DESCRIPTIVE STATISTICS : SUMMARISING THE OBSERVATIONS

4.1 Introduction

In the previous topic we saw how a number of observations of some variable could be summarised by forming a frequency distribution. This distribution will contain a lot of information about the variable, it will show how many high values there are, how many low ones, and by looking at the frequency histogram we can often get some idea of the distribution of this variable in the population. In many situations this is sufficient, but we often find that we need to reduce the amount of information in the frequency distribution even further. If we want to compare two distributions, for example, it can be difficult and confusing to have to look at all the information. In this topic we shall see how we can calculate one or two values that can be considered to represent some feature or property of the distribution. We can then use these values to make comparisons and to form the basis of more complex decisions.

As an example, consider Table 4.1 which shows two frequency distributions of fork length of yellowfin taken by purse-seine vessels.

TABLE 4.1 : COMPARISON OF TWO FREQUENCY DISTRIBUTIONS

Country A		Country B	
Fork length (cm)	Number	Fork length (cm)	Number
30-39	132	Less than 40	167
40-49	219	40-49	345
50-59	253	50-54	369
60-69	126	55-59	492
70-89	61	60-64	318
90-109	124	65-69	160
110-129	182	70-79	114
130+over	135	80-99	281
		100-119	294
Total	1232	120+over	203
		Total	2743

Using the data presented in Table 4.1, comparison is difficult. We have the same variable in each case, but different numbers of observations and different classes. What we need to do is to look at the distributions for the two countries and to find some way of describing certain characteristics of each one, which we can then compare quite easily. There are several different characteristics that we could choose, but in practice we tend to concentrate on just two: the average size and dispersion. We choose these because they have an obvious meaning and most people can understand them, and because in practice we find that they describe the whole distribution effectively. These two measures form the basis of almost all statistical inference, but we shall only be dealing with averages and dispersion as ways of describing or summarising a set of observations.

4.2 Some special notation and concepts

In this topic we shall be concerned with a number of observations of some variable and, as before, we shall only be dealing with one variable at a time. The observations may be grouped into a frequency distribution or they may be in their original state, but the principles in each case will be the same. In order to be able to make general statements that will be true about any set of data, however, we shall need to use some special statistical notation. We can use certain letters and symbols to stand for some items, and these usually will be the same as those introduced in the preliminary session to this course. There are, however, one or two new ideas that we must mention before we can go on to look at average and dispersion in detail.

We will need to distinguish between populations and samples because there will be some important differences. When we are dealing with the whole population we generally use letters from the Greek alphabet to denote values we calculate; in particular, we shall be using the letters μ (mu) and σ (sigma). For a sample, on the other hand, we use ordinary letters to represent values.

In practice we are usually interested in the population values of averages and measures of dispersion, rather than just the sample values. The population values are referred to as parameters of the population to distinguish them from values derived from samples. Very often we do not have information about a population, rather we have a series of values from a sample. What we do is to estimate the population parameters by calculating sample statistics.

4.3 Measures of average values

An average is a measure of the size of a set of variables and it forms the basis of a lot of more advanced statistical work. There are several different types of average that we can calculate, or find, and they have different properties; which one we use in any particular situation will depend upon what we want to do. We shall look at three types of average: the arithmetic mean, the median and the mode; these are the ones most commonly used, although there are other types for more specialised uses.

4.3.1 The arithmetic mean

The most common and widely understood type of average is the arithmetic mean. If we have a sample of values, x_1, $x_2 \ldots x_n$ of some variable x, then the arithmetic mean of this sample, which is denoted by $\bar{x}$ (pronounced 'x bar'), is given by:

$$\bar{x} = (x_1 + x_2 + \ldots + x_n)/n$$

This may be written as $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

As we noted earlier, the use of the subscript 'i' is a convention to indicate the observation under consideration. Thus, $\sum_{i=1}^{n}$ means the sum of all the observations, from the first to the nth; similarly $\sum_{i=2}^{4}$ means the sum of the second, third and fourth observations, and so on. For the rest of this course we will be referring to the sum of n observations. Usually, for the sake of simplicity, we will abbreviate this and just use the symbol Σ on its own, and will omit the subscript i. So we will write the formula for

the mean as simply $\bar{x} = \Sigma x/n$, and we understand that this is really a shorthand way of writing $\frac{1}{n}\sum_{i=1}^{n} x_i$.

We can express the same idea in words by saying that the arithmetic mean of a set of values is given by the sum of the values divided by the number of values in the set. The arithmetic mean is easy to calculate and will always exist for any set of values.

We use the symbol $\bar{x}$ to stand for the arithmetic mean of a sample, and we use the Greek letter μ (mu) to stand for the mean of a population. For a finite population we will have: $\mu = \frac{1}{N}\Sigma x$. Often we calculate $\bar{x}$ and use it to try to estimate μ. Obviously, if we have an infinite population it is impossible to calculate μ, and then there is no alternative but to use $\bar{x}$ to estimate μ.

