STATISTICAL METHODS : SOME IMPORTANT CONCEPTS

2.1 Some Basic Definitions

In statistics, as in any other subject, we need to define some words and phrases so that we can use them to have a specific meaning in a particular situation. We shall try to avoid using much "jargon", but we shall need some technical terms. One possible source of confusion here is that many of these words are used in everyday English, but in statistics their meaning is a little different from normal usage. It is worthwhile, therefore, taking a little time to make sure that these terms are understood since we shall use them a lot throughout this course.

When we collect data, for whatever reason, we need to know exactly what kind of information we want, who or what this refers to, how we are going to obtain the data, and for what group of people or items. We use special terms to refer to each of these things; we talk about observing characteristics for statistical units in some population. The terms observing, characteristic, statistical unit and population all have special meanings and we shall look at all of these to see what their definition is and how the term is used.

Statistical unit - We use this term to mean any person, group of people, item or thing about which we wish to obtain some numerical information. Examples of statistical units are: a person, a family, a household, a village, a building, a province, an island, a fish, a boat, a port, a period of time such as a week or a year, a church, a business establishment, and so on. We could think of many more examples.

Population - When we collect statistical data we are interested in obtaining information about a group of statistical units - we use the word "population" to refer to this whole group. In English we usually use the word "population" to mean a group of people; we talk about the population of Suva or the population of Tonga, for example. In statistics we can use "population" to mean a group of any type of statistical unit; thus we can refer to the population of all households in Apia, the population of fishing boats on Rarotonga, the population of all fish caught in Tuvalu in the year 1983, and so on. When we want to collect statistical data we have to be very careful to define exactly what population it is that we are interested in.

Observation - We use this word to stand for the method we use to collect any particular item of information. Usually an observation will be carried out by a person, sometimes with the help of instruments, but there are examples of some machines which will make observations and record the data automatically. It is important to realise that, in the statistical sense, observation can mean any method of collecting data, not just the physical act of seeing and noting something down. Common methods of statistical observation are: measurement, counting, personal judgement, conducting an interview, copying from existing records, a person completing a questionnaire, using self-recording instruments, and so on.

Characteristic - This word is used to stand for some feature or property of the unit that we are interested in. We could, for example, record or observe the weight of a fish, the area of a farm, the value of all goods imported in a port, the annual income of a household, the number of people living in a village, and so on. In most situations, of course, one unit may have many different characteristics, all or some of which we may observe.

For example, when data are obtained about a pole-and-line fishing trip in a country's waters some of the characteristics of the trip which may be collected are:

port of departure;
country of registration;
gross tonnage of vessel;
number of crew;
days spent fishing in territorial waters;
species of fish taken;
quantity of fish taken;
average weight of fish.

A characteristic can be one of two types. In the first case it may be expressed only in numerical values, and we call this type of characteristic a variable. Other characteristics do not take numeric values, and have to be described in words. This second type is referred to as an attribute.

From the list of characteristics of the pole-and-line trip given above, we can identify the following as variables:

gross tonnage of vessel;
number of crew;
days spent fishing;
quantity of fish taken;
average weight of fish.

The attributes in the list are:

port of departure;
country of registration;
species of fish.

It is often more convenient, especially when using a computer, for statisticians to work in numbers rather than in words. Therefore, we sometimes allocate numerical codes to attributes. For example, we might allocate code 001 to skipjack, 002 to yellowfin, 003 to bigeye, 004 to albacore, and so on. Then we would key into the computer this code number, rather than the name of the species. However, it is important to recognise that, even though this characteristic has now been recorded in numeric form, it is still an attribute, not a variable.

2.2 Notation

During this course we will make use of some special statistical notation. This is the statistician's shorthand way of expressing a concept which would otherwise be cumbersome and long-winded to express. We will try to keep the notation as simple as possible.

This special notation will be introduced progressively during the course, but a few basic symbols should be described immediately.

n, N : The number of observations under consideration is denoted by "n" in the case of a sample, and "N" in relation to the whole population. Thus, if we collect data from a sample of 17 fishing boats for a survey, we say that n = 17. If the fleet consists of 80 boats, we say that N = 80.

x — When one variable is being considered, "x" is used to denote the values of the observation. This symbol is often followed by a subscript to describe exactly which observation is referred to. That is, x_1 refers to the value of the first observation, x_2 to the value of the second observation, and so on up to the final (i.e. nth) observation, which is denoted by x_n.

