Table 5.8 Distribution of student scores in a mathematics examination**

Score %	Frequency $f_i$	Class mid-points $x_i$	$f_i x_i$
0 - less than 10	6	5	30
10 - less than 20	14	15	210
20 - less than 30	20	25	500
30 - less than 40	35	35	1,225
40 - less than 50	67	45	3,015
50 - less than 60	86	55	4,730
60 - less than 70	59	65	3,835
70 - less than 80	37	75	2,775
80 - less than 90	19	85	1,615
90 - 100	3	95	285
<b>Total</b>	<b>346</b>	<b>-</b>	<b>18,220</b>

Source: Illustrative data only

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} = 18,220/346 = 52.7$$

The arithmetic mean score in the examination was, therefore, 52.7 per cent.

**advantages and disadvantages**

The arithmetic mean has a lot of advantages as an average: it is easy to calculate, it always exists for quantitative data, most people understand it, and it is easy to use in more advanced statistical work. It does, however, also have some disadvantages which can produce difficulties in some situations. The value of the arithmetic mean can be severely affected by one or two large observations; this can happen when we have a distribution that has many small observations and a few large ones. In this kind of situation, using the arithmetic mean may be misleading.

**example**

To see what can happen, let us look at the following table:

**Table 5.9 Calculation of an Arithmetic Mean from a Skewed Distribution**

Net monthly income	Frequency	Class mid-points $x_i$	$f_i x_i$
0 - 50	42,872	25	1,071,800.0
51 - 150	1,213	100.5	121,906.5
151 - 250	1,591	200.5	318,995.5
251 - 350	1,861	300.5	559,230.5
351 - 500	1,383	425.5	588,466.5
501 - 700	827	600.5	496,613.5
701 - 900	607	800.5	485,903.5
900 and more	737	(1,000)*	737,000.0
<b>Total</b>	<b>51,091</b>		<b>4,379,916.0</b>

Source: Solomon Islands Statistical Bulletin No. 18/95, Table 3.1.2, p8. \* – assumed mid-point

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} = \frac{4,379,916}{51,091} = 85.7$$

**not an informative measure**

In the above example, although the arithmetic mean net monthly income per household is \$85.70, we can also see that 85 per cent of all households have fewer than this value (that is, they a monthly income of \$0-50). This form of average, therefore, does not seem to be very representative of this distribution and to tell anybody that the average income of households is \$85 may well be misleading. This is a common situation in many income and expenditure distributions. The arithmetic mean income is not very informative as an 'average'.

**other difficulties**

The calculation shown in the above table also illustrates three other difficulties with the arithmetic mean. When dealing with a frequency distribution with an **open-ended class** such as the class '900 and more', we need to make an assumption about the class mid-point. In our example we used \$1,000, but this was only a guess and it may not be accurate. If in reality the true mean average value for this class was \$1,500 then our arithmetic mean would be \$92.94 income per month.

**TIP**

The arithmetic mean we calculate can vary markedly depending on the assumption we make.

**true class means may differ from assumed ones**

The second difficulty with distributions which have many high or low values (that is, skewed distributions) comes from the use of the class mid-points as the assumed class means. In the example above we take the class mid-points as the assumed means, but the true means of each class will probably be less than these assumed means. For example, for class 51 - 150, the true mean is probably substantially less than 100.5 because more observations are probably clustered near the bottom of the class. We therefore tend to slightly inflate the mean value if we use the class mid-points as the assumed class means for positively skewed distributions (that is, distributions which have many low values).

**loss of information by rounding**

The third difficulty that can arise with the arithmetic mean is that we can obtain a value which obviously does not exist. In the example in Table 5.6, we were dealing with discrete data, but we obtained a mean number of destinations of 2.1. Obviously we cannot obtain a value of 2.1 destinations per holiday from visitors and many non-statisticians find such answers extremely difficult to understand. We could round the answer to the nearest whole number, in this case 2, but by doing this we lose a lot of information.