Calculating the arithmetic mean of a frequency distribution

The mean which we discussed above is sometimes called a simple mean, and each value in the set is given the same weight, or importance. In the case of a frequency distribution, the mean must be obtained as a weighted mean.

We will first consider the simpler case of a discrete frequency distribution, and will use as an illustration the data on powered fishing boats per village, which we used in the previous topic. To calculate the mean number of powered boats per village, it is not correct to calculate a simple average of the different numbers of boats (0, 1, 2, etc.) shown in that distribution. There are many more villages with three boats than with one boat, for instance, 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, not much error will be introduced by the way we treat it; in this example we shall assume an average size of 7 boats per village for all units in this last class.

The arithmetic mean in this example is obtained as the sum of the products of the two columns of data, divided by the total number of observations, as follows.

No. of powered boats (x)	No. of villages (f)	Product (fx)
0	20	0
1	7	7
2	12	24
3	28	84
4	17	68
5	10	50
6 or more (Est. = 7)	4	28
Total	98	261

The arithmetic mean number of boats per village is then 261/98 = 2.66. We may note that, if we had gone back to the raw data and calculated the mean of the 98 individual observations we would have obtained almost exactly the same result. The only reason we have to say ´almost´ exactly

is that we do not know the precise values for the 4 villages in the open-ended ´6 or more´ class.

Just as we had a mathematical formula for the arithmetic mean of a set of numbers, so we can derive a similar formula for use with a frequency distribution. In this case we call the number of classes ´k´, the value for the ith class will be denoted by x_i and the frequency of each class by f_i. The formula for the mean $\bar{x}$ is then given by:

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

In words this says that the mean of the distribution is given by the sum of the frequency of each class multiplied by the value of the variable for that class, all divided by the total frequency.

We will abbreviate this formula by omitting reference to i and k, as

$$\bar{x} = \Sigma fx \; / \; \Sigma f$$

In the example above, $\Sigma fx = 261$ and Σf (which is equal to n) = 98. We may note that there are 7 classes in the frequency distribution, so k = 7.

When dealing with a continuous distribution we use the class mark as our values of x, as in the following example using the yellowfin data from the previous topic:

Class (kg)	Class Mark (x)	Frequency (f)	Frequency x Class Mark (fx)
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
Total		$\Sigma f =$ 63	$\Sigma fx =$ 280.35

$$\bar{x} = \Sigma fx / \Sigma f = 280.35/63 = 4.45$$

The arithmetic mean in this case is obtained as the weighted mean of the class marks or midpoints of the class intervals, the weights being the frequencies, f_i, or relative frequencies, $f_i / \Sigma f_i$. What we have done is to assume that all the units in a class interval have the corresponding midpoint as their value.

It follows then that we cannot expect the arithmetic mean we have calculated to be exactly the same as the mean we would obtain by going back to the individual raw data. In fact if we make the calculation of the arithmetic mean from the 63 individual observations of yellowfin weights, we find that $\bar{x} = 278.9/63 = 4.43$.

The arithmetic mean has a lot of advantages as an average: it is easy to calculate, 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 quite affected by one or two large observations, especially in a small sample; this can happen when we have a non-normal distribution. In this kind of situation, using the arithmetic mean may be misleading.

For instance, in distributions of income, which are strongly skewed to the right, it is not unusual for the incomes of a few very rich people to be so large that they pull the arithmetic mean to a higher level than is earned by the great majority of people.

In similar fashion, the arithmetic mean of a bi-modal distribution quite often falls between the two peaks, and is therefore not a good representation of the distribution.

The other difficulty that can arise with the arithmetic mean is that we can obtain a value that obviously does not exist. There is no problem with continuous data: it is easy enough to envisage a mean weight of 4.45 kg, for example. However, for discrete data the situation is different. We calculated the mean number of boats per village as 2.6. Obviously we cannot have 0.6 of a boat, and many non-statisticians find this kind of answer difficult to understand. We could round the answer to the nearest whole number, in this case 3, but we lose a lot of information if we do. What we have to realise is that the arithmetic mean is an artificial concept; we use it because it is convenient, not because it has any natural meaning. If we found that the mean number of boats had been 2.2 in 1980 and is 2.6 in 1984, we could draw some conclusions about trends. We can use the mean to make useful comparisons, but we must not assume that the mean value must actually exist.

4.3.2 The median

The median is a very simple concept which can be quite useful in practice, although it is difficult to deal with mathematically. It is the value of the middle observation of a set of numbers; half the numbers will be larger than this value and half will be less. For data in the raw form all we have to do is to rank the observations in order of size, and the median will be the value of the middle one. If we have n observations, the median will be the value of the $\frac{n+1}{2}$ th observation.

If n is odd there is no problem, but if n is even there are two middle observations; in this case we take the median to be the arithmetic mean of the two values. As an example, consider the following three sets of observations, which have been sorted into size order.

