FREQUENCY DISTRIBUTIONS : REDUCING A LOT OF DATA INTO A MANAGEABLE FORM

3.1 Background and Introduction

When we undertake a statistical investigation we end up with a series of observations of some characteristic for a number of units. Usually, of course, we will have a number of observations of different variables and attributes for each unit, but to keep the situation fairly simple at present we will only look at one variable. Given, then, the series of observations, we want to find out some way we can summarise this information, so that we may begin to make sense out of it. We may want to make some kind of decision, or inference, about the whole group of units, perhaps to compare them with some other group; we may want to make some kind of estimate for the whole population of units of which our group may only be a small part; or we may just want to have some convenient way to summarise the basic data, to reduce the amount of information to a manageable size.

We shall start off by looking at a frequency distribution as a method of summarising the basic data; later on in this topic we shall see how we can develop a theoretical basis. As an example, consider the information given in Table 3.1. This represents a summary of 1,870 observations of the weights of mangrove crabs packed for export. We prepare a frequency distribution by dividing the range of weights we observe into a number of classes and then counting the number of observations in each class.

TABLE 3.1 : EXAMPLE OF A FREQUENCY DISTRIBUTION (NUMBER OF MANGROVE CRABS PACKED FOR EXPORT ACCORDING TO WEIGHT)

Weights of Crabs (g)	Number of Crabs
200 to less than 300	55
300 to less than 400	302
400 to less than 500	540
500 to less than 600	357
600 to less than 800	290
800 to less than 1,000	176
1,000 to less than 1,200	59
1,200 to less than 1,600	52
1,600 and over	39
Total	1,870

This frequency distribution has summarised 1,870 observations into 9 groups or classes, together with a total figure. Obviously this is much easier to comprehend than a list of 1,870 individual values would be, even if those values were sorted into size order. At the same time we have lost some of the original detail; we do not know the actual value of any of the observations.

A typical frequency distribution then, divides the range of values of the characteristic we are considering into different classes and counts the number, or the frequency, of observations within each class. We can construct frequency distributions for both variables and attributes, but procedures for attributes are quite straightforward so in this topic we will concentrate only on variables. Later on we shall distinguish between two basic types of variables.

A frequency distribution is particularly useful if we wish to find out how the values of a variable are distributed. It shows at a glance the range of values, how many high values and how many low values have been observed, what the most frequently occurring values are, and whether the values are symmetrically distributed along the range, or whether they are mostly at one end.

3.2 Construction of a frequency distribution

Preparing a frequency distribution from a list of the basic data consists of three steps:

(a) specifying the classes into which the data are to be grouped;

(b) sorting the data into these classes; and

(c) counting the number of observations in each class.

The last two of these steps are quite straightforward, but it can be quite difficult to decide on the number of classes we need, and on the range of values for each class. In Table 3.1, on crab weights, we chose nine classes, and the class interval (or range) was 100 g for the first four classes, 200 g for the next three, 400 g for the next one, and unlimited for the last class (i.e. it was open-ended, and any observation of 1,600 g or more would have been included in it).

Although in principle we can decide on any set of classes we like for a distribution, and the definition of each class will depend on the purpose of the distribution, there are some guidelines which it is useful to follow:

(a) The classes chosen must span a range sufficient to encompass every observation, from the lowest to the highest.

(b) There should be no gap or overlap in the classes; each should be separate and distinct. It is particularly important that the range of each class should be defined so that each observation can only go into one class. If in our previous example we had carelessly described the classes as 200-300, 300-400, 400-500 and so on, we would not know how to classify a crab weighing exactly 400 gms, as it could go into either of two classes. We must be sure that there is no ambiguity, and that observations which are on the border between two classes will fit into only one of them.

(c) There should not be too many classes (as this will lose the advantage of a frequency distribution over the raw data), nor too few (as too much information will be lost). In general, it is suggested that more than 5, and not more than 16, separate classes are desirable, but these are no firm limits. The number of classes formed will depend on the nature of the data, on the number of observations, and on the type of distribution.

(d) It is advantageous, whenever practicable, for the class intervals, that is, the range (or length) of each class, to be the same. Class intervals of equal length make it much easier to comprehend the distribution and to draw suitable diagrams. If unequal intervals are used it is often difficult to compare one class frequency with another. Sometimes, however, it is impossible to avoid unequal intervals; the variability of the data requires their use.

(e) Whenever possible avoid the use of <u>open-ended</u> <u>intervals</u>, that is classes at the ends of the distribution in which one end value of the class is not stated.

As we have already observed, our frequency distribution of crab weights does not have equal class intervals, as we have suggested in (d) above. We could try to combine data into classes with equal intervals of, say 400 g, and the frequency distribution would then become as shown in Table 3.2.