**bimodal distributions**

A further difficulty occurs with distributions that are 'bimodal' (that is, the frequency distribution has two crests, not one). Here the mean will probably lie between the two crests and not be representative of the distribution (it may be even misrepresentative of the distribution). As an example of this problem, a young man may be a bit disappointed if, after being told you were going to introduce him to three single, attractive ladies with an average age of 20 years, he found out the three were aged 1 years, 2 years and 57 years!

## Calculating the median of a frequency distribution

### two methods

To calculate the median from a frequency distribution, two methods can be used. The first uses a method of calculation direct from the table, the second uses an accurate diagram of the ogive. To illustrate the first method we will use the fish data. This is produced below:

Table 5.10 Distribution of Fish Weights

Class (kg)	Frequency	Cumulative Frequency
2.0–2.9	7	7
3.0–3.9	19	26
4.0–4.9	16	42
5.0–5.9	12	54
6.0–6.9	6	60
7.0–7.9	3	63
<b>Total</b>	<b>63</b>	

Source: Table 3.2

### finding the middle value in grouped data

Since there are 63 observations and the median is the middle value, the median is the value of the 32nd observation in the ordered list (remember that the median observation is the

$$\frac{(n+1)}{2} \text{th observation, not the } \frac{n}{2} \text{th observation.}$$

From the cumulative frequency column of Table 5.10 we see that the 32nd observation falls in the class 4.0 - 4.9 kg (the median class). The cumulative frequency of the group preceding the median class group (denoted as  $c$ ) is 26, so we need to go to the

$$\left\{ \frac{(n+1)}{2} \right\} - c = \left\{ \frac{(63+1)}{2} \right\} - 26 = 6 \text{th observation}$$

in the class group 4.0–4.9 kg. But what value does this observation take?

Table 5.10 tells us that there are 16 observations in the median class 4.0 - 4.9 kg, but we have no information on what values within this class the 16 observations take. They could be all 3.96 or they could be all 4.94.

**TIP**



By convention, we assume that the values within the class are evenly distributed across the class interval.

**assumptions**

To do this, we assume:

- 1 The difference between the ordered observations is equal; and
- 2 The difference between the true lower class limit and the observation with the lowest value and the difference between the true upper class limit and the observation with the highest value is half the difference between ordered observations.

**example**

To see this in a simple example, suppose we have a class with a frequency of 5 and true class limits of 0 and 10. We assume that the 5 observations are evenly spread across the class, that is that they take the values 1, 3, 5, 7 and 9. The difference between ordered observations is equal (two) and the difference between the true lower class limit and the observation with the lowest value and the difference between the true upper class limit and the observation with the highest value is half the difference between ordered observations (one). Note that the difference between ordered observations is equal to the class interval ( $i$ ) divided by the class frequency ( $f$ ) (that is,  $i/f = 10/5 = 2$ , in this case).

**fish example**

Returning to our fish data example, we now have a method of locating the median. The following steps are required:

**method**

start from the true lower class limit (denoted as ' $\ell'$ ') of the median class group (in our example ' $\ell' = 3.95$ ). This will account for  $c = 26$  observations in the ordered list;

add half the difference between the ordered observations in the median class group to locate the first observation in the median class group

$$\text{i.e. add } (\frac{1}{2}) \times (i/f) = (\frac{1}{2}) \times (1/16) = 1/32 = 0.03125;$$

to locate the  $m^{\text{th}}$  observation in the median class group, we then add  $m - 1$  times the difference between the ordered observations

$$\text{i.e. add } (m - 1) \times (i/f) = (m - 1) \times (1/16); \text{ and}$$

since the median is the  $\{\frac{(n+1)}{2}\} - c^{\text{th}}$  observation in the median class group, we need to add

$$\begin{aligned} & ([\{\frac{(n+1)}{2}\} - c] - 1) \times (\frac{i}{f}) \\ &= ([\{\frac{(63+1)}{2}\} - 26] - 1) \times (\frac{1}{16}) \\ &= \frac{5}{16} = 0.3125 \end{aligned}$$