(a) 19, 22, 26, 31, 34, 37, 42, 44, 49, 55, 63

(b) 12, 19, 23, 27, 30, 30, 47, 49, 60, 93

(c) 128, 186, 193, 207, 218, 222, 286, 346

In set (a) there are 11 observations; the median is given by the 6th one, so is equal to 37. In set (b) there are 10 observations; the median is given by the mean of the 5th and 6th ones, and as each of these is equal to 30, there is no problem in obtaining 30 as the median value. In set (c) there are 8 observations so the median value is the arithmetic mean of the 4th and 5th observations, i.e. the median = (207+218)/2 = 212.5.

To calculate the median from a frequency distribution we use the same principle, but we start by determining the class within which the median value lies. If this class contains a single value, then there is no problem. If, however, it contains a range of values, then we have to estimate the median value, using simple interpolation. For the second case, because we are dealing with a range of values, we use the formula n/2, not (n+1)/2 to determine the median observation. This is illustrated in Table 4.2.

TABLE 4.2 : CALCULATION OF THE MEDIAN FOR A FREQUENCY DISTRIBUTION

(I) Using our powered fishing boats example:

No. of boats	Frequency	Cumulative frequency
0	20	20
1	7	27
2	12	39
3	28	67
4	17	84
5	10	94
6 or more	4	98

The median value is given by the (98+1)/2th observation, i.e. by the mean of the 49th and 50th. These lie in the class "3 boats per village", so the median is 3.

(II) Using our yellowfin weights example:

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

In a frequency distribution, the median position is n/2, not (n+1)/2 as in the case of an array. With 63 observations, the median position is the 63/2th or 31.5th. From the cumulative frequency column we see that the 31.5th position falls in the class 4.0 - 4.9 kg with actual class limits 3.95 - 4.95. There are 26 observations prior to the interval beginning 3.95, and 16 in this interval, so we calculate the median weight as 3.95+(31.5-26)/16x1.0. Thus the median is equal to 4.3 kg.

Once again we must recognise that (as for the arithmetic mean) the calculation of the median which we obtain from a frequency distribution is only approximately the same as we obtain from a list of all the individual observations. Indeed if we refer back to our original 63 yellowfin weights and sort them into order, we find that the (63+1)/2th (or 32nd) value is 4.1 kg.

Another way to determine the median is directly from the ogive of the frequency distribution. In this case, though, the accuracy of the median is determined by the accuracy with which the graph is plotted. We draw a horizontal line from the y-axis at the position corresponding to the median observation, and from the point where this line cuts the ogive, we draw a vertical line. The point where this vertical line meets the x-axis gives us our reading of the median value.

If the ogive of our example is accurately plotted, as in Figure 4.1 below, we should be able to observe that the median (i.e. the 31.5th) value is 4.3 kg, which is the figure we have already calculated from the frequency distribution.

4.3.3 Quartiles

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. The most important of these are called the quartiles. Quartiles, as their name suggests, divide the total frequency into four equal parts.

The first, or lower, quartile will then have one quarter of the observations less than this point and three quarters greater. The middle quartile is equivalent to the median. The upper quartile has three quarters of the observations less than this value and one quarter more.

In an array, the lower quartile is the (n+1)/4th value, the middle quartile (which corresponds with the median) is the (n+1)/2th value, and the upper quartile is the 3(n+1)/4th value. For the 63 observations of yellowfin weights we can see that the lower quartile is the 16th value and the upper quartile the 48th. A study of the individual weights, sorted into order, will show that the quartile values are 3.3 kg and 5.4 kg respectively.

Whenever the quartile lies between two values, the value of the quartile is calculated by interpolation as in the case of the median.

