1

AN INTRODUCTION TO MEASURING THINGS

All of statistics is based upon numbers which represent ideas, as I suggested in the Preface. Those numbers are also measurements, taken from the world, or derived from experiments we have undertaken. Therefore, the building blocks of statistics are measurements. It is, therefore, important that you understand the very principles upon which measurements are taken.

Σ LEVELS OF MEASUREMENT

Whenever you measure something, you measure it in a particular way. Some measurements are more sensitive than others, and this is very important for us to understand because it underpins the statistical procedures that we employ. Some data come in a particular form, and that's that. However, occasionally we can dictate the nature of the data depending upon the measurements we take. Therefore, it is crucial that we are familiar with different types of measurement and different types of data, primarily because we can, in theory, change the very nature of what we are studying if we measure and test in different ways.

You are probably confused by that statement, so let us work backwards, and then you can read it again later, and you'll see what it means.

There are four basic types of data, or levels of measurement, as we call them. These are called *nominal*, *ordinal*, *interval* and *ratio*. Data can only be of one type, although statisticians and psychologists might actually argue sometimes about which category certain data fall into.

Nominal data are in the forms of categories or names of things. In fact, sometimes nominal data are referred to as categorical. Nominal data are generally seen as the least informative, and nominal measurements are the least sensitive of those that we can take. Nominal data are essentially in the form of labels. You can't perform standard mathematical operations

on them. For example, bus routes often have numbers attached to them. However, although the Number 10 takes the same route every day, it doesn't actually go half as far as the Number 20. The Number 15 is not 50% more expensive to travel on than the Number 10. The Number 658 isn't the best bus of all, even if it's the highest number that the bus company uses. The numbers are only codes. Other common examples include the numbers on football players' shirts, postal or zip codes, and people's names. I can count up how many Sarahs in a certain population, and I can count how many Janes there are, but I can't divide Sarahs by Janes, and I can't work out the average name by adding them up and dividing by the number of numbers. A common example of this type or error is when students ask a statistical package to automatically work out the average of a set of scores. If your scores are simply labels for different groups of people, this is meaningless. That is, imagine that you have coded men as 1 and women as 2. If you work out the averages of your scores, you might discover that the mean sex of participants is 1.4. Now what on earth does that mean? The mean sex of a participant in your sample is, it seems, halfway between male and female, but ever so slightly more male. There is an old joke that runs like this: if I put my head in the refrigerator and my feet in a bowl of hot water, my temperature *on average* is just about right. No? A friend and I are out walking when we come to a junction where we can go one of three ways. There is a path to the left, one to the right, and one in the middle. Unbeknown to us, only the middle path leads where we want to go. But, I take the left path, he takes the right, and, *statistically*, that means that *on average* we took the middle path and got to our intended destination. No? Right: I think you get the point here about nominal data. This type of data is very basic, and there is not much you can do with it.

If we put the types of data in order of their sensitivity and usefulness, then the next one in line is called ordinal data. This is pretty much what it sounds like. Ordinal data tell you what order things are in, but not a lot more. Therefore, world chess rankings are ordinal. The order of finishers in a race (without their race times) is also ordinal. Now, you can probably see why this kind of data is better than that which is nominal, but still carries some strong limits. After all, knowing what order things are in is good, but knowing how far apart they are is better. I might beat you in a race today, so that I finish first and you finish second. If I beat you by only one second, you would be right to suggest that this is not necessarily very meaningful, and that on another day you might be able to finish first instead. However, if I beat you by minutes, or even hours, that would probably give you a good idea of who was the better runner. Ordinal data fall short of giving us this crucial information. This leads us onto interval data.

Interval data are measured in such a way that one does know the distance (if you like) between points on our scale. We not only know the order of the finishers of the race now, but also their actual race times, which allows us much more scope to understand the relationships between things we can measure in such a way. Many data used in psychology are interval in nature. The classic example of this kind of data is temperature in degrees Celsius. The difference between 10 and 20 is the same amount as the difference between 20 and 30, and so on. However, there is one characteristic missing, which sets interval data apart from ratio data. The Celsius scale does not have a true zero. Zero degrees does not mean 'no temperature'. It is simply a marker. For convenience, we chose the freezing point of water to correspond with this point on the scale. But, we have to bear in mind that there is still a temperature going on! Some people in the past have asked me about 'absolute zero' on the Kelvin scale of temperature. This corresponds to –273 degrees centigrade. It's basically the coldest it is possible for something to be. Now, the thing that is important here is that absolute zero is a true zero, because the Kelvin scale is an index of heat energy present in a substance. No heat energy is zero degrees Kelvin. There really is no heat energy present at all, which is why it can't get any colder. Kelvin, therefore, is a *ratio* level of measurement, which we can now turn to.