**median**

So the median is:

$$3.95 + 0.03125 + 0.3125 = 4.29$$

$$\ell + \left(\frac{1}{2}\right) \times \left(\frac{i}{f}\right) + \left(\left[\left\{\frac{(n+1)}{2}\right\} - c\right] - 1\right) \times \left(\frac{i}{f}\right)$$

**formula**

From a) to d) above, we can see the median is given by the formula:

$$\begin{aligned}\tilde{x} &= \ell + \left(\frac{1}{2}\right) \cdot \left(\frac{i}{f}\right) + \left(\left[\left\{\frac{(n+1)}{2}\right\} - c\right] - 1\right) \times \left(\frac{i}{f}\right) \\ &= \ell + \left(\frac{i}{f}\right) \cdot \left[\left(\frac{n}{2}\right) + \left(\frac{1}{2}\right) - c - 1 + \left(\frac{1}{2}\right)\right] \\ &= \ell + \left(\frac{i}{f}\right) \cdot \left[\left(\frac{n}{2}\right) - c\right]\end{aligned}$$

**notation**

- $\ell$  is the true lower limit of the median class;
- $i$  is the class interval of the median class;
- $f$  is the class frequency of the median class;
- $c$  is the cumulative frequency of the class preceding the median class; and
- $n$  is the number of observations

**example**

and applying this formula to our fish data example gives:

$$= 3.95 + \left(\frac{1}{16}\right) \cdot \left[\left(\frac{63}{2}\right) - 26\right]$$

$$= 4.29$$

**discrete distribution**

For a **discrete distribution** the calculation is much simpler. For example, consider the data on the number of rooms per housing unit. This had as a frequency distribution:

**Table 5.11** Number of Destinations on Holiday by Number of Persons (Cook Islands Visitor Survey 1991)

Number of Destinations	Persons	Cumulative Frequency
1	1,811	1,811
2	683	2,494
3	342	2,836
4	273	3,109
5	137	3,246
6 or more	171	3,417
<b>Total</b>	<b>3,417</b>	

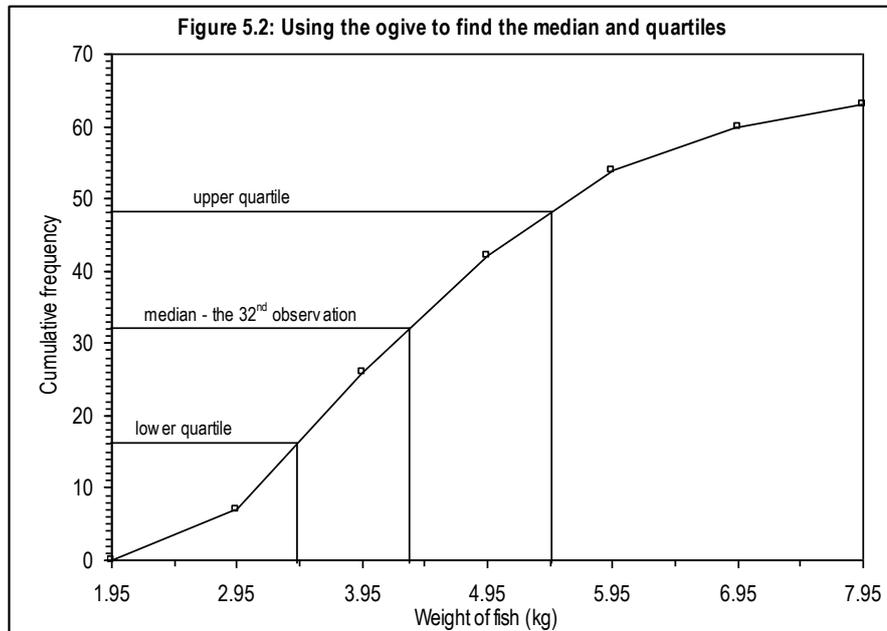
Source: Cook Islands Visitor Survey 1991, Survey Report No. 13, TCSP, Table 23, p. 31.