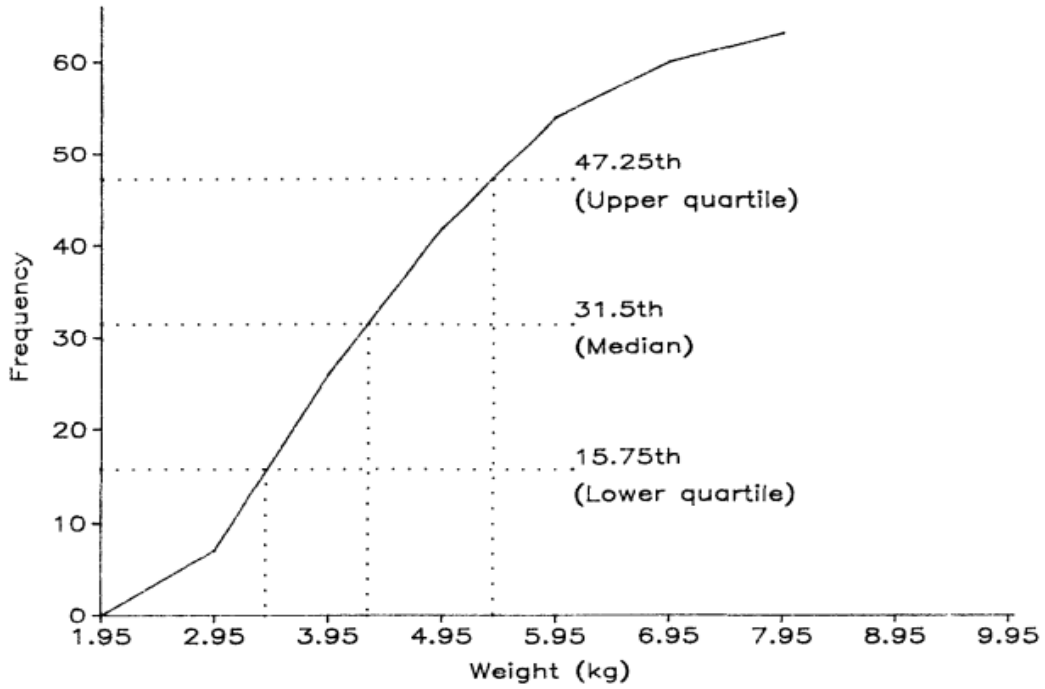
A similar method is used even for a frequency distribution, interpolating (as in the case of the median) within the quartile class. However, the quartile positions become n/4th, n/2th and 3n/4th. In our example Q_L is the 63/4th or 15.75th position. This falls in the 3.0 - 3.9 class, being the 15.75-7 or 8.75th item into this class. Therefore $Q_L = 2.95+(15.75-7)/19 = 3.41$, or 3.4. A similar calculation will give us a value for the upper quartile of 5.4 kg. These values can be compared with those obtained from the raw data, as shown above.

Later in this topic we will use the quartiles to determine a measure of the spread or dispersion among the observations.

Apart from splitting the total frequency into quarters we could use other fractions, or percentages, and we associate the word percentiles with these. There are also special names for certain commonly used fractions, e.g. deciles (which split the total frequency into 10 groups) and quintiles (into 5 groups). We should note here that these are not averages, since they do not give us a central value to represent the distribution, but it is convenient to discuss them at the same time as we discuss the median, which is an average. All of these measures can be read

directly from an ogive, and we illustrated that in Figure 3.6. The derivation of the median and quartiles for the yellowfin data is shown in Figure 4.1.

FIGURE 4.1 : CUMULATIVE FREQUENCY OF YELLOWFIN WEIGHTS SHOWING POSITION OF THE QUARTILES



The median and other percentiles are very useful when we want to describe what is happening in certain types of distribution. Quite often, for example, we would like to find out about part of a distribution, the poorest 20 per cent of fishermen, the largest 20 per cent of skipjack, and so on. By calculating percentiles we can do this, and we can compare values between distributions.

We also find that for skewed distributions (ones that are not almost symmetrical), the median is often a better measure of the average value than the mean. Figure 4.2 illustrates this; it shows the position of the mean and the median for a symmetrical and a skewed distribution. For a distribution that is almost symmetrical, the value of the median and the mean are very similar, and both are good measure of the average. For a distribution skewed to the right, however, the mean will be to the right of the median. In the sense of actually representing the data the median may be more useful; it is more stable, is not affected by the inclusion of a few very large values, and hence is probably better to use for the purposes of comparison.

4.3.4 The mode

If a population distribution has a peak in its distribution function at a certain point then there is said to be a mode at that point. Like the arithmetic mean and the median, the mode is a type of average.

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 which occurs most often. For instance, in our earlier illustration of the number of powered boats, the modal value is 3 boats per village, because more villages had 3 boats than any other number. We may consider that this is a more useful summary of our information than is the arithmetic mean of 2.66 boats. It is interesting to note that in this case the median value is also 3.

For a continuous distribution the determination of the mode is rather complicated, and so for our purposes we shall be concerned only with the modal group or class. This is the class with the highest frequency (that is, 3.0 to 3.9 kg in the yellowfin data, for example), in a distribution which has equal class intervals.

However, determining the modal group for continuous frequency distributions, particularly where the distribution is a sample from a population, often produces problems. The modal group will very largely depend 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 group. The smaller the sample, the more likely this is to happen. For instance, if we had grouped the yellowfin data into 3 classes, namely, 2.0-3.9 kg, 4.0-5.9 kg and 6.0-7.9 kg, we would find that the class with the highest frequency is 4.0-5.9 kg, so we would have quite a different modal class. Various different groupings, e.g. into a large number of smaller classes, would give us different results again. For this reason, the mode is of limited value, and should be used with care.

Diagrams illustrating the relationship between the arithmetic mean, the median and the mode for the most common distributions are given in Figure 4.2.

