Chapter 8: Summarising Your Data – Measures of Central Tendency and Dispersion

Progress Questions

1. (a) There is a very significant difference between the two measure of *mean* and *median* and it is this that determines the value of each measure. The median represents the middle value in a range of values. In other words it is the mid-point value and half the number of values lie above and below this point. The median is therefore determined by its ranking within the range. The mean, on the other hand is a calculated value and is determined not by ranking but by the sum of all the values and the number of values.

    1. The *standard deviation* and *interquartile range* are related to the measures of central tendency described in (a). Both represent the distribution of values around a measure of central tendency. The *interquartile range* is probably easier to understand so let's look at it first. The median represents the half way value in the range of values as described above. Another way of describing the median is by using the term *second quartile*. There are three *quartiles* in regular use and they represent discrete groups of 25% of values. The *first quartile* represents the value below which 25% of cases fall, the median is 50% and the *third quartile* is 75%. The interquartile range represents the middle 50% of cases by their lower and upper values based on the first and third quartile. It is a very useful measure because it is easy to understand and calculate. Its value, like that of the median is not affected by anything other than the ranking of cases, unlike *standard deviation* that does make assumptions about the distribution of data (i.e. it assumes the distribution conforms to *normal distribution* pattern). Interquartile range can therefore be used with any data set irrespective of distribution. Standard deviation cannot be used like this. It represents distribution of data in a rather different way, but if the data distribution does conform to the normal curve, it is a very powerful statistical tool and is used to express confidence limits through the use of measures such as *standard error of measurement*. The standard deviation represents the range either side of the mean that captures a particular proportion of cases.

2. Assuming your sample has been drawn from the population as a whole using probability sampling, the mean of 70 is calculated on the basis of the data you collect. However, if you draw a second sample from the popluation using probability sampling techniques, these will be different individuals (although it is clearly possible that someone from the previous sample is drawn again). Because the sample is made up of different people, the mean score of that sample may be different simply through natural variation. This will be particularly likely if the sample is actually a very small proportion of the population as a whole. The *standard error* is a measure that recognises this and attempts to give a reasonable estimate of the range in which the *real mean score* of the *entire population* lies. A SE of 2 means that the actual population mean has a 68% probability of lying within a range of ±2 from the sample mean – i.e. between 68 and 72. There is a 95% probability of it lying between ±4 (i.e. 2 x SE from the sample mean) – i.e between 66 and 74. There is a 99% probability of the population mean lying between 64 and 76 (i.e. ±3 x SE of the sample mean).

3. The answer to this question was hinted at in the answer to question 1. The principle can be demonstrated with a simple example. Take the following two sets of numbers:

SET A:	2	3	4	5	6	Median = 4,	Mean = 4
SET B:	2	3	4	5	50	Median = 4,	Mean = 12.8

The effect of the extreme value, 50, is to elevate the value of the mean in a way that makes it unrepresentative of the main body of values. The interquartile range is similarly less sensitive to these extreme values because it is based on ranking rather than calculation.

4. ***Standard error of measurement*** is based on the assumption that the sampling process introduces errors – sampling errors. The statistic is the theoretical standard deviation of the means of many samples taken from the population. Therefore, if you calculate the mean value of any interval measure (e.g. age) you need also to determine the standard error of measurement (a simple formula for this is given in the chapter) so that you can generate confidence limits for the real mean. Returning to the question, as the mean age of the sample is 25 and the standard error 1.5 years, then we can be 95% confident that the population mean actually lies within ±2 standard errors – i.e. 25 ±3 or between 22 and 28 years.

5. Remember the formula for calculating standard error is

$$S.E. = \text{(standard deviation)} / \sqrt{\text{(sample size)}}$$

On the basis of this:

$$S.E. = (3) / \sqrt{(900)}$$

$$S.E. = 3/30 = 0.1$$

In other words, there is a 99% likelihood that the mean consumption of chocolate bars each month within the population as a whole lies between ±0.3 (i.e. 3 standard errors) above or below the sample mean.