TABLE 3.2 : NUMBER OF MANGROVE CRABS, ARRANGED IN CLASSES WITH EQUAL INTERVALS

Weights of Crabs (g)	Number of Crabs
Less than 400	357
400 and less than 800	1,187
800 and less than 1,200	237
1,200 and less than 1,600	52
1,600 and over	39
Total	1,870

It will be seen that this distribution does not contain as much useful information as the first table, as one class now contains nearly two-thirds of all cases (1,187 out of 1,870). The breakdown of this class into smaller classes, as we had originally shown, is desirable to give additional information on the size of crabs in this large class. Generally it will be found that the more evenly the observations are spread, the easier it will be to construct a frequency distribution with equal class intervals.

It will also be noted that the final class of our distribution is open-ended, despite the recommendations of principle (e) above. However, it is often quite difficult, or even virtually impossible to avoid their use. For example, there may have been one or two very large crabs, of say 3,000 g, and we would have had to break down the final class into several classes, with different intervals, in order to avoid the open-ended class we have used. Since there are only two per cent of all observations in this class, a further breakdown would not have been desirable.

We will now use a different set of data to illustrate how a frequency distribution is constructed. The data represent the weight in kilograms of a sample of 63 yellowfin tuna caught by pole-and-line method (Table 3.3). For this exercise it is assumed that all weights are rounded to the nearest 1/10 kg.

TABLE 3.3 : WEIGHTS OF 63 YELLOWFIN (in kilograms)

4.6	3.9	2.8	6.6	4.2	3.7	3.7	5.9
3.2	2.2	3.2	4.1	3.1	3.0	4.8	4.1
2.1	4.2	5.0	4.6	5.4	2.4	6.3	2.9
5.3	4.0	4.7	3.6	3.3	6.9	4.5	2.5
5.4	5.7	3.8	4.1	5.6	6.2	3.0	3.3
5.0	5.4	3.4	4.4	4.0	3.6	5.0	4.1
4.8	7.2	6.4	3.0	3.5	5.8	7.7	3.9
2.6	7.9	3.3	5.5	4.3	3.9	6.3	

To determine the different classes we first of all need to know the range, from the lowest to the highest value. In this case the lowest value is 2.1 kg, the highest 7.9 kg, the range therefore is 5.8 kg. We wish to split the range up into a number of classes, and the observations are so evenly spread that there seems no reason with this data why we should choose classes with unequal intervals. A suitable distribution, then, is given in Table 3.4.

TABLE 3.4 : FREQUENCY DISTRIBUTION OF YELLOWFIN WEIGHT DATA

Weights (kg)	Frequencies
2.0 - 2.9	7
3.0 - 3.9	19
4.0 - 4.9	16
5.0 - 5.9	12
6.0 - 6.9	6
7.0 - 7.9	3
Total	63

3.3 Definition of terms

We need to define some terms when talking about frequency distributions. Some of these we have used already; in this section we shall define the terms more closely.

(a) Class frequency

The class frequency in a distribution gives the number of observations falling within that particular class. When presenting a frequency distribution in tabular form, the classes always go in the left hand column, with the class frequencies on the right.

(b) Class limits

The smallest and largest values (rounded where necessary) that can go into any given class are termed its class limits. In the yellowfin weights table the class limits are 2.0, 2.9, 3.0, 3.9, and so on. We usually differentiate between the lower class limits (2.0, 3.0, 4.0, etc.) and the upper class limits (2.9, 3.9, 4.9, etc.).

(c) Class boundaries

These represent the actual, or true limits to a class. There is a fine distinction between class boundaries and class limits, and it is important to be clear on this distinction. In our example we may note that a fish weighing (say) 2.96 kg will be recorded in the survey as weighing 3.0 kg. The class boundaries in this example are actually 1.95, 2.95, 3.95, and so on.

(d) Class marks

The class mark is the mid-point of the class, and is obtained by taking the arithmetic mean of the upper and lower class limits. In the example, the class marks are 2.45, 3.45, 4.45, etc. These are often also referred to as mid-marks, mid-points, mid-values, etc.

(e) Class interval, or range

The class interval is the length of any class, the range of values it contains. The class interval of a class is the difference between the lower class limit of that class and the lower class limit of the next class. If all the intervals are equal then it is also equal to the difference between successive class marks. For example, the class interval for the yellowfin weights is 1.0 kg and is equal for all classes. Note that the class interval is not necessarily the difference between the upper and lower limits of the class. (In our table this is equal to 0.9 kg.)

3.4 Continuous and discrete data

At this point in the study of frequency distributions, we need to distinguish between two different types of variable, because the problems of constructing a frequency distribution and drawing diagrams of the distributions are somewhat different for each type. The first is where the variable is allowed to take any value within a specified range, and the second where the variable can only take certain values. The first type we call a continuous variable and we also refer to continuous data; the second type we call discrete. Examples of continuous variables are:

(a) fork length of fish;
(b) weight of fish;
(c) water temperature of sea surface.

Examples of discrete variables are:

(a) number of canoes in a village;
(b) number of longline sets;
(c) crew number.

In most cases discrete variables take whole number (or integer) values, although this is not essential.