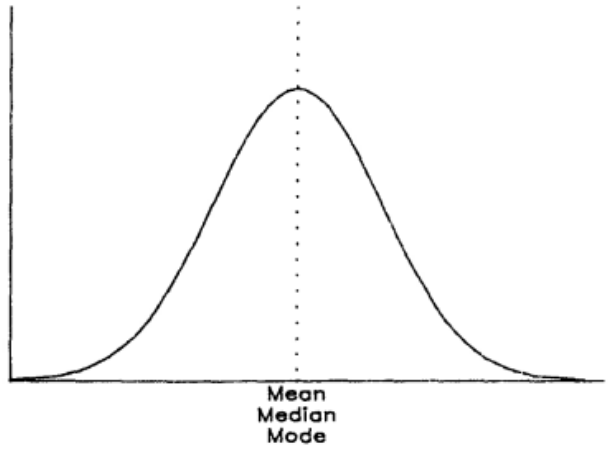
4.3.5 Summary of the different types of average

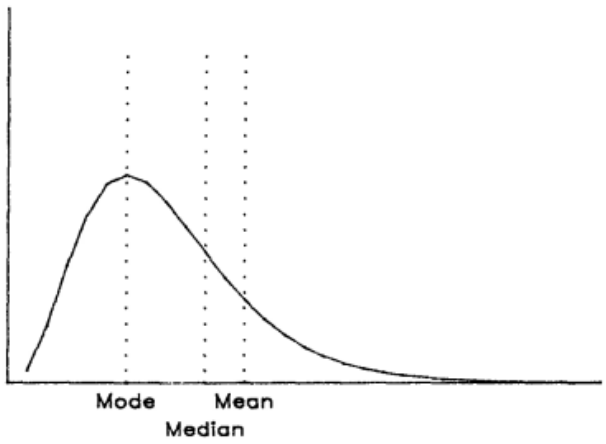
We have looked at three main types of average: the arithmetic mean, the median and the mode. All three of these have both advantages and disadvantages when used to describe or summarise a set of data. The mean is the most widely known and most widely used, but can be misleading when dealing with skewed distributions. In this situation the median, and various percentiles can be more useful, particularly when making comparisons between distributions. The mode has limited value, and should not be used with small samples.

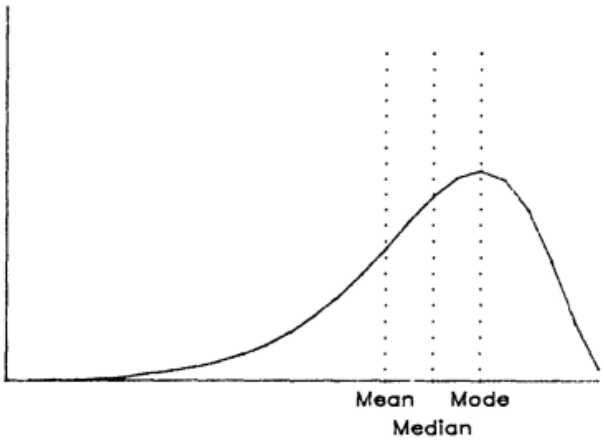
When we come to the problems of statistical inference, however, we almost always use the arithmetic mean. The reason for this is purely mathematical convenience. It is much easier to deal with the mean to derive more complex results; the ways the median and the mode are defined make these much more difficult to use. We tend, therefore, to concentrate on the mean just because this helps us when we want to study more complicated areas of statistics.

There are also other types of average, which we will not discuss in this course. The best known are the geometric mean (which is most useful for measuring rates of change) and the harmonic mean.