**median calculation**

Since there are 3,417 observations, the median is the  $(3,417 + 1) / 2 = 1,709^{\text{th}}$  observation in the ordered list. From the cumulative frequency column we can see that the  $1,709^{\text{th}}$  observation has a value of 1, i.e. the median number of destinations per holiday is 1.

**use the ogive**

Another way to approximately determine the median is directly from the **ogive** of the frequency distribution. Consider as an example the 'less than' ogive of the fish data.



Source: Table 5.10

**read off the horizontal scale**

The median is the  $32^{\text{nd}}$  observation so we simply read off the horizontal scale the position corresponding to the frequency 32 on the vertical scale, as shown in Figure 5.2. We find that the median is 4.4 kg which approximately agrees with the value obtained from the first method.

**tends to over estimate**

In determining the median from the ogive, we will slightly overestimate the median. In drawing the ogive we in effect assume that the data within a class are equally spread across the class, with the largest value coinciding with the true upper class limit and the difference between the true lower class limit and the lowest value being the difference between data values. This is equivalent to replacing the  $n/2$  term in the formula for the median used earlier by the term  $(n+1)/2$ . However, since determining the median from the ogive is an approximation (partly due to the inaccuracy of drawing the ogive), such a minor overestimation is unlikely to be of importance.

**Quartiles****four equal parts**

The median is the value of that observation which divides the total frequency into two equal parts. In the same way we can determine other values which divide the frequency into other fractions. **Quartiles**, as their name suggests, divide the total frequency into four equal parts.

**lower and upper quartiles**

The **lower (or first) quartile** will have one quarter of the observations less than this value and three quarters of the observations greater than this value. The **upper (or third) quartile** has three quarters of the observations less than this value and one quarter of the observations greater than this value.

**formula**

In terms of the number of observations, then, we have:

$$\begin{aligned} \text{Lower quartile is the } & \frac{n+1}{4} \text{th value of the ordered list.} \\ \text{Median is the } & \frac{n+1}{2} \text{th value of the ordered list.} \\ \text{Upper quartile is the } & \frac{3(n+1)}{4} \text{th value of the ordered list.} \end{aligned}$$

**example**

In the case of the fish data, where  $n$  is equal to 63, these correspond to:

$$\begin{aligned} \text{Lower quartile} & \quad 16^{\text{th}} \text{ value of the ordered list} \\ \text{Median} & \quad 32^{\text{nd}} \text{ value of the ordered list} \\ \text{Upper quartile} & \quad 48^{\text{th}} \text{ value of the ordered list} \end{aligned}$$

The **first or the lower quartile ( $Q_1$ )** can be calculated using the formula:

$$Q_1 = l + \left(\frac{i}{f}\right) \cdot \left[\left\{\frac{(n-1)}{4}\right\} - c\right]$$

and the **third or the upper quartile ( $Q_3$ )** can be calculated using the formula:

$$Q_3 = l + \left(\frac{i}{f}\right) \cdot \left[\left\{\frac{(3n+1)}{4}\right\} - c\right]$$

**same logic as median**

These formulae can be derived using the same logic as that used to derive the formula for the median earlier in this section.

**can also use the ogive**

We can also derive approximate values of the quartiles from the **ogive** in the same way as the median. In Figure 5.2 the lower quartile was found by reading off the horizontal scale the position corresponding to the frequency 16 on the vertical scale. The upper quartile was found in the same way, finding the  $x$  axis value for the 48<sup>th</sup> observation. These values are 3.4 kg and 5.4 kg respectively.