In practice, the dividing line between continuous and discrete data is often very difficult to discern. Continuous data will not normally be recorded in an absolutely continuous way, as there are limitations to the accuracy with which we can measure or record a variable. For example, we may be able to measure the weight of a fish to the nearest 1/10 kg - or even to the nearest gram if we had accurate enough equipment - but that is as fine a breakdown as we could hope to achieve.

Likewise some discrete data can be dissected so finely that it looks like continuous data. Revenue from fish sales could be recorded to the nearest dollar, or even to the nearest cent, and we would have such a spread of recordings that we could treat this as though it were continuous. There is a fine distinction, however. Fish weights could take any value whatever in a range, and it would still be meaningful; revenue on the other hand cannot be expressed in fractions of a cent, because a cent is the smallest unit of currency which exists.

While quite often the division between continuous and discrete data is blurred, there are situations where the distinction is important. For instance, if we are discussing the number of fishing units in a village, or the number of landing sites on an island, we will have a discrete distribution which does not look at all like a continuous distribution.

We can illustrate this by looking at two examples of constructing frequency distributions from discrete data. The situation is simple as long as we have one value of the variable only in one class. For example, we can quite simply present data summarising the number of villages with 0, 1, 2, 3, 4, etc. powered fishing boats as in Table 3.5.

TABLE 3.5 : NUMBER OF VILLAGES, BY NUMBER OF POWERED FISHING BOATS PER VILLAGE

Number of powered boats	Number of villages
0	20
1	7
2	12
3	28
4	17
5	10
6 or more	4
Total	98

Often, however, the range of values is so great that we have to combine values in each class. For example, a country may be interested in the distribution of visits by longline vessels to its territorial waters, classified by number of longline sets made during a visit. The distribution might be as shown in Table 3.6.

TABLE 3.6 : NUMBER OF VISITS BY LONGLINE VESSELS, CLASSIFIED BY NUMBER OF SETS PER VISIT

Number of longline sets	Number of visits
1 - 5	17
6 - 10	25
11 - 15	23
16 - 20	39
21 - 25	49
26 - 30	33
31 - 35	15
36 - 40	4
Over 40	5
Total	210

Although this table looks like the table for the distribution of weights of yellowfin, we must be careful to remember that the data are discrete. We cannot talk about a vessel making 18.3 longline sets.

3.5 Cumulative frequency distributions

A frequency distribution provides information on the number of observations in the different classes; we can tell at a glance from the table how many small-sized and large-sized observations there are. Often, however, we have a situation where slightly different information is required. What is of interest in this case is to find out how many observations are larger than some specified value, or how many observations are less than a certain amount. For example, in our earlier exercise on crab weights, we may well be interested in how many crabs weighed less than 500 g, how many weighed 800 g or more, and so on. This type of information can be readily obtained from a cumulative frequency distribution.

Table 3.7 shows how we can construct a cumulative frequency distribution for the data on longline sets per visit.

TABLE 3.7 : EXAMPLE OF A CUMULATIVE FREQUENCY DISTRIBUTION

No. of longline sets per visit	No. of visits by longline vessels	Cumulative frequency (less than)	Cumulative frequency (greater than)
1 - 5	17	17	210
6 - 10	25	42	193
11 - 15	23	65	168
16 - 20	39	104	145
21 - 25	49	153	106
26 - 30	33	186	57
31 - 35	15	201	24
36 - 40	4	205	9
Over 40	5	210	5
Total	210		

The cumulative frequency distribution is obtained by calculating the progressive totals of the frequencies in each class. This can be done in one of two ways as illustrated in Table 3.7. The cumulative frequency distribution (less than) is calculated by starting with the first class, and then adding the cumulative frequency in each class until the last. The cumulative frequency distribution (greater than) is calculated by starting with the last class and working upwards.

As their name suggests, the two cumulative distributions are used to answer the questions: how many observations are greater than a certain value? Or, how many observations are less than a certain value? From the table we can see at a glance that 65 of the vessels made 15 sets or less during a visit to the country´s territorial waters, and 106 made over 20 sets in a visit.

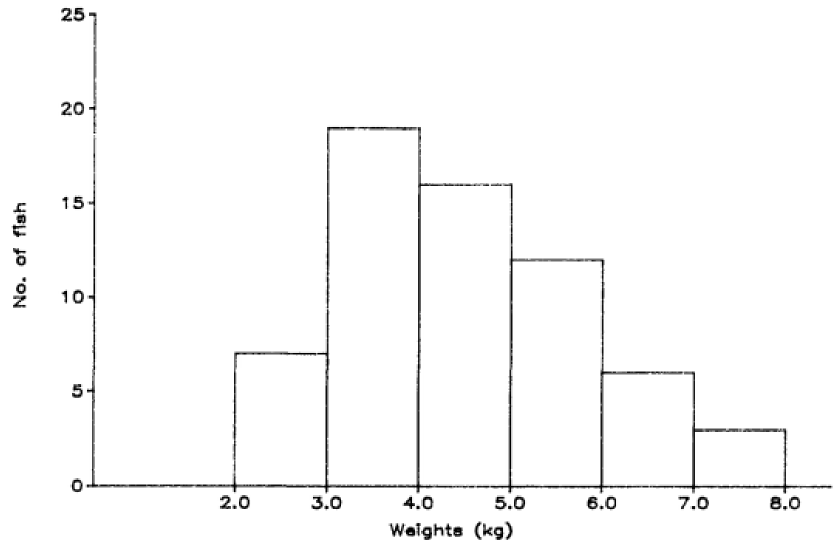
3.6 Diagrams of frequency distributions

A frequency distribution gives information on the way that a number of observations of a particular characteristic are "distributed" or spread out over a range of values. As well as preparing tables of these distributions, it is also important to have methods of representating them graphically, since in this way the different patterns in the data can be seen at a glance.