Ratio data are the 'best' kind of data of all, at least in terms of sensitivity in statistical testing. Ratio data are interval data measured on a scale with a true zero. The number of items recalled in a memory test is ratio data; it is possible to get zero correct, if you don't remember anything. Furthermore, when you get 2 correct, it is half as much as someone who gets 4. There is a perfect mathematical relationship between the items. For reasons I won't go into, this fits very neatly into a whole bunch of statistical tests we have invented over the years. One nice and easy reason for you to ponder is that ratio data have a tendency to end up being normally distributed, at least more often than the other forms of data, some of which actually can't ever be! Therefore, given that we have developed our tests to compare distributions of scores with approximately normal shapes, ratio data are most likely to provide that.

Σ MEASURES OF CENTRAL TENDENCY

Staring at sets of numbers can, from time to time, be quite useful, but there's a limit to that usefulness. Also, the larger the sets become the more difficult it is to get any sense of their properties by 'eyeballing', as it is

called. Therefore, we need some way of characterising and summarising data, and that is why we have developed the descriptive statistics known as measures of central tendency. Why 'central tendency'? The answer is that, in this case, 'central' refers to average, in the sense that numbers tend to cluster, and that clustering we can call a 'centre'. Note that I said 'tend'... Hence central tendency. Measures of central tendency are simply ways of telling us what the numbers in a set are like, on average.

There are three main kinds of average, and although the mean tends to be used the most, the mode and the median are also handy at times. Sometimes, you cannot use a mean, for example because it completely misrepresents the data. We can examine the detail of three the types of measure now.

Mean

The mean is the kind of measure of central tendency that most people are referring to when they use the word 'average'. Simply, you summate all of the numbers in a set, and then divide by the number of numbers in the set. Mathematically, this takes all of the numbers into account, and there is a kind of 'tug-o'-war' that occurs in the formula. Numbers pull each other back and forth, but in the end the mean settles at the point where most of the values lie. A very high or very low number will not influence the mean greatly because the mean is normally found where the majority of values are to be found. The mean is best used when a set of numbers are normally distributed (that is, where most people score in the middle, and a few people score very highly or very low), so you should always check the frequency distribution before you automatically calculate the mean. If the distribution doesn't look like this, then it might be a good idea to avoid relying on the mean.

The mean should not be used when the data are categorical. It doesn't make sense to calculate the average sex of your participants, when you have labelled men as 1 and women as 2. You'd probably end up with a mean of somewhere around 1.5. Of course, it's nonsensical. What sex is 1.5? Additionally, don't calculate means when you appear to have a bimodal distribution, that is where there seem to be two peaks of scores (Figure 1.1). If there is a bunch of low scores and a bunch of high scores, with virtually nothing in the middle, the mean is not going to adequately depict the scores. What happens in this case is that the mean ends up being somewhere in the middle. Think about it: how can you have an average score which no-one actually achieves? It's a contradiction in terms. The

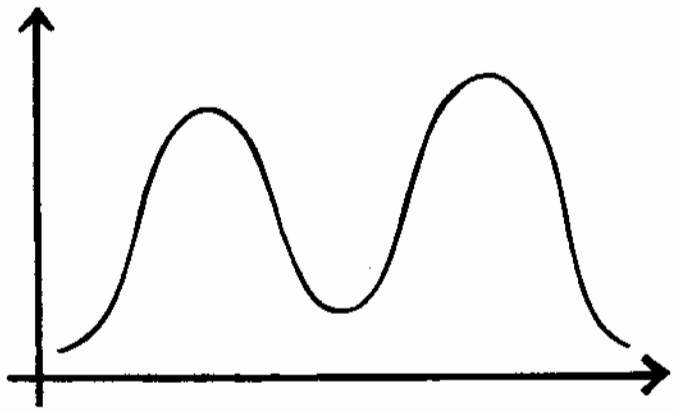

Figure 1.1 A largely bimodal distribution.

average must tell you what most scores are like, not what none or a few are.

Mode

A mode is used in a few rare circumstances. All you have to do to identify the mode is look for the most commonly occurring number in your set. Not only is this a good way to eyeball small sets of data, it also works well for categorical data. If you want to know which of three choices of steak sauce is the most popular, you can simply count up the numbers of people choosing each. The most commonly occurring is the most popular steak sauce, and precisely because it is the most popular sauce it also is what the average diner will choose, if you like. This is no more than common sense!