The values derived for the quartiles from the ogive are approximations for the same reasons as the median.

## The geometric mean

### formula

The **geometric mean** of  $n$  numbers  $x_1, x_2 \dots x_n$  is defined as:

$$\text{Geometric Mean} = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n}$$

### notation

A shorthand way of writing this is:

$$\text{Geometric Mean} = \left( \prod_{i=1}^n x_i \right)^{1/n}$$

### product

The sign  $\Pi$  (the Greek letter pi) means the product of a set of numbers in the same way as  $\Sigma$  means the sum of a set of numbers. As an example, the geometric mean of 3, 7 and 9 is:

### example

$$(3 \times 7 \times 9)^{1/3} = (189)^{1/3} = 5.74$$

### can be difficult to calculate

The geometric mean can be quite difficult to calculate, especially for a long series of numbers. Some calculators have special functions to help with this calculation and if a computer is available the calculations are quite simple.

### cannot calculate for certain data

We cannot calculate the geometric mean if any of the values are either negative or zero. Its use is fairly restricted because of this but it is quite useful when dealing with index numbers and growth rates.

### example

For example, suppose we have the following annual increases in a retail price index for a country: 1992 to 1993, 16.7 per cent; 1993 to 1994, 10.8 per cent; 1994 to 1995, 2.1 per cent, and 1995 to 1996, 4.5 per cent. The average annual retail price index increase from 1992 to 1996 is given by the geometric mean of these numbers and not the arithmetic mean. The average growth rate is calculated by first calculating the geometric mean of the following numbers; 1.167, 1.108, 1.021 and 1.045:

$$\text{Geometric mean} = (1.167 \times 1.108 \times 1.021 \times 1.045)^{1/4} = 1.0838$$

### used for index numbers

So the average growth rate is 8.38%. Note that the arithmetic mean would give an answer of 8.53%.

The geometric mean will always give an answer less than or equal to the arithmetic mean. It is used in preference to the arithmetic mean in calculating average growth rates since growth rates compound over time.

## Summary of additional types of averages

### Quartiles

#### ADVANTAGES

- 😊 fairly easy to calculate;
- 😊 not affected by extreme values; and
- 😊 can be determined in all situations except for nominal data.

#### DISADVANTAGES

- 😞 is not an 'average' but a measure of location;
- 😞 does not use all the values in the sample;
- 😞 can produce an measure which cannot occur in practice; and
- 😞 is difficult to deal with algebraically.

### Geometric mean

#### ADVANTAGES

- 😊 use all the data in a sample;
- 😊 less affected by extreme values than the arithmetic mean;
- 😊 can be manipulated algebraically; and
- 😊 particularly useful to calculate averages of growth rates, index numbers and percentage change.

#### DISADVANTAGES

- 😞 can be difficult to calculate;
- 😞 is of no use when there are zero or negative values. Is greatly affected by low values; and
- 😞 is not generally well understood.



# Excel – functions

## What is a function?

A function is a predefined formula that performs calculations by using specific values, called arguments, in a particular order, called the syntax. For example, the SUM function adds values or ranges of cells, and the AVERAGE function calculates the arithmetic mean for selected ranges of cells. There are approximately 150 functions in Excel.

## Count

Count is a useful function in Excel for finding out how many records (or rows) are in the data or for checking that you have all the records you should have (e.g. that you imported the records correctly from IMPS). The COUNT function is also a 'sub-function' of the AVERAGE function.

Format:     **= count(cell range)** will count all the cells with numbers (including 0) but NOT blanks or empty cells.

**= counta(cell range)** will count cells with an entry (i.e. numbers AND text) but NOT blanks.

Example     **=count(A1:A234)** will return the number of cells with data in this range.

You don't have to use the count function to find out how many records are in your dataset. The quick way to count the number of records is to double click on the bottom of the active cell to go to the last row in the data. Position the mouse so it looks like the following illustration and double click:

1	d+altobwbc
2	\$6.10
3	\$21.13

Then subtract one from the last row number (because you have a row containing titles) and that is the number of records in your Excel worksheet.

