DESCRIPTIVE STATISTICS - FREQUENCY DISTRIBUTIONS AND AVERAGES:

Inferential and Descriptive Statistics:

There are four main stages in doing research: designing a study, collecting the data, obtaining descriptive statistics and perhaps performing some inferential statistics.

Statistics are a set of tools for obtaining insight into a psychological phenomenon. *Descriptive* statistics summarise the data, making clear any trends, patterns etc. which may be lurking within them; they consist of visual displays such as graphs, and summary statistics such as means. *Inferential* statistics attempt to make inferences about the parent population on the basis of the limited samples actually obtained. The term inferential statistics is usually reserved for the various statistical tests used for comparing two or more groups of subjects within an experiment, etc. (Although in effect, most of statistics in psychology involves a process of trying to make inferences about humanity on the basis of a few undergraduate subjects!)

Samples and Populations:

In statistics, the terms ***sample*** and ***population*** have special meanings. A sample is a limited subset from a population, and a population is the entire set of things under consideration. What constitutes a population will vary according to what it is you are trying to make extrapolations to: usually the population to which we are trying to generalise from our sample consists of the entire human race. However, if you are interested only in making claims about the behaviour of three-toed Amazonian sloths, then you might be able to get access to the entire population of sloths if there weren't very many of them in existence. Normally, we have to obtain data from a limited sample of subjects because it is impossible to obtain data from the entire population to which they belong.

Types of Data:

It's important to know what kind of data you are dealing with, as certain statistics can only be used validly with certain kinds of data.

(a) Nominal (categorical) data:

Numbers are used only as *names* for categories. Therefore they are not really acting as numbers, but just serving as labels. If the numbers can be replaced by letters or names, then you are dealing with nominal data. So, if you had four groups, you could label them "1", "2", "3" and "4", but it would mean just the same as if they were entitled "A", "B", C, and "D". You cannot do any statistics or arithmetic on nominal data, in just the same way as you cannot add together A and B. The numbers on footballers' jerseys are the classic example of nominal data: adding together players at positions 1 and 2 does not give you the player at position 3!

(b) Ordinal (rank) data:

Things can be *ordered* in terms of some property such as size, length, speed, time etc. However, successive points on the scale are not necessarily spaced equally apart. To rank Grand Prix racers in terms of whether they finish first, second, third, and so on, is to measure them on an ordinal scale: all that you know for certain is the order of finishing - the difference between first and second-placed finishers may be quite different from the difference between those placed second and third. In psychology, most attitude scales (e.g., seven-point scales which run from "1" for "completely disagree" through to "7" for "completely agree") are making measurements on an ordinal scale.

(c) Interval data:

These are measured on a scale on which measurements are spaced at equal intervals, but on which there is no true zero point (although there may be a point on the scale which is arbitrarily named "zero"). The classic examples are the Fahrenheit and Centigrade temperature scales. On the Centigrade scale, a rise of "one degree" represents an increase in heat by the same amount, whether one is talking about the shift from 5 degrees to 6 degrees, or from 90 degrees to 91 degrees. However, because the zero point on the scale is arbitrary, one cannot make any statements about ratios of quantities on the scale: thus it is not permissible to say that 26 degrees is twice 13 degrees or half of 52 degrees.

(d) Ratio data:

This scale is the same as the interval scale, except that there is a true zero on the scale, representing a compete absence of the thing in question which is being measured. Measurements of height, weight, length, time are all measurements on ratio scales. With this kind of scale, the intervals are equally spaced and we can make meaningful statements about the ratios of quantities measured. Thus, an increase in length between 9 inches and 10 inches is the same as an increase in length between 90 inches and 91 inches; and additionally, we can say that 6 inches is twice 3 inches, and half of 12 inches.

The concept of levels of measurement may seem rather abstract and confusing now, but it will prove important in helping us to decide which statistics are appropriate for our data. For example, some statistical tests can only be used with nominal data, and others can only be used with interval or ratio data.