FIGURE 4.2 : THE RELATIONSHIP BETWEEN THE ARITHMETIC MEAN, THE MEDIAN AND THE MODE







4.4 Measures of dispersion

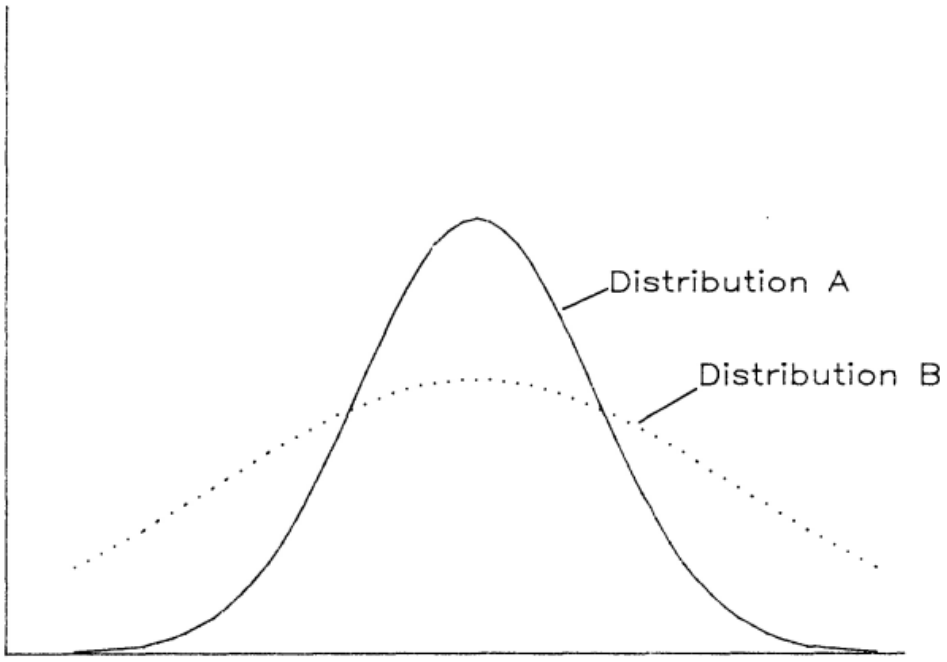
4.4.1 Basic principles

In the two previous sections we have discussed ways in which we can summarise statistical data, and can present it in a straightforward way which will be fairly readily understood. The frequency distribution is a method to summarise information in tabular or graphical form, while an average (such as the arithmetic mean) summarises this information into one single number.

We must recognise, however, that while in summarising we are attempting to distil from a mass of data the essential features which need to be highlighted, in so doing we are always losing some of the information. We have to be very careful that in the process we do not go too far, and leave out of our summary information which is necessary for a proper understanding of the situation. We will see in this section that an average, on its own, is often insufficient to describe a population adequately. In particular when we are endeavouring to compare the characteristics of different populations, some further measure in addition to the average is usually required.

Consider, for example, Figure 4.3 where there are two distributions shown. Both have the same average value, whether measured as a mean, a median or a mode, but 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. From the diagram we can see that distribution B is much more spread out than distribution A; in this section we shall look at different ways of measuring this spread, or dispersion.

FIGURE 4.3 : COMPARISON OF TWO DISTRIBUTIONS



We want to measure dispersion for two main reasons. In the first place we may well be interested in the actual level of dispersion and in comparing this with another distribution. The second reason for wanting to measure dispersion is that, even when we only want to compare average values, we still need to take variability into account. We want to be able to distinguish between differences that might have just happened by chance and those that indicate some real change.

In this topic we shall consider four different measures of dispersion, which are basically of two types:

(a) measures of the distance between certain representative values of the population; and

(b) measures of the deviation of every member of the population from some specified central value.

As examples of the first type of measure of dispersion we shall look at the range and the interquartile range, while under the second type we shall consider the mean deviation and the standard deviation (or the square of this, the variance).

4.4.2 Measure of the distance between selected points of the distribution

The most obvious way of measuring the dispersion in a set of observations is to calculate the range, which is just the difference between the smallest and the largest values. This is 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 type of distribution over the range. It has some obvious drawbacks which tend to restrict its use in practice; some of the more important disadvantages are:

(a) Because the range is the difference between the smallest and the largest values, it is very sensitive to very large or very small observations; the inclusion of just one freak value will affect the range.

(b) The range depends 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) The range ignores most of the observations; for example, the following sets of data all have the same range, even though we can see that the degree of dispersion is different.

(i) 3, 5, 7, 9, 11, 13, 15, 17
(ii) 3, 3, 3, 3, 17, 17, 17, 17
(iii) 3, 3, 3, 3, 3, 3, 3, 17

(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 will obviously depend on our definitions of the classes, and is impossible if we have an open-ended class.

We can get round most of the disadvantages of the range as a measure of dispersion by using other points in the distribution rather than the two extremes. An obvious choice would be to measure the inter-quartile range - the difference between the upper and lower quartiles. Another alternative would be to use the difference between the 10th and the 90th percentile. As measures of dispersion, both these are quite useful. They are not affected by one or two wild 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 frequency distributions, we can nearly always calculate these distances, the only problem being when one of the percentiles or quartiles falls in an open-ended class.

The inter-quartile range in particular is a fairly good measure of dispersion, that is reasonably easy to calculate and which most people find fairly simple to understand. It can be used to measure the amount of dispersion and to make simple comparisons between distributions. Quartiles are far enough from the ends of the distribution to make it extremely unlikely that they will fall within an open-ended class. In fact, if there is an open-ended class in a distribution, the inter-quartile range will probably be the only one of our four measures of dispersion which we can calculate accurately. All the other measures will require some assumption to be made about the open-ended class.

In practice, the quartile deviation is often quoted; this is defined as one half of the inter-quartile range and provides a result that is comparable with other measures of dispersion. The major drawback, however, comes when we want to undertake more advanced statistical work. It is difficult to deal mathematically with quartiles, so in practice we tend to concentrate on other measures of dispersion.

4.4.3 Measures of deviation from a specified central value

With this type of measure of dispersion we use every value in the distribution and find the average distance between every observation and some central point. In theory, we could use any central point we like, the median, the mode, or whatever, but in practice we use the arithmetic mean for reasons of mathematical convenience. What we need to do, then, is to find the difference between each observation and the mean, and then calculate the average of these distances. There is, however, one immediate problem which we can illustrate with the following simple set of data:

3, 5, 7, 9, 11, 13

The arithmetic mean is (3+5+7+9+11+13)/6=8 and the differences, or dispersions, of each observation from the mean are:

-5, -3, -1, +1, +3, +5.