Thus, if we were measuring fork length of fish in centimetres, and the length of the first fish in our sample was 62 cm, we would say that $x_1 = 62$. (The whole set of n observations can be described by reference to x_1, x_2, x_3, ... x_n.)

y — When we are considering two variables, the second variable will be denoted by y with subscripts as required. So if we were conducting length-weight comparisons, and the first fish weighed 3.8 kg, we would say that $x_1 = 62$ and $y_1 = 3.8$.

i — For convenience any particular observation is referred to as the ith observation. So in order to refer to the value of the first observation we say i = 1 and so on. We will find that "i" is most often written as a subscript. So that for instance x_i means the ith observation of our variable x.

Σ — This is the Greek letter, capital sigma, and means simply "the sum of". It must not be confused with σ, which is the ordinary Greek letter sigma, and which will be introduced later in the course.

2.3 Diagrams

In the following topics in this manual, we will be representing a number of statistical concepts in diagrammatic form. We will use three types of diagram - a scatter diagram, a graph and a bar chart. At the same time we will make brief mention of pie charts, which, although not specifically required in the later topics, are a very useful way to portray information diagrammatically.

The topic of diagrams is a very important one, and the way diagrams are used to portray results of statistical surveys can greatly affect the understanding of the results. However, for the present we are not going to explore this topic in detail; we will simply touch on the basic principles of constructing these types of diagram, as a lead-in to the following topics.

A diagram is used to demonstrate the relationship between characteristics. The main components are:

(a) Heading: essentially a diagram number and a title, describing what the diagram represents;

(b) Two axes: a vertical or "y" axis and a horizontal or "x" axis. These meet at the origin ("0"). Each axis must be clearly labelled and values of the variable or attribute are plotted along it according to some scale;

(c) The data: plotted on the diagram, depending on the type of diagram being used.

2.3.1 Scatter diagram

When we draw a scatter diagram, we are looking at the relationship between two characteristics. In general, we shall have a number of observations of both of these for a series of statistical units. We shall be concerned here almost exclusively with variables. We shall assume that we have a sample of n units, and for each unit we shall observe two variables which we can denote by x and y. Thus, for the first unit, the observations can be written as x_1y_1, for the second unit x_2y_2, and so on. In general, for the ith unit, our observations will be written as x_iy_i and there will be n such pairs. If we have a graph with two axes to represent the two variables, then each pair of observations can be plotted as a point; the co-ordinates of the pairs will be the values (x_i, y_i). This kind of graph with all n observations plotted as a number of points is called a "scatter diagram". It is a very useful first step in looking at the relationship between x and y.

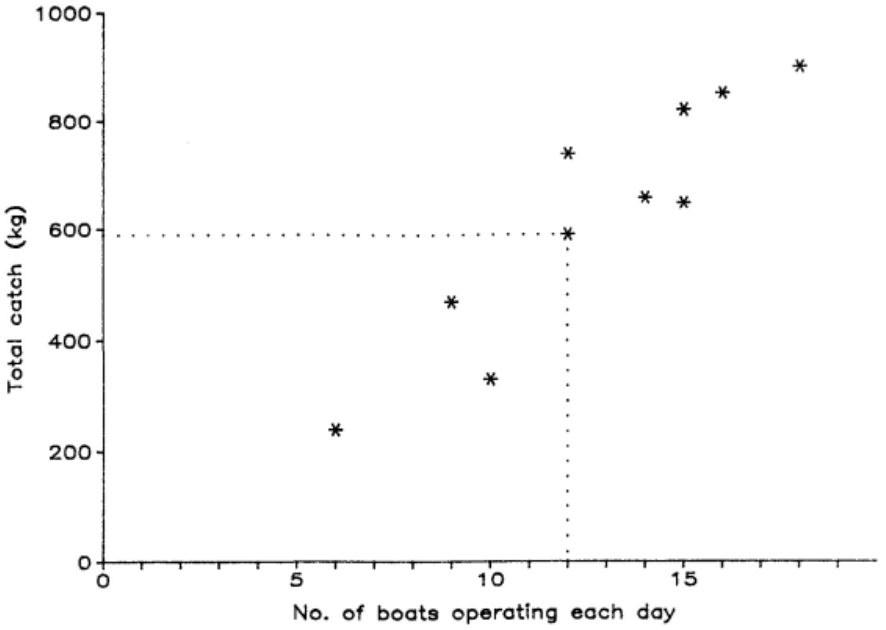
For example, suppose that on 10 selected days we recorded details of the number of boats fishing, and the total daily catch from an artisanal fishery (Table 2.1):

