1

THE SCOPE OF STATISTICS

1.1 Broadly speaking, statistics is the numerical study of a problem. Unless the problem can be reduced to quantities measurable on a scale or capable of being expressed as a number, it is impossible to make a statistical study of it. Statistics is not, however, just concerned with the counting of individuals or the measuring of items. Its ramifications are far wider than that, and include the study of what are the right figures to collect and the correct interpretation to be placed on them. The politician trying to envisage the effects of different forms of taxation needs to know the estimated yields of each form of taxation proposed; and the local town councillor must be able to appreciate how the local rate is split up into various headings. The citizen today is deluged with White papers, Economic surveys and a multitude of reports not only from the Government but from banks, insurance companies and industrial firms, all of which present, and argue from, a mass of statistical data. An understanding of statistics and the treatment of numerical data is therefore essential and only by a patient study of the part played by figures in such reports can good decisions be made and policies understood and, if need be, criticised. It is necessary to recognise, moreover, the power and the limitations of statistical arguments, to learn how to obtain the full information from a set of figures and how to avoid the pitfalls which await the unwary. If the figures are worth analysing at all they are surely worth the form of analysis that yields the maximum amount of information. There are a large number of statistical tools, and to use a steam-hammer where a light tap is required would be not only wasteful but often misleading. No one universal rule can be made and both knowledge and experience must be gained if the best possible results are desired.

The whole subject of statistics has taken tremendous strides since the beginning of the century, and the last war gave a big impetus to the further study of statistical methods, since the use of such methods often led to large savings of time, materials and personnel. Before a study of the basic methods of the subject, a

number of examples will be given to illustrate some of the many fields of its application. Whilst these examples are in no way exhaustive they are nevertheless instructive in that they give some idea of the multitude of problems that confront the statistician.

1.2 Government. For effective government and the shaping of policy it is necessary to have accurate statistical knowledge of the exact composition of the population. From this need has sprung in the first place the compulsory registration of births, marriages and deaths, and secondly the census which was inaugurated at the time of the Napoleonic wars and is normally taken in Great Britain once in every ten years. The information derived from these sources gives an instantaneous picture, as it were, of the population divided up by age, sex, housing and so on. This information is needed in order to see, for example, how many schools are required, how many workers there are, and how many persons are pensioners and hence no longer productive workers; estimates can then be made of important quantities, such as the working population ten years hence. By collecting details of housing conditions (number of rooms, washing facilities and so on) the census also gives valuable information about the social well-being of the nation.

Table 1.1. *Acreage in Great Britain utilised for agriculture in June 1967*

Type of agriculture	Area in thousand acres
Arable land under crops	12,354
Arable land under grass	5,934
Permanent grassland	12,328
Rough grazing	17,639

The well-being of a nation, however, depends not only on the population but also on the land and such natural resources as minerals, forests and livestock, and on capital equipment in the form of buildings, factories and machinery. To obtain information on all these subjects the various government departments make large-scale surveys at regular intervals. Much of the material obtained by the government is published in the *Monthly Digest of Statistics*, from which the information in table 1.1 is taken. This

publication is probably in your local library. A study of it will give you some idea of the vast amount of information used by the government in making decisions concerning the day-to-day policy of the country.

1.3 Industry. The wealth of the country does not depend primarily on the government. It has to be generated elsewhere, and industry is one of the main sources of that wealth. The production of marketable goods of adequate quality provides innumerable statistical problems of which the following are typical examples.