(a) Frequency histogram

The most common method of representing a frequency distribution is by drawing a frequency histogram. A histogram for our earlier data on yellowfin weights would be drawn as shown in Figure 3.1.

FIGURE 3.1 : FREQUENCY HISTOGRAM OF WEIGHTS OF 63 YELLOWFIN



To draw a frequency histogram we observe the following general principles:

(a) Magnitudes are drawn along the horizontal axis.

(b) Frequencies are plotted along the vertical axis.

(c) The general practice of all diagrams should be followed. Where applicable, headings, footnotes and full details of both axes should be provided.

(d) Frequencies should only be represented by a rectangle covering the whole of an interval if the data is continuous; discrete histograms are drawn somewhat differently, as will be shown a little later in this section.

(e) When plotting a continuous distribution it is the area of each rectangle and not the height that is proportional to the frequency. It is only in the case of equal intervals, as in the previous diagram, that the frequency is proportional to the height.

It is worth noting here that we have shown the weights of the classes of yellowfin in the previous diagram as 2.0, 3.0, 4.0, etc. These values represent the lower limit of each class. At the same time, we must remember that all the weights have been rounded to the nearest 0.1 of a kilogram and that the true class limits are 1.95, 2.95, etc.

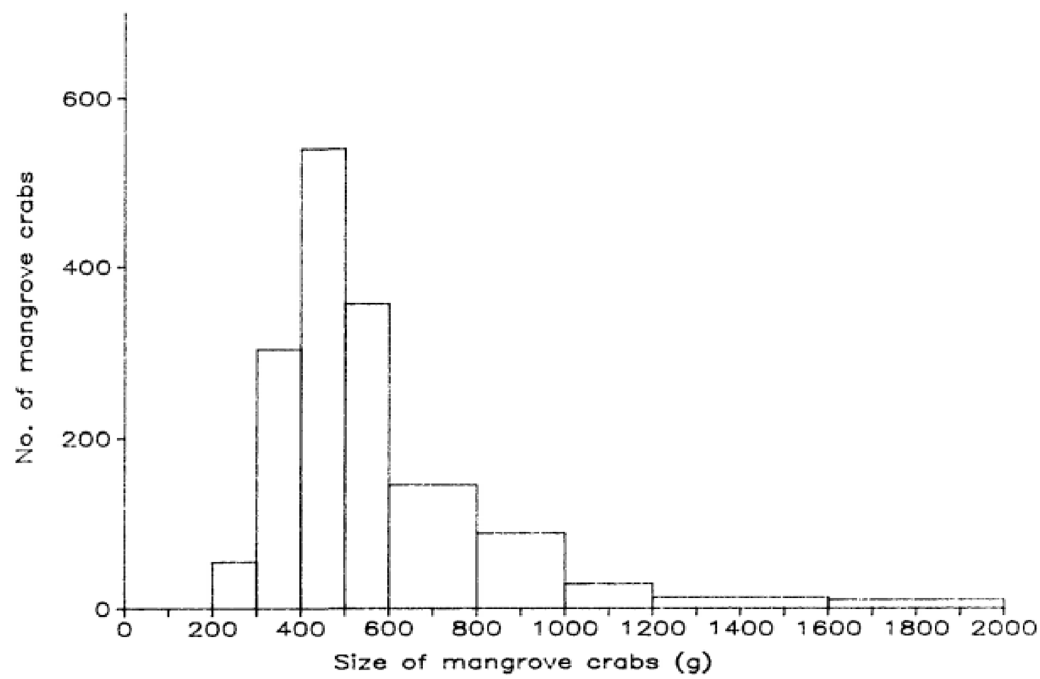
To illustrate how we should deal with a distribution with unequal intervals, we will return to our earlier example of mangrove crabs. In this case we have several different intervals for the crab weights. We must draw each rectangle so that its width is proportional to the class interval, but so that its area is proportional to the observed frequency. Perhaps the best way to achieve this is to calculate for each class a frequency which is an equivalent for the smallest class interval in the table (i.e. 100 g). Thus we can say the the 290 observations for the class 600 to less than 800 g is equivalent to 290 x 100/200 = 145 observations per 100 g interval. In this way all figures are brought to a common basis, and the heights in the diagram will be proportional to these. The calculations are as shown in Table 3.8.