## Average

When working with numeric data you can use Excel to calculate the average of the data. The average function in Excel calculates what is technically called the arithmetic mean. Excel adds up all the values in the specified data range, counts how many values it added, and divides the total by the number of values it counted.

### WARNING



You can have BIG problems with the COUNT part of the average function. If the cells contain text, logical values, or empty cells, the values are ignored in the count; BUT, cells with the value zero are included.

When averaging cells, keep in mind the difference between empty cells and those containing the value zero. Empty cells are not counted, but zero values are.

Format:     **= average(cell range)** will calculate the arithmetic mean with  $n$  as the COUNT of all the cells with numbers (including 0) but NOT blanks or empty cells.

Example     **=average(A1:A234)** will return the arithmetic mean of the cell range.

## Median

The median is another common statistical summary measure used for data analysis. Remember that the arithmetic mean can be distorted by extremely large or small values, but the median is not. As a general guide, if you are analysing income, expenditure and sales values – any 'currency' number you should calculate both the average and the median and see which better summarises the distribution in your data.

The median represents the mid-point in your data SORTED from smallest to largest (i.e. ascending). That is, ½ the values in your data occur above it, and ½ below it.

Format:     **= median(cell range)** will calculate the median of all the cells with numbers (including 0) but NOT blanks or empty cells. You do NOT have to sort the data.

Example     =median(A1:A234) will return the median of the cell range.

## Rounding

When you use decimal place options to format cells you are changing the way Excel DISPLAYS the data, not how the data is STORED. To change the way the data is stored you have to ROUND the data.

Format:     **= round(cell range, number of decimal places)** will round the number to the specified number of decimal places. To round a number to the nearest integer (i.e. to round a number with decimal places to the nearest whole number) you enter 0 as the number of DP argument.

Example     =round(A1,2) will round the number in cell A1 to 2 decimal places.



**NOTE** To round data to a whole number, enter a negative value as the number of decimal places. For example the formula =round(\$3,456,789,-5) would return \$3,500,000.

## Vertical Lookup (VLOOKUP)

The VLOOKUP function is very useful in Excel based data collection and processing (collections like the Consumer Price Index). You use VLOOKUP to assign descriptors to classification codes. So for the CPI, you might have entered the Item Code. You would use VLOOKUP to enter the description for the code.

Basically the LOOKUP function takes the contents of a cell, and looks for the matching cell in another location, and returns the value you tell it to. So for the Item code, the VLOOKUP would lookup the data entry Item Code and compare it to the full Item Classification. You would return the description for the item.

The VLOOKUP function moves VERTICALLY down the rows of a lookup table, looking for matching information in the first column of the other location.



**WARNING** The code list with the description (i.e. your full classification) **MUST BE IN SORTED IN ASCENDING NUMERIC ORDER** for the VLOOKUP to work. If the classification is not sorted, the VLOOKUP will return rubbish, so you will know that you have to fix it.

The format of the VLOOKUP function is:

**=VLOOKUP** (**lookup\_value**, **table\_array**, **col\_index\_num**)  
 The number you want to assign the code to      The cell range of the item and descriptor (i.e. the classification)      The value you want returned – if your classification is code and descriptor, you will return the second number.