TABLE 2.1 : BOATS OPERATING AND DAILY CATCH AT AN ARTISANAL FISHERY

Day	Boats operating (x)	Total catch (kg) (y)
1	12	590
2	15	820
3	10	330
4	12	740
5	18	900
6	14	660
7	6	240
8	15	650
9	16	850
10	9	470

In a scatter diagram we will show number of boats on the x-axis and catch on the y-axis. The result for the first day is plotted as a point, where the vertical distance above the x-axis is equal to 590 kg and the horizontal distance from the y-axis is equalivalent to 12 boats. This point represents the pair of observations x_1y_1, as shown in the diagram. The dotted lines are also drawn in to show exactly how this point was located. In practice, however, in constructing a scatter diagram we are interested in showing just the distribution of points. Figure 2.1 is the scatter diagram showing all 10 pairs of observations.

FIGURE 2.1 : NUMBER OF BOATS OPERATING AND TOTAL CATCH PER DAY

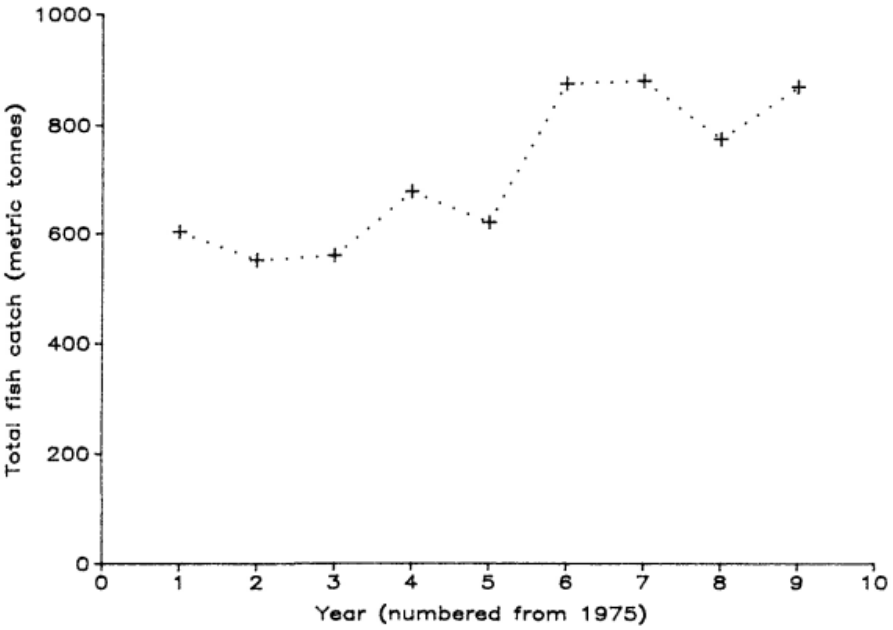


2.3.2 Graphs

A graph is very similar to a scatter diagram, showing the relationship between two variables, but with the points linked up by lines, to show the trend in the relationship.

Graphs are very useful when we show how some variables change over time, as in Figure 2.2.

FIGURE 2.2 : TOTAL ANNUAL FISH CATCH IN COUNTRY ABC



We may note that drawing lines to link up points has a real meaning in this situation, because the slope of each line shows whether catch is going up or down from one year to the next. However, linking up points in the previous diagram would not make sense. It would not be showing a trend.