TABLE 3.8 : CALCULATIONS FOR PLOTTING A HISTOGRAM WITH UNEQUAL CLASS INTERVALS

Class	Frequency	Equivalent frequency per 100 gram interval
200 and less than 300	55	55
300 and less than 400	302	302
400 and less than 500	540	540
500 and less than 600	357	357
600 and less than 800	290	290 x 100/200 = 145
800 and less than 1,000	176	176 x 100/200 = 88
1,000 and less than 1,200	59	59 x 100/200 = 30
1,200 and less than 1,600	52	52 x 100/400 = 13
1,600 and over	39	- -

The histogram of this then can be drawn as in Figure 3.2.

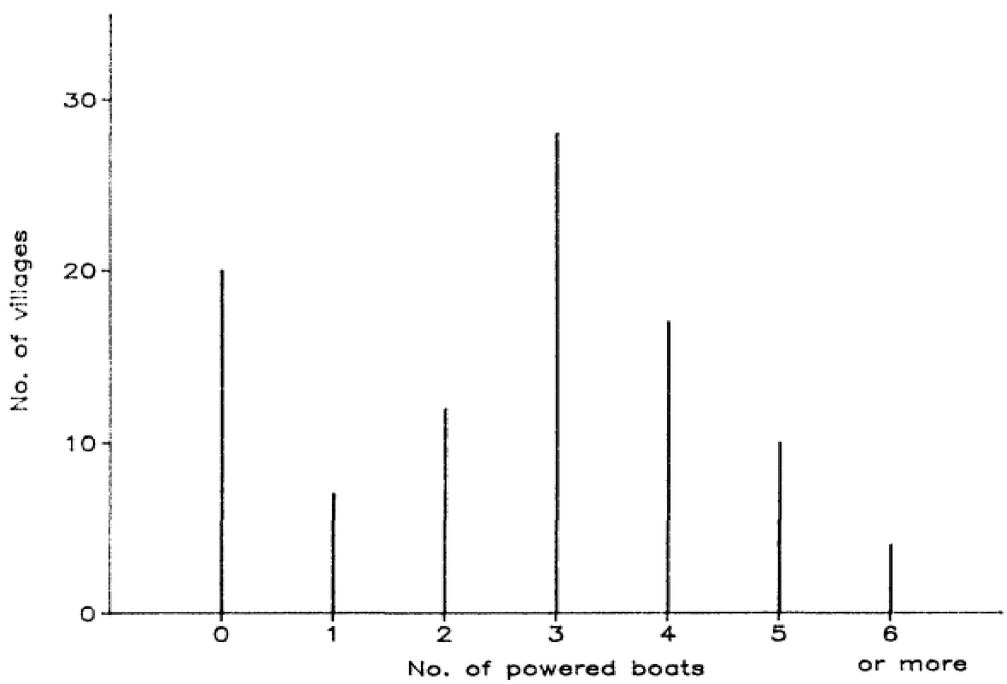
FIGURE 3.2 : NUMBER OF MANGROVE CRABS PACKED FOR EXPORT ACCORDING TO WEIGHT



Apart from the fact that this distribution has unequal intervals, there is one other point that is interesting. There is a problem in deciding how to deal with open-ended classes, in this case the class "1,600 and over". Since we do not know the class width we cannot calculate the height of the rectangle to represent this part of the total frequency. It would be wrong to leave it out altogether, so we have to decide what to do. Basically the problem can be dealt with in one of two ways. Firstly, we can assume an upper limit for the distribution, and draw the rectangle accordingly. The second alternative is to draw the rectangle with a nominal height but leave it open, as in Figure 3.2. This indicates an open-ended interval. This second method only works when the frequency in the open-ended interval is small (as is often the case), and in these circumstances, it is probably the better presentation.

When we are dealing with discrete data we can sometimes consider it to be approximately continuous if the unit of measurement is small compared with the size of the observation. Thus, for example, we may have a distribution of the number of villages according to the number of people in each village. Strictly speaking this is a discrete distribution since we cannot have fractional parts of a person. In practice, however, we can probably treat the data as continuous since the unit of measurement, one person, is small compared with the range of the data, which might be 1,000 people for instance. When we have a discrete distribution where the unit of measurement is large compared with the range of values, such as in our earlier example of the number of powered boats in a village, then we cannot draw a continuous histogram. Instead, we plot the frequencies by means of a simple bar chart of the type we studied in the previous chapter. Sometimes a single line instead of a bar is used as in the example in Figure 3.3.