Summarising the data, by using Frequency Distributions:

The normal consequence of doing a psychology experiment is to end up with a score or set of scores for each of the individuals that took part in the experiment. The first thing we need to do is to summarise these data in some way, so that we can see the "wood" for the "trees" - in other words, so that we are more likely to spot any trends or effects that are lurking within all those numbers. Imagine we were interested in how well psychology students coped with doing a statistics exam. For each student, we might have their exam score. We might want the answers to various questions: for example, how well people were doing generally; what the highest and lowest scores were; within what range the bulk of the scores fell; etc. Presented simply as a table of unprocessed numbers, answering questions like this would be difficult. One of the goals of statistics is to make it easier to see what's going on in your data. So, the first step might be to construct a ***frequency distribution*** of the scores: we could simply count how many people obtained each of the scores that it was possible to obtain.

This is an improvement over our original unordered data-set, but things could be better: we are still bogged down in the details of the data. To get a better picture of the overall trends that are present, we could produce a ***grouped frequency distribution*** of the scores: we could construct a number of categories, each of which corresponded to a range of scores. We could then count how many scores fell into each of these categories, or ***class intervals***. Thus, instead of recording that four subjects scored 52, two subjects scored 53 and one subject scored 54, we could summarise these data by noting that seven subjects scored between 52 and 54. The number of class intervals is fairly arbitrary, but most statisticians suggest between 10 and 20 gives the best results: too few, and you lose too much information about what's going on in your set of numbers. Too many categories, and you are back to where you started, unable to see the overall picture because you are bogged down in the excessive details.

Sometimes we might want to compare frequency distributions which are based on different totals. Imagine that 50 people take a statistics exam one year, and 100 take it the next year, but in both years 25 people fail. 25 failures out of 50 is a lot more serious than 25 failures out of 100. In the former case, half of the students have failed, whereas in the latter case only a quarter have failed! How can we compare frequency distributions when they are based on different totals? The solution is to turn our "raw" frequencies into ***relative frequencies***. To turn a raw frequency into a relative frequency, divide the raw frequency by the total number of cases, and then multiply by 100. Thus (25/50)*100 = 50%, and (25/100)*100 = 25%. By converting frequencies to relative frequencies in this way, we can more easily compare frequency distributions based on different totals. If you display data in the form of relative frequencies (i.e., as percentages of the total), ideally you should also show the raw frequencies, so that readers are aware of how many cases your percentages are based on. At the very least you should show the total number of cases, so that readers can work out for themselves how large a sample your percentages are based on. The reason for this is that it is easy to use relative frequencies to make your data look more impressive than they really are. For example, claiming that "75% of those interviewed said they would vote for Tony Blair" is impressive if we are talking about 75% of a large sample, but less impressive if 75% merely represents 3 out of a total of 4 people!

Summarising data with single numbers: measures of central tendency and dispersion:

Often we want to summarise our results in a more succinct form than is possible with tables and graphs; we want to talk about the "average" performance on a test.

Measures of central tendency:

These are single numbers which attempt to convey some impression of what constitutes "typical" performance. The three most commonly encountered are the mean, median and mode. Each has advantages and disadvantages as a summary description of our data.

(a) The Mode:

This is the score which occurs most frequently in a set of scores. Thus, for the set of scores 5, 6, 11, 22, 22, 96, 98 the mode is 22, since there are more instances of 22 than there are of any of the other scores.

Advantages of the mode:

(i) It is simple to calculate, and intuitively easy to understand.

(ii) It is the *only* measure of central tendency which can be used with nominal data.

Disadvantages of the mode:

Unfortunately, the disadvantages outweigh these benefits for any serious statistical use.

(i) The mode may easily be unrepresentative of the bulk of the data, and so produce a misleading picture. Imagine if we had the following set of scores: 3, 4, 4, 5, 6, 7, 8, 8, 96, 96, 96. Here the mode is 96 - but *most* of the scores are low numbers, and so 96 is unrepresentative of them.

(ii) There may be more than one mode in a set of scores. For example, in the set of scores 3,3,3,4,4,4,6,6,6, there are three modes!

(iii) If data are grouped into class intervals, the mode is very sensitive to the size and number of class intervals used; it can easily be made to "jump around" by varying the limits of the class intervals.

(b) The Median:

This is the middle score of a set of scores, when those scores are arranged in order of size. Thus the median of 3, 4, 4, 5, 6, 7, 8, 8, 96, 96, 96 is 7. If you have an even number of scores, there is no middle score; in that case, the median is the average of the middle two scores. Thus, if we had the scores 3, 4, 4, 5, 6, 7, 8, 8, 96, 96, 96, 96, the median would be (7+8)/2 = 7.5.