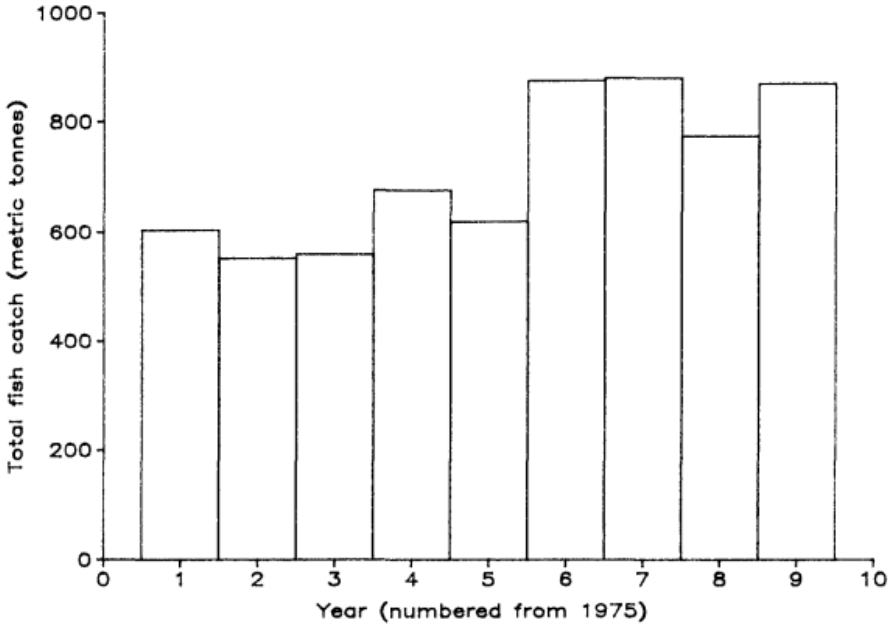
Graphs are not necessarily constructed by straight lines joining up a series of points, as in that illustration. Often we have curves, to represent the shape of different distributions, and we will study that in the next topic.

2.3.3 Bar charts

If we wish to prepare a diagram of data classified by some attribute, rather than by a variable, then a line graph is not suitable. For example, if we have statistics on production of fish by district, we cannot place districts along the x-axis and join up a series of points by a line. Such a line would be meaningless. In this situation the best form of presentation is a bar chart.

However, the use of this type of diagram is not restricted to attributes. We can also depict relationships between variables on a bar chart. Figure 2.3 shows the same data of fish catch over several years, which we previously depicted by a line graph, in the form a bar chart.

FIGURE 2.3 : TOTAL ANNUAL FISH CATCH IN COUNTRY ABC



The bars can be drawn fairly close together, or further apart, and may be shaded or cross-hatched to improve the appearance. A little later, we will look at a particular type of bar chart, called a histogram, in which data is represented in a series of bars which are contiguous.

2.3.4 Dependent and independent variables

Having chosen the type of diagram we require in order to best illustrate the data, we next need to decide which characteristic will be plotted on the x-axis, and which on the y-axis. To determine which way round to draw the diagram, we need to see if there is likely to be any form

of relationship between the characteristics. In many cases we can say that we are interested in seeing how one variable changes as another variable or attribute changes. In our example of total fish catch in Figures 2.2 and 2.3, we are trying to show how the level of catch changes as time changes. In Figure 2.1, we are interested in how the catch varies according to the number of boats engaged.

In Figures 2.2 and 2.3, we may say that the total catch depends on time; in Figure 2.1, that catch depends on the number of boats engaged. More formally we say that we have a dependent variable and an independent variable (or attribute). In these situations we cannot reverse the relationship. It would be silly to say we were looking at how time varied depending on the level of fish catch, for example.

Given this kind of independent-dependent relationship we always plot the dependent variable on the y-axis and the independent variable or attribute on the x-axis. This is a mathematical convention, it is used for convenience, and it makes diagrams easier to understand.