FIGURE 3.3 : NUMBER OF VILLAGES BY NUMBER OF POWERED FISHING BOATS PER VILLAGE

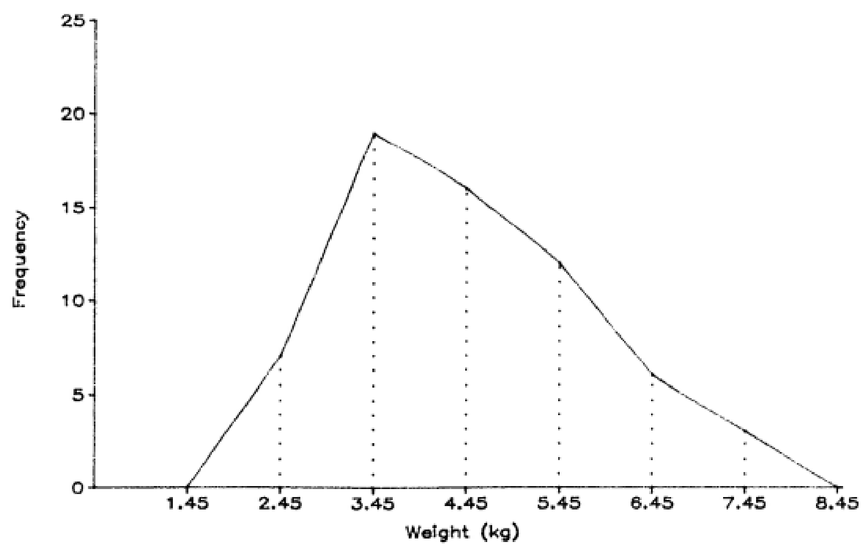


Notice the way the problem of class "6 or more" has been dealt with.

(b) <u>Frequency polygon</u>

An alternative type of diagram, suitable for continuous data, or for discrete data which can be considered to be approximately continuous, is a <u>frequency polygon</u>. In this type the frequencies of each class are plotted at the class mark, and successive points joined up by straight lines. An example of such a polygon for the yellowfin tuna data is given in Figure 3.4.

FIGURE 3.4 : FREQUENCY POLYGON FOR YELLOWFIN WEIGHT DATA

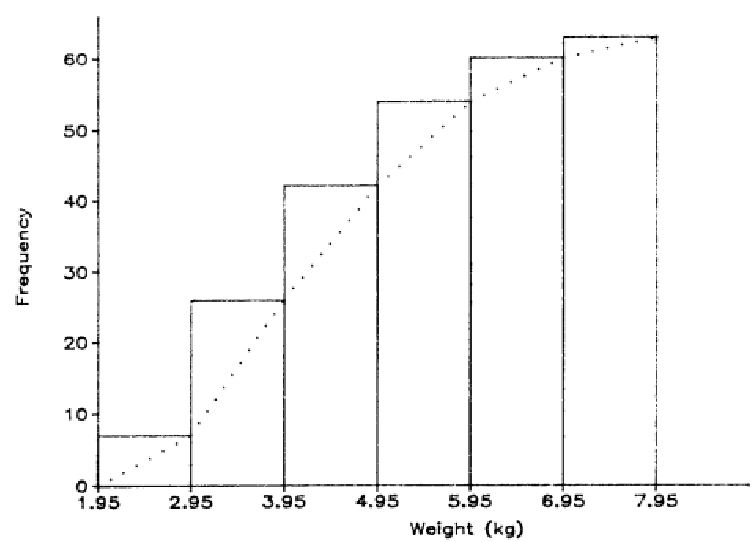


The beginning and end of the polygon should be extended to the x axis, at the mid-points of the classes below and above those covered by the distribution. The area under the polygon then is equal to the area of the rectangles in the earlier diagram. This area, as in the histogram, represents the total frequency.

(c) The ogive

Just as for some purposes it is better to construct a cumulative frequency distribution, so we also find it useful to represent cumulative distributions graphically. If we constructed a cumulative frequency distribution for the yellowfin data, and then draw a histogram, this would appear as a series of rectangles as shown in Figure 3.5.

FIGURE 3.5 : CUMULATIVE FREQUENCY HISTOGRAM OF YELLOWFIN WEIGHTS



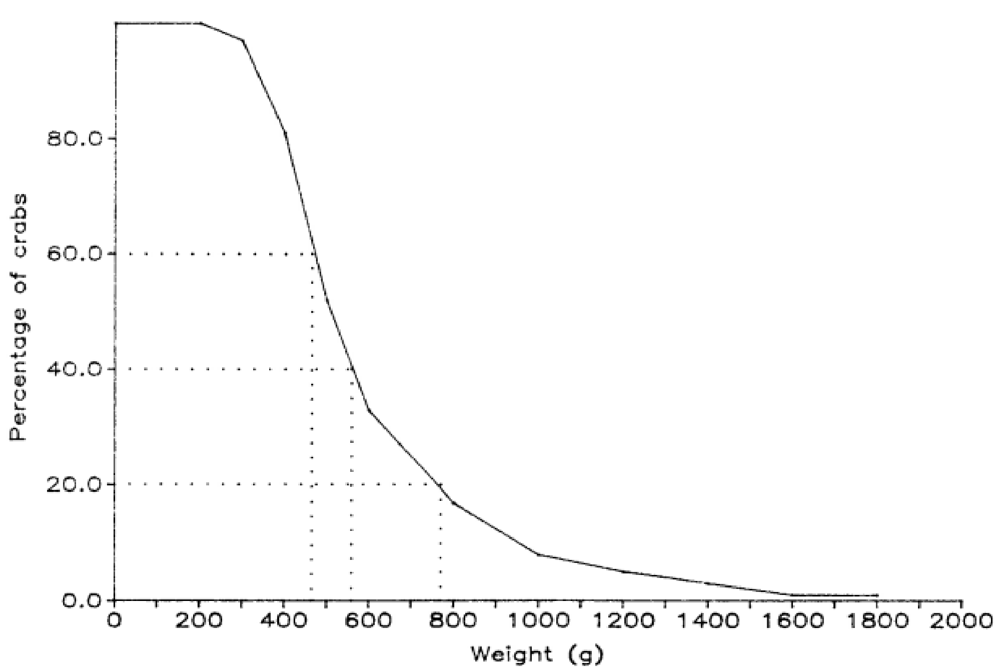
The dotted line in Figure 3.5 represents the cumulative frequency polygon; this is also called the ogive. This joins up the top right hand corner of each rectangle. This diagram represents the cumulative frequency distribution (less than). We can also draw a similar ogive for the cumulative frequency distribution (greater than).

