Lesson One

Descriptive Statistics

Aims

The aims of this lesson are to enable you to

- know the difference between the population and sample
- learn the terms qualitative, quantitative, discrete and continuous and what a variable is
- explain the difference between primary and secondary data
- use information technology effectively

Context

Much of this lesson will seem very familiar to most students as the topics will have been covered at GCSE or elsewhere in 'A' level studies.

The ideas behind this lesson are in Chapter 1 of *Statistics 1*, pp.1-6.

Studying Statistics

"There are lies, damned lies and statistics!" (Sir Benjamin Disraeli)

How much truth is there in this popular quotation? Before you are tempted to use it yourself, consider the following points:

- This quotation is used by people who feel that there is something deceitful about the subject and that a lot of it is "mumbo-jumbo".
- Hopefully anyone completing this course will disagree and be able to recognise when somebody is attempting to pull the wool over their eyes by abusing statistics.
- The word **statistics** has two meanings. Either it is the title of the subject being studied or it is a general term used for facts and figures or data, as in the case of the "government unemployment statistics".
- Much of this subject is concerned with understanding and simplifying a **sample** of data taken from some large **population**. Working on the sample only, we can then proceed to try and understand the whole population. It's as though we have a barrel full of raffle tickets and take some out but are only allowed to look at these ones to predict the likely make-up of the rest of the tickets in the barrel.
- Since an element of chance is involved as to which tickets are drawn, the mathematical theory of **probability** lies at the heart of all advanced work in statistics.

Statistical Data

Statistics deals with the collection, organisation, analysis and interpretation of numerical data. Statisticians must decide carefully which group of people they intend to study and for which purpose the study is to be made. The group of people or things to be studied is known as the **population**. Ideally, one should carry out a census and study every member of the population, but unless the population is very small this will be very time-consuming and expensive. For this reason, it is normal to select a representative **sample** of that population. The size of the sample depends on many factors.

Statistics are used in many walks of life, e.g. sport, industry, finance, government, etc.

At a General Election we are bombarded with statistics. Statistics are also used in quality control testing in industry. For example, a tyre manufacturer will find it useful to know the average life of tyres. However, there is no way that every tyre can be tested to destruction otherwise there would be nothing left to sell. The only way out of this problem is to take samples from the population.

When dealing with materials, this is quite simple in that every 100^{th} item could be tested. When dealing with people, however, much greater care is needed. Public opinion pollsters often use the Electoral register and take, for instance, every 10^{th} or 100^{th} name. Some statistics are gathered in the street. It is here that bias can easily creep in. An interviewer is more likely to stop a reasonable (in their opinion) person than one who looks scruffy or belligerent.

Variables

A variable is simply a quantity or an item which is subject to change or variation; examples would be:

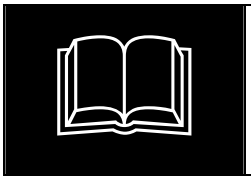
- the type of pet kept, such as cat, dog, rabbit
- The number of people living in the various households in a particular street
- The heights of a group of people

Each of these examples is a different type of variable.

Qualitative variables are those which are not numerical. A survey on family pets indicates a type of animal, as in the first example. Although the data will show the numbers of people keeping each type of pet, the pets themselves are described by words, not numbers. **Quantitative** variables are those that are numeric.

Discrete variables are those which can be expressed by certain numbers within a range. For example, the numbers of people living in the houses in your street are measured in integers only. You cannot have 2.4 people living in a house, so the numbers which you can use are separate or discrete.

Continuous variables are those which can take any value within a given range. For instance, the height of an adult human may vary between about 0.80 and 2.40 metres. Within this range, any height is possible, even 1.789654 metres. The fact that a person of this height would describe himself or herself as being 1.79 metres tall is irrelevant: such a figure is only an approximation anyway. Even 1.789654 metres would count as an approximation. For continuous variables, we usually group the data, so we might use groups of 1.50 to less than 1.55m, 1.55 to less than 1.60m, and so on. This helps us to make sense of the data collected.



Now try your first exercise on p.3, Exercise 1a.

Primary and Secondary Data

Primary data is original data collected by you. You must collect it yourself, and design a way of doing so. For example:

- asking people where they shop
- collecting shoe sizes from a class of pupils
- traffic count

Secondary data is data collected by someone else. For example:

- a census
- a table of unemployment figures
- population statistics

Random Samples

The term "fair sample" has been used earlier. What is a fair sample? Well, the term "fair" is rather vague and. statistically, we require something which can be defined more clearly. So it is usual to talk of a **Random Sample**. This is one in which each member of the population has an equal chance of being chosen to be part of the sample.