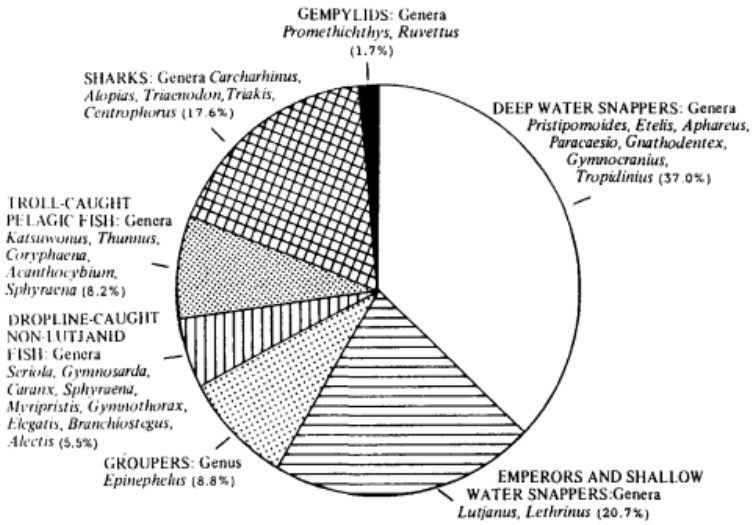
Where there is no clear direction of dependence between the characteristics then it does not matter much which one is plotted on which axis. This situation will not occur very often in practice, as we will usually find that one variable can be considered to depend on the other.

An example, which we will study later, is the relationship between the length and weight of fish. It may be argued that each one depends on the other, and that there is no clear dependent/independent relationship. Even here though, there is a convention, and it will be found that weight is always plotted on the y-axis, and length on the x-axis.

2.3.5 Pie charts

A basic pie chart consists of a circle divided into a number of sectors. Each sector is used to represent a particular value of a characteristic, the area of each sector being proportional to the share of that characteristic to the total. Either variables or attributes can be portrayed in this manner, but a pie chart is especially useful for attributes, as in Figure 2.4.

FIGURE 2.4 : CATCH COMPOSITION BY WEIGHT (GENERA WITHIN GROUP LISTED IN DECREASING ORDER OF IMPORTANCE)



The normal practice, as shown in this figure, is to commence at the top of the chart (the "12 o´clock" position) and to work clockwise from there, with the largest or most important sector being shown first.

A pie chart such as this is quite easy to prepare. The area of any sector of a circle is proportional to the angle at the centre between the two radii. Since the value of the characteristic has to be proportional to the area, all we have to do is to draw a number of sectors with angles proportional to the value of the characteristic. The sum of the angles will of course be 360 degrees. The calculation for any category is then quite simple.

Notice also, that since we are dealing with proportions we can prepare the pie chart either from the actual data or from a percentage distribution. With too many categories a pie chart becomes confused and difficult to read; as a general rule eight is about the maximum number that should be included.

2.4 Rounding of numbers

During later topics we will encounter situations where we need to round numbers to a certain number of significant digits, or to the nearest one decimal place. It may also be necessary in publishing survey data to present results rounded to the nearest tonne, or to the nearest thousand, etc. To ensure that this is done in a consistent way, we need a standard rounding procedure.

The basic principle is to round to the nearest significant digit. Thus, if we wish to round 428,548 to the nearest thousand, we would record this as 429,000. When we need to round a number which is exactly halfway between two significant digits, we adopt a convention of rounding so that the last significant digit is even. So we would round the number 428,500 to 428,000, in preference to 429,000.

We should note here that, when a series of numbers and the total of those numbers are rounded, it may happen that, after rounding, the sum of the components is not equal to the total. For example, let us consider the following, where a set of numbers is to be rounded to the nearest thousand.

128,613	rounded to	129,000
428,548		429,000
37,924		38,000
595,085		?

Clearly the total should be rounded to 595,000 according to our rules, and yet the sum of the three rounded numbers is 596,000.

This gives us a problem in how to present the data in rounded form, and as similar situations will often arise in practice, we need a convention to deal with it. It is recommended that each number be rounded correctly according to the rules (the total in the example being rounded to 595,000), so that the sum of the components may not be exactly equal to the total. This disadvantage of this is that users may notice that the total does not correspond exactly with the sum of the components, and may conclude that an error has been made. To counteract this it is normal practice to include in publications a note, such as "Any discrepancy between totals and the sum of components is due to rounding".