Note that the ogive joins up points on the Figure which represent all the observations up to that point. The first class must therefore terminate at 2.95 kg, and not 3.0 kg; a yellowfin weighing (say) 2.97 kg would not be included in this first class.

This shows the need for care and precision, and highlights the importance of defining exactly what we mean by all our terms. In our crab weights example, we used a different approach and defined classes as "400 and less than 500 g", etc., so in that case the true class limits are 400 g etc. We could have adopted the same approach for the yellowfin, and defined classes as "2.0 kg and less than 3.0 kg", etc. In that case weights would not be rounded, and a fish weighing 2.97 kg would be included in that class. If we had done so, the ogive would be drawn through points at 3.0, 4.0 kg etc. (Our class marks would also be different - at 2.5 kg, etc.) We can record our data either way, whichever is more convenient for us, but we must then take care to make all our calculations accordingly.

Instead of plotting the actual frequencies on the y axis, as shown in Figure 3.5, we could convert this to percentages, and show the relative frequencies instead. The shape of the ogive would be exactly the same, but the y axis would be marked in percentages, from 0 to 100, instead of in numbers. The ogive (greater than) for our previous example of mangrove crab weights, expressed in percentages, is given in Figure 3.6. We can see from the lines drawn on this graph that it is very easy to derive estimates that (for example) 60 per cent of crabs weigh more than 485 g, 40 per cent weigh more than 560 g and 20 per cent weigh more than 755 g.

FIGURE 3.6 : PERCENTAGE OGIVE (greater than) FOR DATA ON WEIGHT OF MANGROVE CRABS



It is worth noting here that all figures of frequency distribution have exactly the same shape regardless of whether they are expressed in absolute values or in percentages. It is often preferable, particularly when attempting to draw general conclusions from data, to express these diagrams in percentages.

Ogives are useful for many purposes. We will see in Topic 4 that they can be used to calculate certain measures, particularly the median, very easily.

3.7 The distribution of the population

Sometimes when we construct a frequency distribution the data we use comprises the whole population that we are considering. Very often, however, the data we have is only part of the population, and we wish to make assumptions about the population from our sample. For example, we are not only interested in the frequency distribution of the 63 yellowfin we have been discussing, but we would hope to be able to say something about the whole population of yellowfin which are caught by pole-and-line method.

In this section we shall see in a very simple way how we can extend the idea of a frequency distribution to a population which has no limit on its size. We call such a population infinite, which indicates that we cannot count all the items in it. The same ideas will also apply to finite populations with a very large number of units; in many cases they can be assumed to be infinite.

Consider the frequency polygon for the yellowfin data which we drew earlier (Figure 3.4). Imagine that instead of six classes we have twelve, but that we double the size of the sample. If we draw a frequency polygon of this data we would expect it to have roughly the same shape as before, but because there are more lines, the shape would appear smoother.

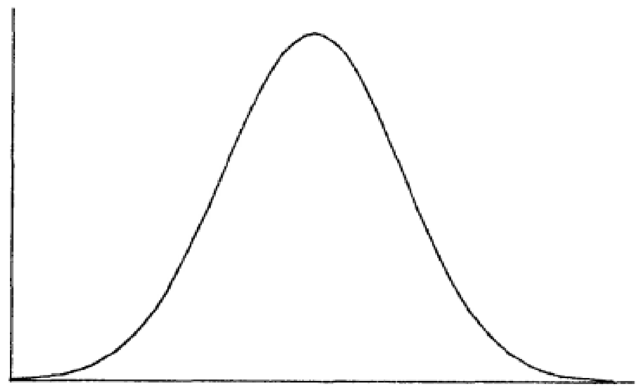
Let us suppose that we repeat the process again, taking a `arger number of classes, but also a larger number of observations. The frequency polygon would be made smoother still. Eventually, as we take more and more classes, but increase the number of observations as well, we could expect to end up with a smooth curve. This curve represents the distribution of the whole population.

The idea of a population distribution is very important in statistics; it forms the basis for a great deal of more advanced statistical theory. We have no time to go into this theory in detail; all we can do is to introduce some of the concepts.