Problems of Sampling

In practice, choosing a random sample is rarely easy since some subjective bias almost inevitably creeps in. For example, if you ask various people to write down a series of 5 digits, chosen at random, most people automatically write down what they think are 5 random digits and it is unlikely that they will include a repeated digit. Yet the probability of five digits produced at random being all different is only about $\frac{3}{10}$.

Parameters and Statistics

A good picture to have in your mind of what often happens in Statistics is to think of a large number of various coloured raffle tickets inside a big barrel. The raffle tickets are then the

population. A number of tickets are randomly drawn from the barrel. These are a **random sample** drawn from that population.

Any numerical property of the population is a **parameter** and the corresponding value for the sample is a **statistic.** For instance the proportion of red tickets in the barrel is a parameter, whereas the proportion of red prize-winning tickets is a statistic.

Pervading the whole barrel model of statistics is an element of luck as to which tickets come out. This is represented by the theory of probability which you will meet in Lesson Three.

Below and on the next page is an example of some data which analyses groups of Statistics students over the years at a college and looks at the AS and A level results in statistics as well as the individual mock results. It is very difficult to be able to say what this data is saying by just looking at this table.

One of the roles of a statistician is to bring the data alive by perhaps illustrating the data where appropriate with diagrams (e.g pie charts and histograms) and also to calculate various numerical measures such as means and standard deviations as a way of summarizing the data.

In fact this data could be appropriately named "Statistics statistics". The first use of the word Statistics is the subject and the second use of the word is for facts and figures. We shall return to this data set at various other points in the course.

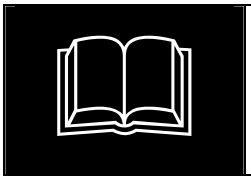
AS AND A/L STATISTICS RESULTS

Year	Student	Sex	A/S Mock	AS Result	A/L Mock	A/L Result
2001	1	m	64	B	44	B
2001	2	m	60	B	89	A
2001	3	f	57	C	32	C
2001	4	f	58	C	28	E
2001	5	m	60	C	66	B
2001	6	f	58	C	60	D
2001	7	f	72	B	60	A
2001	8	f	59	C	60	C
2001	9	m	79	C	60	B
2001	10	m	85	B	64	A
2002	11	f	23	C	47	N
2002	12	f	64	B	61	C
2002	13	m	36	B	82	A
2002	15	m	53	C	65	C
2002	16	m	53	A	85	B
2002	17	m	56	B	65	C
2002	18	m	58	A	65	B
2003	19	m	69	B	56	C
2003	20	m	76	A	75	A

Year	Student	Sex	A/S Mock	AS Result	A/L Mock	A/L Result
2003	22	m	91	A	94	A
2003	23	f	67	D	50	E
2003	24	f	69	B	67	B
2003	25	m	56	A	67	B
2003	26	m	50	E	56	N
2003	27	f	75	A	67	B
2003	28	m	71	A	67	A
2003	29	f	54	C	67	C
2003	30	m	48	C	50	E
2003	31	m	63	D	78	E
2004	32	f	74	A	61	A
2004	33	m	64	B	42	C
2004	34	m	46	C	23	B
2004	35	m	76	D	41	E
2004	36	m	70	B	58	E
2004	37	m	79	A	70	A
2004	38	m	65	C	70	C
2004	39	f	38	C	29	N
2004	40	f	43	C	40	C
2004	41	m	55	D	63	C
2004	42	m	42	C	33	D
2004	43	m	41	D	30	E
2005	44	f	68	B	74	D
2005	45	m	53	C	38	C
2005	46	f	81	A	77	A
2005	47	m	72	B	75	C
2005	48	m	39	E	58	E
2005	49	m	45	D	39	C
2005	50	f	70	B	65	B
2005	51	f	63	C	74	B
2005	52	m	68	A	61	B
2005	53	m	39	N	48	N
2005	54	m	62	C	60	D
2005	55	m	73	A	65	C
2005	56	m	47	C	34	D

Using Information Technology for Statistics

In the workplace a lot of practical statistics is done using specialist software. It is not part of this course to have to use any software. However if you have access to Excel and are proficient in its use, you can aid your understanding in some topics if you wish by using Excel.



Now try Ex 1B and 1C on pp4-5.

Self-Assessment Test One

1. What is the difference between quantitative and qualitative data?
2. What is the difference between discrete and continuous data?
3. What is the difference between a census and a sample?
4. What is the difference between a parameter and a statistic?

Answers to self-assessment tests are to be found at the end of the module.