The total dispersion is zero, and in fact this will always be true. Because of the way we define the mean, the total dispersion of all the observations from that value will always be zero; it is a check on the accuracy of our calculations. We cannot, therefore, use the values exactly as they are.

What we are interested in, in fact, is the actual size of the dispersion, regardless of the sign, and so one possible solution would be to take the average value disregarding all signs. In our simple example, then, our total dispersion would be:

5+3+1+1+3+5=18, and, since we have 6 observations,
the average will be 18/6=3.

This is a good measure of dispersion, and we call it the mean deviation; it is the average deviation of all observations from the mean, disregarding all signs. For a general sample, $x_1\ x_2\ \ldots x_n$ we can write the formula for mean deviation as

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \qquad \text{or, more simply} \qquad \frac{1}{n}\Sigma\,|x - \bar{x}|$$

The symbol $|\ldots|$ stands for modulus, or mod for short, and it means - take the absolute value, ignore all signs.

Although the mean deviation is a good measure of dispersion, and one that most people find quite easy to understand, we do not use it much. The reason for this is that it is difficult to manipulate the modulus of a number mathematically, which means that the mean deviation cannot be easily used in more advanced statistical work.

4.4.4 Standard deviation

Instead, we calculate the standard deviation, which we obtain as follows:

As before, we work with the deviations from the arithmetic mean; in our example we had:

-5, -3, -1, +1, +3, +5.

In this case we square these deviations, which will give us the following:

25, 9, 1, 1, 9, 25.

All these numbers are positive. We now take the average of these squares, i.e.

(25+9+1+1+9+25)/6 = 70/6 = 11.67

Since we have squared all the deviations, we should return to the magnitude of the original units, and so we take the square root of this result, i.e.

standard deviation = $\sqrt{11.67} = 3.4$

If we are concerned with a population, we use the symbol σ (sigma) to stand for the standard deviation, and in general terms σ is given by:

$$\sigma = \sqrt{\sum_{i=1}^{N}\frac{(x_i - \mu)^2}{N}} \qquad \text{which we will simplify to:} \quad \sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

where μ, of course, is the mean of the population.

For a sample, the situation is a little different; we can use the symbol s to stand for the sample standard deviation and this is given by:

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

Here, we use the divisor n-1, where for a population we use N. If the sample size is large, this will make little difference. The reason that we use n-1 is that, in this case, s provides an unbiased estimate of σ. In other words, if we took many different samples and calculated s for each one, the mean of these values would approach the population value. This would not be true if we used the divisor n.

The square of the standard deviation is called the variance, and we sometimes use this to avoid having to take square roots. The population variance is denoted by σ^2, and equals $\frac{\Sigma(x-\mu)^2}{N}$, while the sample variance, s^2 equals $\frac{\Sigma(x-\bar{x})^2}{n-1}$.

As it stands, it is quite a cumbersome procedure to calculate the standard deviation of 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, add them up and take the square root of the result. Even with a calculator this will require each value to be entered twice, and can take some time.

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

$$S = \sqrt{\frac{1}{n-1}\left(\Sigma x^2 - \frac{1}{n}(\Sigma x)^2\right)}$$

and for a population

$$\sigma = \sqrt{\frac{1}{N}\left(\Sigma x^2 - \frac{1}{N}(\Sigma x)^2\right)}$$

Although this second formula looks more complicated than the first, it is in fact much easier to use with a calculator. We can observe that in this formula we do not have to start by calculating the arithmetic mean, so we can save one step in the calculation process. Using the memory function on the calculator, we can now calculate the standard deviation without having to write down any intermediate results.

We can use the second version of the formula to give us a fairly simple method for calculating the standard deviation of a frequency distribution. We shall use as an example the yellowfin data to illustrate this. The relevant calculations are as follows:

Weights (kg)	Class Mark (x)	Frequency (f)	fx	fx^2
2.0 - 2.9	2.45	7	17.15	42.02
3.0 - 3.9	3.45	19	65.55	226.15
4.0 - 4.9	4.45	16	71.20	316.84
5.0 - 5.9	5.45	12	65.40	356.43
6.0 - 6.9	6.45	6	38.70	249.61
7.0 - 7.9	7.45	3	22.35	166.51
Total		63	280.35	1357.56

The standard deviation, s

$$= \sqrt{\frac{1}{\Sigma f-1}\left(\Sigma fx^2 - \frac{1}{\Sigma f}(\Sigma fx)^2\right)}$$

$$= \sqrt{\frac{1}{62}(1357.56 - 1247.56)}$$

$$= 1.33$$

We use $\Sigma f-1$ which is another way of writing n-1, as the denominator, because the data are from a sample.