A manufacturer is making electric light bulbs and according to the design and specification the bulbs should burn for 2,000 hr. Due to slight differences in manufacture the bulbs will not all have exactly the same length of life, but will vary amongst themselves. A batch of 500 bulbs has just been produced and the manufacturer is going to make tests in order to see whether the bulbs are up to the standard and have a burning life of at least 2000 hr. Quite clearly he cannot test every bulb by measuring the time it takes to burn out for there would then be none left for sale. Hence it is essential to use some method whereby a few of the bulbs are examined and inferences made from them about the whole batch. If the whole batch could be examined it would be possible to make a categorical statement such as 'the bulbs all have a life of at least 2,000 hours'. As only a few of the bulbs can be examined the statement must take the form 'the bulbs almost certainly all have a life of at least 2,000 hr.' or 'the bulbs are very unlikely to have a life of at least 2,000 hr.', or something in between these two statements. But provided that the selection and examination of the bulbs is in accordance with the principles given later, then the latter forms of statement may give as much information as is required at a fraction of the cost of a complete examination, which in any event is an impossibility in this case.

A similar situation would arise in the determination of the breaking strength of a batch of steel wires, where a test destroys the wire, and the desired information must be obtained by performing tests on a selection of the wires. Such a procedure is necessary in order to check whether the quality of the product is being maintained, as a failure could be dangerous and might well result in a falling-off of sales.

Industrial problems also arise in the trial of new processes, as in the following example. Two batches of cloth are made by different processes, A and B, and the resistance of the cloth to acid is tested by taking four pieces as samples from each batch and measuring the length of time for which they resist the acid. The results, in hours, of a typical experiment were as follows:

	Process A	Process B
Piece 1	40·9	43·8
2	40·7	41·7
3	41·6	43·1
4	41·2	44·2

Process B is a new process and the question is whether it is better than Process A? It is true that all four results for B are higher than those of A, but only just. Notice that the four values for B vary much more amongst themselves than do those for A. This suggests that the samples from B are not so consistent as those from A. The final decision must reconcile all these factors and this type of problem is analysed in detail later in this book.

1.4 Road Safety. Industrial goods have to be moved to the places where they are needed and much of the transport is by road. In recent years there have been more and more accidents on the roads, with the result that new safety measures are constantly being devised. The only sound basis for judging the relative effectiveness of suggested precautions for the reduction of road accidents is a statistical one. Thus a common form of statement is that road junction X is much more dangerous than road junction Y. A statement of this sort cannot be based merely on the number of accidents that occur at the two junctions. One junction may carry much more traffic or may have more traffic at dusk, which is a peak period for accidents. Or again, at one junction all the traffic may go straight across whereas at the other a large proportion may make a right turn.

Other common assertions are that high speeds are the most frequent cause of accidents, or that different parts of the country have different accident rates, or that one form of road surface is more dangerous than another. To substantiate these statements a considerable amount of statistical evidence is needed, since the opinions of individuals are widely conflicting, and only by a careful

numerical study of the various factors involved is it possible to arrive at an impartial decision.

Statistical work is also necessary in solving the problem of traffic congestion in towns. To make vast and expensive alterations to the system of traffic control without making sure the alteration will have the desired effect is both useless and wasteful. For instance, the common type of roundabout will take only a certain amount of traffic per hour, and any attempt to force more traffic to use it will only result in large-scale congestion. Hence before installing a roundabout a detailed study of the volume of traffic coming into the junction at various times of the day must be made. If this is not done a hoped-for improvement could have the opposite effect.

London and most other cities have large and complicated networks of public transport and, although such networks may appear somewhat haphazard to the uninformed, there is scope for an enormous volume of statistical work in the background. The organisers want to know how people get to work, or to the shops, or to places of entertainment, and how long it takes them; they want to know the effect of alterations in fares on total receipts, and on the pattern of the journeys. All these things and many more must be studied in order to choose the most useful routes, and arrange the most convenient time-tables, consistent with keeping the running costs as low as possible.

1.5 Insurance. The citizen comes into even more direct contact with statistics in the field of insurance. For example, Mr *A* insures his house against damage by fire by paying an annual sum or premium to an insurance company. The amount of this premium is not arbitrary, but is governed by the numbers of fires that do occur in houses of similar type and the amount of damage they cause. There are factors that enable the financial risk of a fire at Mr *A*'s house to be directly assessed. Again to assess the correct premium to charge for insuring a motor car requires knowledge of the claims made on that type of car, for the particular age of the driver, his occupation, previous claim record and so on. All this information needs to be collated and carefully assessed so as to yield the important factors that determine the correct premium to charge.