When we collect data from many different sources and then construct frequency distributions and draw histograms and polygons, in many cases we find that the shape of the distribution is quite regular. The distributions we have looked at so far all have quite a simple shape. This leads statisticians to seek a simple mathematical function that will describe or "fit" this shape.

Such a mathematical function, in the case of continuous data, will be smooth and will describe the distribution of the population. If we find that many different populations are distributed in more or less the same way, then we can use these mathematical functions to answer important questions about the population. For example, are the fish in one part of the sea generally larger than those from another part? Does one group of people have a larger income than another? What is the likely range of catch per unit effort from a certain fishing technique? and so on.

One of the most important shapes of a population distribution is known as a "normal distribution". A typical normal distribution is shown below:



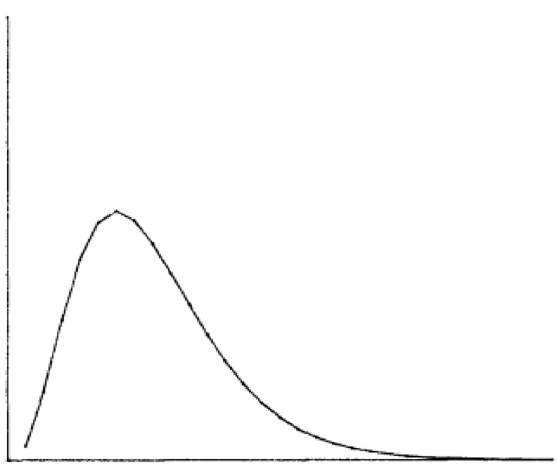
Normal

The normal distribution is symmetrical, that is, both sides are of the same shape. Some examples of distributions which have this kind of shape are:

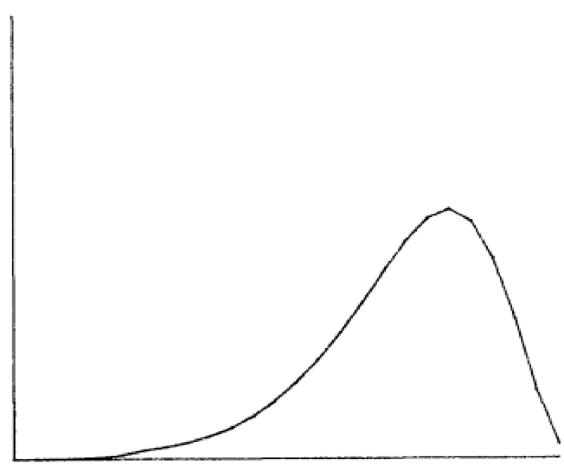
(a) numbers of adult men of different heights;
(b) the number of rainy days in a year;
(c) length frequency of a species of fish for a certain age class.

In these examples, observations will have some high and some low values but a predominance of "average" values.

There are a number of other shapes of population distributions which occur quite frequently, probably the most common being the skewed distribution. This looks like a normal distribution which has been pushed out of shape sideways so that it is no longer symmetrical, as depicted below:



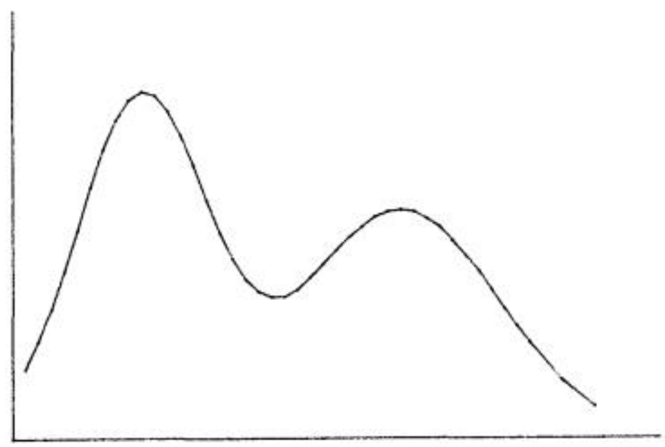
Skewed to right



Skewed to left

An example of a distribution which would almost certainly be skewed to the right is income distribution of the population; in many countries it will be found that there is a heavy concentration of incomes at fairly low levels, and then a long "tail" in the graph stretching out to the right, representing a fairly small number of people with very high incomes. Distributions of catch per unit effort are also usually strongly skewed to the right.

Another distribution quite often encountered is a "bi-modal" distribution. This has two distinct peaks, viz:



Bi—modal

Perhaps the best-known bi-modal distribution is of deaths by age. There is a first sharp peak at age 0, representing infant death, and then (as would be expected) a second, broader peak of deaths in higher age-groups. The distribution of fork length of yellowfin tuna taken by purse-seiners in the Pacific is also bi-modal.

If we examine the diagrams of distributions presented earlier in this topic, we find that they have shapes similar to the types described above. The distribution of crab weights is strongly skewed to the right; that for yellowfin taken by pole-and-line is much closer to a normal distribution, but still skewed a little to the right; and the distribution of villages by number of powered boats is bi-modal.