You can't use the mode if there are lots of decimals in your data, or if there are many different numbers and none the same. Imagine a dataset containing numbers to five decimal places, such as 0.00452. The same number might never occur twice, in which case there can't be a distinct mode. Either that, or in a dataset of 140 different numbers there are really 140 modes! Since the purpose of a measure of central tendency is to summarise the essential property of a set of numbers, having to list them all over again hardly works.

In fact, the mode is extremely limited. It also won't do the job when there are just a small number of identical data, and many other, different scores. In the following dataset, 3 is the mode, but it really doesn't represent the

set properly because most of the other figures are much higher: 2, 2, 3, 3, 3, 3, 38, 45, 46, 48, 50, 55, 65, 71, 71, 72, 73, 74, 75, 80, 99.

Median

The median is a very simple measure of central tendency to work out because it doesn't involve any calculation. All you have to do is put the numbers in order of magnitude and then find the one right in the middle. With an even number of scores, you have to choose halfway between the two scores nearest the centre. Nothing more, nothing less. This can be a very useful descriptive statistic, especially when you have some extreme scores or outliers which you don't want to trim out of your data, for whatever reason. The median is not affected by extreme scores because it is based upon finding whatever is in the middle of the set, making the tails of the distribution almost irrelevant.

For example, let us take these numbers: 1, 1, 5, 6, 8, 9, 10, 24, 25. The median is 8. Now, imagine what would happen if the lowest score was now 0 and the highest score was 11,475. That is, 0, 1, 5, 6, 8, 9, 10, 24, 11,475. The median is still 8! The means for the two sets vary massively, however. The first has a mean of 9.89, and the second a mean of 1282. Here's an example not only of the reason why medians are helpful statistics, but also why means of small samples can be very misleading, since extreme scores can really pull the mean away from where it might more sensibly lie. The mode in the first set of figures, by the way, is 1, since that number occurs twice, and the others only once each. So, for the second set we have no mode (or nine modes, depending on how you look at it), a mean of 1282 and a median of 8. Quite different, aren't they? Since the whole purpose of descriptive statistics are to describe the data, it isn't difficult to identify which is the most accurate at doing so.

Σ NORMAL DISTRIBUTION

When you collect data, it is possible to look at the scores you have in terms of a frequency distribution. A frequency distribution is simply a way of seeing how the data are spread out along the range of scores. If you gave people a visuo-spatial test of reasoning, it might be possible to score as little as zero, and as much as 100, for example. We might want to know just how many people score highly, or how many people score below average, and so on. A frequency distribution can reveal this. It is common to plot

such a distribution using a bar graph and categories of data, for example 0–10, 11–20, 21–30, and so on. In a frequency distribution, the number of people gaining scores in that particular band are shown by the height of a bar on the bar graph.

Frequency distributions can vary considerably in their shapes. The most basic form, and the most useless for statisticians and psychologists, is what we would call a monomial distribution. This is where everyone achieves exactly the same score. Sometimes, certain data are what we called binomially distributed. This time, people either get one score or they get another. Binomial distributions of data should be a warning sign that you are really dealing not with a continuous variable, but rather a categorical one. If everyone seems to score either high or low, and little else, then to all intents and purposes you should not consider the data to be spread across a scale.

Most scores that we collect in psychology are likely to be distributed normally. This is set by definition, since a normal distribution means a distribution which occurs as a norm, i.e. what normally happens. Normal distributions have a clearly defined shape. They look like a bell, which is why they are also called 'bell curves'. They have a hump in the middle, and small tails at each extreme, with scores trailing off steadily towards those 'tails'. What this means is quite simple: most people are found to be scoring in the middle range of collected data, a few people achieve very high scores, and a few people achieve very low scores. The average score is right in the middle of the hump, and this doesn't matter whether you use the mean, the mode, or the median as your measure of central tendency.

The possession of normally distributed data is an ideal situation for most statistical analyses. It is what we call an *assumption* of many statistical tests. You might legitimately ask why this is so important. Well, without boring or confusing you with the fine detail, suffice it to say that that most parametric, inferential, statistical tests are constructed to be fed normally distributed data. If your data aren't normally distributed, you will run up against problems. It's the same as cars which run on diesel fuel instead of petrol. If you want to know why you have to use diesel, you'll have to understand engineering, but most diesel vehicle owners simply accept that the engine was made that way and get on with the process of filling up with the right kind of fuel. A car with the wrong fuel in the tank is a bit like a statistical test fed the wrong kind of data. It won't work.

Normally distributed data are to be found all over the natural world. Height, weight, blood pressure, finger length, width and depth of rivers are all normally distributed. Most people are of 'average' height, and there are a small number of very tall and very short people. Most people are of

'average' weight, with a few very heavy and very light ones making up the tails of the distribution. The same goes for the other variables I mentioned above, and literally millions of others, including psychological ones such as vocabulary score, intelligence and problem-solving ability.