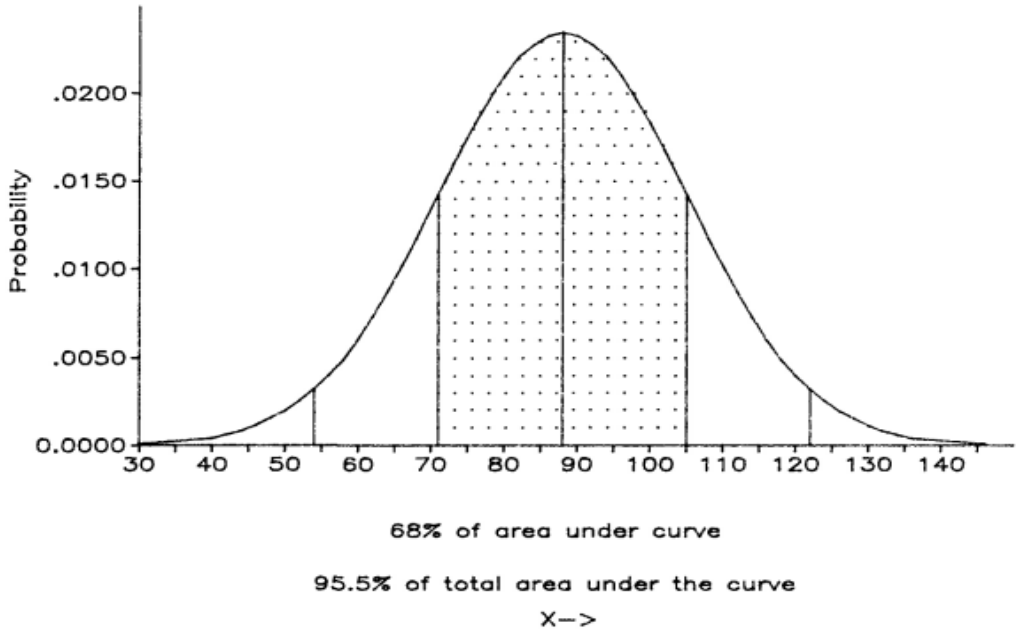
If we wanted to write out the formula for the variance of a frequency distribution in full, it would be:

$$s^2 = \frac{1}{\sum_{i=1}^{k} f_i - 1} \left(\sum_{i=1}^{k} f_i x_i^2 - \frac{1}{\sum_{i=1}^{k} f_i} \left(\sum_{i=1}^{k} f_i x_i \right)^2 \right)$$

in the case of a sample.

One interesting feature of the standard deviation in respect to the normal distribution may be mentioned here. If the population is distributed normally about the mean, then approximately 68 per cent of all values will lie within one standard deviation of the mean, and about 95.5 per cent will lie within 2σ. Thus, for a normal distribution with a mean of 88 and a standard deviation of 17, about 68 per cent of all values will lie in the range 88-17 to 88+17 (i.e. between 71 and 105) and 95.5 per cent will be within the range 54 to 122. We can demonstrate this graphically, as shown in Figure 4.4, by plotting the normal curve, and vertical lines drawn from the x-axis at values 71 and 105 would enclose 68 per cent of the total area under the curve. This particular property will hold true for a normal distribution, no matter how widely spread the values are. It will be very useful in our understanding of standard errors, which will be discussed later in the course.

FIGURE 4.4 : NORMAL PROBABILITY DISTRIBUTION, MEAN 88, S.D. 17



The standard deviation is by far the most widely used of the four measures of dispersion. As we will see it is also used in the calculation of sampling errors. Although it is so widely used, this does not mean that it is superior in every respect. Its main weakness is that it is very

greatly affected by extreme values, much more so than is the mean deviation. This occurs because the deviations from the mean (which are already large in the case of extreme values) become very large indeed when they are squared, as they are in the calculation of the standard deviation.

4.4.5 Summary of the different measures of dispersion

There are two ways we can measure the degree of dispersion in a set of observations: we can look either at the difference between two points in the distribution or at the average deviation of all the observations from some central point. Examples of the first type are the range and the quartile deviation. These are fairly easy to calculate, have an obvious meaning, but ignore quite a large part of the data. The range, in particular, is unstable and is affected by wild observations; the quartile deviation is more stable and better to use in practice. Both measures are difficult to deal with mathematically. Examples of the second type of measure are the mean deviation and the standard deviation. In practice we use the standard deviation because it is, mathematically, more convenient.

Average value and dispersion are not the only properties of distributions we can measure, but we generally concentrate on just these two. The reason for this is that, for a large class of symmetrical or almost symmetrical distributions with a single mode, we can fit a normal distribution quite easily. This will allow us to make many important inferences concerning the data. One very important property of a normal distribution is that the only things we need to know are the mean and the variance (or standard deviation). Once these are determined, the distribution is fixed. So, to fit a normal distribution to a set of data, all we have to do is to calculate its mean and variance.