An alternative form of insurance policy is that for an annuity.

According to the prospectus of one large assurance office they will grant an annuity of £11.30 to a man now aged 60 if he will pay them £100. This means that if Mr *B*, now aged 60, were to pay over to the office the sum of £100 they would pay Mr *B* the £11.30 every year for the rest of his life. If Mr *B* dies in two years' time he will receive two payments, but if he lives to be a centenarian he will receive forty payments. In order to determine the amount of the annuity it can offer the assurance office must have accurate information on the number of years a man aged 60 is likely to live, and on the rate of interest they may expect to earn on invested money. A study must therefore be made of the distribution of ages at death of all men over 60 in the past, and examination made of the rates of interest that are obtainable on various forms of investment. Then the rate of annuity is calculated by spreading the risks over all men aged 60 who buy annuities.

1.6 Market Research. All the goods and services available to the public have to be shown or demonstrated to them in order that people shall know of their existence. The demand has to be measured and the reactions of the public noted for future developments. Imagine that a firm is putting on the market a new type of washing machine. Obviously it is desirable to know what will be the approximate demand for the machine. If the demand is only going to be of the order of 100 machines a week it is very unprofitable to set up plant and machinery designed to turn out washing machines at the rate of 5000 a week. The concept of gauging the demand for a product is the basic problem in market research. In its simplest form it is answered by questioning a proportion of households on their intentions, that is, the likelihood of their purchasing a washing machine of this kind. A similar procedure is followed by the B.B.C. in its Listener research organisation, which continuously investigates the popularity or otherwise of its programmes. The statement that 8,000,000 people listen to such and such a programme does not mean that every person in Great Britain has been questioned. In practice only a few thousand will have been questioned, but care will have been taken to see that these are representative of the whole population, so that valid deductions about the listening habits of the whole population may be made.

1.7 The fields of application of statistics are boundless and cannot all be mentioned. In agricultural work the decision as to whether one variety of wheat is better than another will be made on the results of a carefully planned series of statistical experiments. In biology the effects of two drugs on rabbits, may be compared and the effects of the two drugs assessed on some numerical scale. Again, if a company makes steel bars of a specified length, the bars turned out in practice will inevitably vary in length. Management must determine the average length for which the process should be set so that the total scrap—resulting from trimming long bars plus that resulting from scrapping the short ones—is at a minimum.

All these examples show how statistical methods are essentially a guide to action or decision of some form. It replaces a vague personal impression such as 'I think that Drug *A* is better than Drug *B*' by a well-defined and clear-cut statement of the form 'Drug *A* is 40 per cent more powerful than Drug *B*'. Everyone argues from general impressions, and the art of conversation would be very difficult if sweeping statements such as 'aeroplanes are more dangerous than cars' were inadmissible without supporting statistical evidence. Nevertheless general impressions are often misleading and sometimes untrustworthy. For example, people are apt to say that they are always getting wrong numbers when using the telephone. A statistical count would, in all probability, show that the proportion of wrong numbers was really very small. As is so often the case, the times when things go wrong are remembered but the numerous times when all goes well are forgotten.

The examples of the applications of statistics given above illustrate the usual pattern of a statistical investigation in which there are four phases, namely:

(*a*) Statement of the problem to be investigated.

(*b*) Collection of the data needed, either from available sources or by performing experiments.

(*c*) Analysis of the experimental results or data.

(*d*) Interpretation of the results of the analysis.

Thus to see how many schools will be needed in Manchester in 1980 is to state the problem as in (*a*). Next, from census figures and local figures relating to movements of families in and out of the city, together with a forecast of birth rates, the numbers of children in various age-groups expected to be in Manchester in 1980 must be

obtained. This constitutes (*b*). The data must now be sifted down in order to produce the comprehensive figure required for stage (*c*) and finally in stage (*d*) the decision is made on the number of schools required.