The Case of the Normal Distribution

When you have a perfectly normally distributed set of data, conforming exactly to the classic bell curve, something rather special happens to the measures of central tendency. They converge to the same point, which is the peak of the hump in the middle of the distribution, or, the vertical axis of symmetry, as you can see in Figure 1.2. That is, they are all the same. In a perfect normal distribution, the mean, mode and median are exactly the same number.

Now, this is something that is worth remembering for the future. As you'll see elsewhere in this book, and beyond, statisticians have a tendency to be a little obsessed with normal distributions. They base so much of what they do on them, and researchers almost pray for them to arise in their data. This is one of the reasons. It makes life so much easier when you don't have to make the choice between measures of central tendency. You don't have to know which one is best, or carefully examine the data to work it out, because they are all the same. Oh, and here's a little tip. One way to quickly check if a dataset is likely to be normally distributed is to get SPSS or some other software to calculate the three main measures of

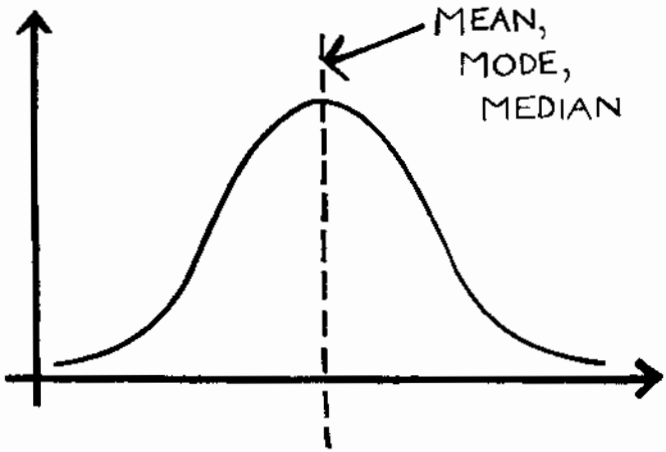

Figure 1.2 Measures of central tendency in a normal distribution.

central tendency for you, and to see how close to each other they are. If they are only a whisker away from each other, there's a very high chance that you have normally distributed data. Don't forget to plot a frequency graph, however, to check the shape visually.

There are other ways to test for normality. We can calculate coefficients which tell us about the size of the tails of the distribution in relation to the hump in the middle, how pointy or flat or asymmetrical the distribution is. If we do this, rather than just looking at the shape of the curve on a graph, we can be that bit more certain of what it is we are dealing with. In statistics, we almost always prefer a number over a picture. Pictures can tell us that something might be wrong, whereas a statistical test can tell us just how wrong, and how likely it is that we are right when we say it is wrong! (Sorry if you are confused. All I mean to convey to you is that just because something looks wrong doesn't mean that it is, and just because something appears right isn't enough to feel confident that it is.)

There are two main tests of normality, and these are the Kolmogorov–Smirnov and Shapiro–Wilk tests. SPSS will compute them very easily for you, and what they essentially do is compare your data with an hypothesised or 'idealised' normal distribution and look for a difference. They give you a probability value like other statistical tests, but this time you are looking for high probabilities rather than low ones. A value of P greater than 0.05 indicates that there is no difference between your data and a normal distribution. That's a good thing, because if there's no difference, that means that you have, to all intents and purposes, a normal distribution! However, if the value of P is less than 0.05, your data are significantly different from a normal set, which of course means that they are not normal.

One word of warning. When you have very small samples (less than about ten), tests of normality are generally very insensitive to violations of normality. So, rather ironically, tiny samples tend to pass the normality test when in fact they aren't normal, and it's almost impossible for them to be normal anyway because there aren't enough numbers present to make up a normal curve!

The problems that such tests can bring about make many statisticians avoid them altogether, choosing instead to use 'rules of thumb'. They differ on these, and there are no official regulations, as it were! Look at the statistics for skew and kurtosis that SPSS gives you, and look at the standard errors that come next to them. Compare each statistic with its standard error. If the statistic is more than twice the standard error you should think of this as indicating a problem with non-normality. Well, that's what Coolican (2004) suggests. However, some say that you needn't worry

about skew unless the statistic is greater than plus or minus one. Some say that as long as skewness is acceptable you needn't worry too much about kurtosis at all. Furthermore, most of this debate is about enabling the statistician to choose between different tests to use: the parametric and non-parametric ones (which you'll come across later in the book). Therefore, it is not uncommon for statisticians simply to run both tests on the data and then compare them. If they don't really differ, the chances are you have a result you can be sure of.