Advantages of the median:

(i) It is resistant to the distorting effects of extreme high or low scores: notice that the median in the examples above was relatively unaffected by the few very high scores.

(ii) It can be used with ordinal, interval or ratio data; it cannot be used with nominal data, since categories have no numerical order.

Disadvantages of the median:

(i) It only takes account of the relative ranks of scores with respect to each other; it takes no account of the actual numerical values of the numbers themselves, which is rather wasteful if the data are on an interval or ratio scale.

(ii) It is more susceptible to sampling fluctuations than the mean (see below).

(iii) It is less mathematically useful than the mean.

(c) The Mean:

To find the mean of a set of scores, add them all together and then divide this total by the number of scores. In statistics notation:

$$\overline{X} = \frac{\sum X}{N}$$

Advantages of the mean:

(i) It is the only measure of central tendency that uses the information from every single score.

(ii) It has certain mathematical advantages; it is very common in statistical formulae, in one form or another.

(iii) It is the measure which is most resistant to sampling fluctuation. If you take repeated samples of scores from a population of scores, you will find that each sample is somewhat different from the next. If you work out each sample's mean, median and mode, it turns out that the mean is the measure which will vary the least from sample to sample. Since we usually want to make extrapolations from samples to populations, this is important.

Disadvantages of the mean:

(i) The mean is susceptible to distortion from extreme scores (e.g., "mean income" in the U.K. is a highly misleading statistic, because the few millionaires that contribute to this mean have a disproportionate effect, biasing the mean upwards from what it would otherwise be).Just one or two high or low scores can seriously distort the mean.

(ii) The mean can only be used with interval or ratio data; it cannot be used with ordinal or nominal data.

Measures of Dispersion or Spread:

A complete description of a set of data requires a measure of how much the scores vary around our chosen measure of central tendency.

(a) The Range:

This is the difference between the highest and lowest scores. (i.e., range = highest - lowest).

Advantages of the range:

It is quick and easy to calculate, and easy to understand.

Disadvantages of the range:

(i) It is unduly susceptible to extreme scores at either end. Thus the range of the set of scores 3, 4, 4, 5, 100 is (100-3) = 97. Replace the 100 with a 5, and the range is now (5-3)=2! variation in just one score can make an enormous difference to the range.

(ii) Because the range uses only the two extreme scores, it conveys no information about the spread of scores between these two points. Very different distributions of scores might have the same range. Thus 2, 2, 2, 2, 2, 20 and 2, 20, 20, 20, 20, 20 both have exactly the same range (18), but they are clearly very different sets of scores!

(b) The Semi-Interquartile Range:

To calculate the SIQR, arrange the scores in order of magnitude (as with the median). Find the score which is a quarter of the way from the bottom, and the score which is a quarter of the way from the top. These two scores are called the lower and upper quartiles respectively. Subtract the lower quartile from the upper quartile, and the result is the semi-interquartile range.

Advantages of the SIQR:

It is less susceptible than the range to distortion from extreme scores.

Disadvantages of the SIQR:

It still uses the information from only a couple of the scores in the distribution of scores.

(c) The Standard Deviation:

This is effectively an "average difference from the mean", which commonly accompanies the mean. The bigger the s.d., the more the scores differ from the mean and between themselves, and the less satisfactory the mean is as a summary of the data.

To calculate the standard deviation, you do the following:

(i) find the mean of all the scores;

(ii) subtract the mean from each score (keeping a note of the sign of the bit that remains each time);

(iii) square each of these "deviation scores";

(iv) find the mean of these scores (i.e., add them all up and divide by the total number of scores). This mean is called the **variance**.

(v) Find the square root of the variance; this is called the standard deviation.

Two steps in this sequence (iii and v) require explanation. The reason we square the differences in step (iii) is in order to get rid of the minus signs (anything squared, whether positive or negative, becomes a positive number). Some differences from the mean will be positive, and some will be negative. If we didn't get rid of the differences' signs, and simply added them all together, the pluses and minuses would all cancel out and we would always get zero! This squaring is therefore just a piece of arithmetical trickery to get rid of the minus signs. The reason we take the square root of the variance in step (v) is that the variance is okay, but due to the squaring trick, it's not in the same units as our original mean; it is nicer to have a measure of the spread of the scores which is in the same units as our mean, and this is where the standard deviation comes in.