Or again the insurance company in section 1.5 wishes to find how much annuity it can grant for £100 to a man aged 60. This is the problem under (*a*). Next, the company collects together all the available information as to how much longer men aged 60 have lived in the past, together with the rates of interest at which the company is likely to be able to invest its money. This is stage (*b*) and leads to stage (*c*), the sifting of this information to give the required figures. Since there may be alternative estimates of the annuity from the analysis in (*c*) the interpretation of the results under (*d*) will require the exercise of judgment and experience in order to decide what annuity can be offered.

In the remainder of this book these subdivisions of any investigation will be discussed at some length, and the powerful aid that statistical methods can give to clear thinking and rational decisions will be demonstrated. Since each problem that turns up will be slightly different from the next one, practice is essential in order to acquire facility, and the student is urged to work through as many of the examples and exercises as is possible.

2

THE COLLECTION OF DATA

2.1 In the opening stages of a statistical inquiry the investigator will need to collect a large amount of raw material or data from which to extract the quantities relevant to the purposes that he has in mind. Thus the market research investigator may have to collect data on consumer preferences after four different sets of advertisements have appeared; the botanist may have to spend several days in a grassland area counting the number of shoots of *Solidago glaberima* per square foot; the traffic investigator may have to count traffic at a busy crossing; the agriculturalist may have to collect data concerning the quantity of fertiliser applied to wheat crops on all the farms in Sussex. The method and care given to the collection of this raw material is important. The strength of a chain lies in the strength of its weakest link and it is useless to reach intricate conclusions from insufficient or inaccurate data. Before making any form of elaborate analysis it is essential to know the limitations and accuracy of statistical material and to be aware of the kind of errors that can arise. In this chapter two of the most common sources of data will be considered in some detail. These are:

(i) *Questionnaires*. The data here are obtained by forms designed by the statistician and completed by the general public.

(ii) *Observations*. The data here are collected by the investigator himself recording the results of a series of observations but not necessarily relying on the public at large for his information.

2.2 Questionnaires and their completion have to some extent become a part of the daily life of the citizen in this country, though there have been for some time special forms concerning each citizen's history. For example, his birth must be registered at a local registrar's office, within 42 days, on a form somewhat similar to that shown in fig. 2.1. This registration is required by British law, and on marriage and at death similar types of form have to be completed. These forms provide the raw material for a large number of studies made by the government into the size, age-distribution and marital status of the population. The record of

a birth, marriage or death can be inspected at Somerset House in London, where the records have been kept since this type of registration became compulsory in 1874. It is sometimes necessary to prove one's age (in order to sit for certain examinations, for example, or to receive a legacy conditional on being twenty-one) and the birth certificate provides a ready means for this purpose. However, these records of births, marriages and deaths are not always by themselves sufficient for many statistical problems. They give no idea, for example, of the population of a town at a given time or of the occupations followed by its residents. To answer such questions some further source of information is necessary and some more continuous check on movements would therefore be necessary. This was attempted during the last war by means of identity cards which were useful also in operating a fair system of rationing. In normal times, however, this continuous check is regularly made by carrying out a census.

Birth in the district of.....in the County of.....

When born	Name	Sex	Name and surname of father	Name and maiden surname of mother	Profession of father	Date of registration	Signature of registrar

Fig. 2.1. Abbreviated form of a birth certificate

2.3 A census is a very comprehensive affair and in peace time is carried out simultaneously throughout the length and breadth of Great Britain every ten years. The last complete census was held on 23 April 1961, when the head of every household was responsible for the correct filling up of a schedule asking for certain particulars for every member of his household. There were in all thirty-six questions on the schedule but many of them did not apply to everyone. The questions fell, broadly speaking into the following categories:

1. Name.
2. Sex.
3. Age.
4. Whether married, etc.