The following is an example of how to use the vlookup function. Imagine you have a worksheet which looks like this:

	A	B	C	D	E	F	G	H
1	household #	age	sex	Marital st	School Attendance			
2	1	43	2	1	2			
3	1	49	2	1	2			
4	1	15	1	5	3			
5	1	14	2	5	4			
6	1	57	2	1	1			
7	1	38	2	1	2			
8	1	26	1	1	2			
9	1	21	2	5	2			
10	1	21	1	5	1			

You decide to assign the descriptors to the Marital st variable. The first thing you have to do is enter the descriptor for each code IN NUMERICAL ORDER. So in Columns G and H you type in the Marital Status classification like this:

	A	B	C	D	E	F	G	H	I
1	household #	age	sex	Marital st	School Attendance		Marital status		
2	1	43	2	1	2		1	Married	
3	1	49	2	1	2		2	Widowed	
4	1	15	1	5	3		3	Divorced	
5	1	14	2	5	4		4	Separated	
6	1	57	2	1	1		5	Never married	
7	1	38	2	1	2				
8	1	26	1	1	2				

You insert a column beside the Marital St one and label it Marital code. In cell E2 you type the vlookup formula: =vlookup(d2,\$h\$2:\$i\$6,2) and press Enter. The descriptor Married should be displayed:

E2		= =VLOOKUP(D2,\$H\$2:\$I\$6,2)									
	A	B	C	D	E	F	G	H	I	J	
1	household #	age	sex	Marital st	Marital code	School Attendance		Marital status			
2	1	43	2	1	Married	2		1	Married		
3	1	49	2	1		2		2	Widowed		
4	1	15	1	5		3		3	Divorced		
5	1	14	2	5		4		4	Separated		
6	1	57	2	1		1		5	Never married		
7	1	38	2	1		2					

Position the mouse over the fill handle in the bottom right of cell E2 so it is shaped like a cross and double click to AutoFill the formula:

E2		= =VLOOKUP(D2,\$H\$2:\$I\$6,2)									
	A	B	C	D	E	F	G	H	I	J	
1	household #	age	sex	Marital st	Marital code	School Attendance		Marital status			
2	1	43	2	1	Married	2		1	Married		
3	1	49	2	1		2		2	Widowed		
4	1	15	1	5		3		3	Divorced		
5	1	14	2	5		4		4	Separated		
6	1	57	2	1		1		5	Never married		

The formula will be copied down the data series (NOTE that AutoFill will stop when it finds a blank cell to the LEFT so be careful with blank cells). The sheet now looks like this:

	A	B	C	D	E	F	G
1	household #	age	sex	Marital st	Marital code	School Attendace	
2	1	43	2	1	Married	2	
3	1	49	2	1		2	
4	1	15	1	5		3	
5	1	14	2	5		4	
6	1	57	2	1		1	

**NOTE**

Use the \$ sign in your cell reference to make the VLOOKUP only look in that cell range. If you do not use the \$ sign, the VLOOKUP will continue looking down the columns your reference is in.

# One step further ...

## ***Creating age from date of birth***

You can perform calculations on dates in Excel, providing that the cell format is set for date information. The NOW() function in Excel returns the current date from your computer.

1. Open up a new workbook.
2. Select Column A and set the format of the column to date.
3. In cell A1 enter your date of birth in the format dd/mm/yy.
4. Select Column B and set the format of the column to number.
5. In cell B1 type the formula =now()-A1. This returns how old you are in DAYS. You now convert this to years.
6. In cell C1 type =B1/365.25. Cell C1 now contains your age.

## Summary

To	Do this
<b>Count data</b>	Enter the formula for the count with the cell range you want to count. The count formula = count(cell range). Note that the count will count cells with 0, but not cells with text or 'blanks'.
<b>Average</b>	Calculate the arithmetic mean by entering the formula = average(cell range). Note that the average uses the count function, so empty cells are not included in the count of values.
<b>Median</b>	Calculate the median by entering the formula = median(cell range).
<b>Round data</b>	Round data to the specified number of places by entering the formula = round(cell reference,number of decimal places). Round a whole number by using a negative value as the number of decimal places.
<b>VLOOKUP</b>	Use VLOOKUP to assign descriptors to codes. Enter the formula = vlookup(lookup value,table_array,column to return). Note that the table_array is where the classification is stored and it MUST be sorted in ascending order for the VLOOKUP to work.