This is what the formula looks like in statistical notation:

$$s = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{n}}$$

(but see below, in the section entitled "complications.." for further details about this formula).

Advantages of the standard deviation:

Like the mean, the s.d. uses the information from every score.

Disadvantages of the standard deviation:

It can only be used with interval or ratio data.

Complications with using means and standard deviations:

The distinction between *samples* and *populations* needs to be taken into account when considering means and standard deviations. There are three conditions which can occur:

(a) when the mean and s.d. are intended to describe only the sample on which they are based (i.e., they are used purely as descriptive sample statistics);

(b) when the mean and s.d. are intended to describe the entire population (and *every* member of the population can be measured - very rare in psychology);

(c) when the mean and s.d. are obtained from a sample, but one wishes to use them as *estimates* of the mean and s.d. of the parent population (probably the most common situation in psychology).

It can be shown that the best estimate of the population mean is the sample mean; the sample mean is said to be an *unbiased estimate* of the population mean. Consequently, exactly the same formula can be used for the mean, in all three of the circumstances just listed.

Unfortunately, the same is not true for the standard deviation; it can be shown that the sample s.d. tends to *underestimate* the size of the population s.d. So, if you were to try to use your limited sample from a population in order to make inferences about the characteristics of that population, your sample s.d. would tend to be a bit smaller than the "true" s.d. of the parent population. To correct for this underestimation, we simply add a small bodge to the s.d. formula: instead of dividing the sum of the squared deviations by n, the number of scores, we divide it by n-1 (the number of scores minus one). We thus divide the top part of the formula by a smaller number, and hence end up with a slightly larger s.d. than we would otherwise have obtained. In short, if you want to work out the s.d. for conditions (a) or (b), divide by n; but if you want to make inferences from your sample to the parent population from which it is believed to have originated (i.e., condition c), then you need to use the n-1 formula.

(a) sample s.d. (used purely as a decscription of your particular sample):

$$s = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{n}}$$

(b) sample s.d. (used as an estimate of the population s.d.):

$$s = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{n - 1}}$$

(c) population s.d. (if you could get scores from *every* member of the entire population):

$$\delta = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{N}}$$

Note that the first and last formulae are identical; only the notation differs, to show that in one case you are referring to a sample, and in the other case you are referring to a population. On Casio calculators that are capable of working out standard deviations, you will usually find two keys: one is labelled σ_n and the other is labelled σ_{n-1}. The first corresponds to the first formula above, and will give you the standard deviation of the scores that you have entered, purely as a description of that particular sample; the second is equivalent to the second formula above, and will therefore give you a standard deviation which would serve as an estimate of the standard deviation of the population of scores form which the sample is thought to have been taken. (If you want the calculator to work out the third formula, the population standard deviation itself, use the σ_n key).

Some useful terminology:

Mathematical notation is a language with which many of you are unfamiliar, and which most of you find off-putting. Although it will be kept to a minimum in this course, you will find knowledge of the following symbols useful:

Σ = an uppercase "sigma". This symbol means "add up the following".
δ = a lowercase "sigma". This symbol is often used to represent the population standard deviation.
μ = "mu". This symbol is often used to represent the population mean.
$\overline{X}$ = "x bar", usually used to denote a sample mean (the mean of the scores in sample X, as opposed to Y, Z or any other samples of scores you might have).
s = standard deviation of your sample of scores.
N is the number of scores in a *population*.
n is the number of scores in a *sample*.

The following two formulae for calculating the mean look different because of the notation, but they involve the same operations - it's just that one is for a sample mean and the other is for a population mean:

$$\text{population mean : } \quad \mu = \frac{\sum X}{N}$$

$$\text{sample mean: } \overline{X} = \frac{\sum X}{n}$$

You might find it useful to know that arithmetic operations are performed in the order: stuff in Brackets first; then Division; Multiplication; Addition; Subtraction.

Note that ΣX^2 and $(\Sigma X)^2$ mean completely different things. The former means "square each value of X *first*, and *then* add together all of these squared values". The latter means "add together all of the X values, and *then* square this total" (i.e. you do the thing in the brackets first).

$\Sigma\left(X - \overline{X}\right)^2$ means "take each value of X in turn; subtract the mean of X from it, to get a difference score; square this difference score; and lastly, add together all of these squared difference scores".

You can see that one advantage of statistical notation is that it has the potential to save a lot of paper!