DESCRIPTIVE STATISTICS

Introduction to Descriptive Statistics

In attempting to answer most research questions, it is common to collect large amounts of data that often represent individual participant's scores on variables of interest. For example, in order to be able to help in the recovery of patients who have undergone heart surgery, exercise researchers might be interested in knowing how important factors such as age, previous exercise habits and self-confidence on adherence to a cardiac rehabilitation programme. To study the influence of these variables on adherence we would first have to obtain a measurement of each participant's age, previous exercise habits and self-confidence, along with measurements of adherence to the rehabilitation programme. But rather than deal with every individual score on the variables of interest, data reduction techniques would be utilised to refine the data and reduce it to a manageable form. This allows us to make summary statements about the sample as a whole, rather than refer to each participant's scores. We can therefore see that descriptive statistics aim to provide a shorthand description of large amounts of data.

To illustrate what this means let's consider a rather less complex exercise-related example. A researcher who was interested in the exercise behaviour of students at a local university distributed a questionnaire to a sample of students who regularly used the university health club to determine their patterns of use.

The data in Table 2.1 represents part of the information collected by the researcher and it illustrates the number of times per month that a sample of students participated in aerobic exercise classes at the student health club. A sample of 19 students provided the researcher with information on their exercise behaviour.

In statistics the number of observations is denoted by the letter n. Therefore in this case,

$$n = 19$$

As it is displayed in this raw form, it is difficult to comprehend any features of the data and answer any questions about how often people engage in aerobic exercise. One useful process is to arrange the observations in order from the smallest to the largest. This process will produce what is termed an **array** (Table 2.2).

It can now readily be seen from this arrangement that the scores in this data set range between 1 and 13. Therefore, it can be concluded that the minimum amount

Table 2.1 **Raw data**

7	4	5	13	8	10	8
7	3	5	9	8	10	
6	1	6	12	11	9	

Table 2.2 **Array of data**

1	5	6	8	9	10	13
3	5	7	8	9	11	
4	6	7	8	10	12	

of exercise a student engages in is one aerobic session per month, and the maximum number of sessions undertaken is 13. This range equals the highest score minus the lowest score plus one. The reason for the plus one is because this range is inclusive of both the highest and lowest scores:

$$\begin{aligned}\text{Range of data} &= 13 - 1 + 1\\ &= 13\end{aligned}$$

Knowing the number of observations ($n = 19$) and having a rough measure of the dispersion of these scores (range = 13) are useful but they are not enough to provide an accurate shorthand description of the data. For example, they do not provide any indication of how the 19 scores are distributed across this range of 13. Furthermore, it is not known whether most students exercise once or twice a month with only an odd few exercising 12 or 13 times a month, or whether most exercise 12 or 13 times a month with only an indolent few exercising once or twice per month. To find out how the frequency of exercise is distributed across this range of 13 the data needs to be classified into what is known as a **frequency distribution**.

To undertake this task the range of scores must be divided into a number of different classes ranging from low attendance to high attendance. Then, the frequency of scores falling within each specific level of attendance can be counted in order to produce a frequency distribution.

Frequency Distribution

If data are to be divided into a number of classes then the first thing to do is to decide how many classes there should be. There is in fact no rule to guide this process. However, it ought to be clear that the more classes there are, the narrower the width of each class. In this example we have decided that the number of classes is to be 7. The number of classes is denoted by the letter k:

$$\begin{aligned}k &= 7\\ \text{Range of data} &= 13\\ \text{Therefore the width of each class} &= \frac{13}{7}\\ &= 1.86\end{aligned}$$

The resulting class width of 1.86 is an awkward number to deal with if the frequency distribution is to be worked out by hand. Fortunately, just as the number of classes is arbitrary, so also is the point at which to begin the lowest class and the point at which to end the last class. For example, instead of starting at 1, the lowest score, and ending at 13, the largest score, we could alter this so that the starting point is 1 and the end point is 14. This would ensure the class width becomes a whole number. To see the consequences of this decision, let us repeat the above process:

$$\begin{aligned}k &= 7\\ \text{Range of data} &= 14 - 1 + 1\\ &= 14\\ \text{Therefore width of each class} &= \frac{14}{7}\\ &= 2\end{aligned}$$

As you can see, the effect has been to increase the class width from 1.86 to 2. A class width of 2 is much easier to work with if this exercise is to be calculated by hand.

In the above example seven classes have been created, each with an equal width of two participating points. The first class starts at 1, known as the lower class limit, and ends at 2, known as the upper class limit. If you now consider the array of data in Table 2.2 only one score, which in this case is 1, falls into this first class. Let's now consider the second class. This will have a lower class limit of 3 and an upper class limit of 4. Looking once again at the array of data in Table 2.2 it can be seen that two scores, in this case 3 and 4, fall into this class. This process can be repeated for all seven classes and tabulated in the form of the frequency distribution given in Table 2.3.

Table 2.3 **Frequency distribution of student participation in aerobic exercises per month**

Class	Class limits	Class frequency (f)
1	1–2	1
2	3–4	2
3	5–6	4
4	7–8	5
5	9–10	4
6	11–12	2
7	13–14	1
		Total (Σ) = 19

If the only information available to the reader was this table, then it would be possible to conclude that one student attended one or two aerobic classes per month, two students attended three or four classes per month and so on.

The term **class limits** in the table refers to the smallest and largest value that are permitted in each class. The point half-way between the class limits is termed the **class mark**. So, for example, the class mark of the first class would be 1.5 and of the seventh class 13.5.

Sometimes it is more convenient to define a class not by its upper and lower limits but by its boundaries, that is where it begins and ends. Looking back at our frequency distribution we can see that the first class's upper limit is 2 and the second class's lower limit is 3. Therefore, the boundary between these classes must be between 2 and 3. The definition of where this boundary is does not matter as long as we are consistent. The easiest thing would be to put it half-way between the two, that is at 2.5. This point is known as the **class boundary**.

Graphical Representations

The distribution of the data across the classes can often be more clearly seen from a graphical presentation. A graphical representation of a frequency distribution is known as a **histogram**. A histogram of the above frequency distribution is shown in Figure 2.1.

If a line is drawn that joins all the class marks a **frequency polygon** will be constructed.

As with the frequency distribution this histogram illustrates that one student attended between one and two aerobic classes per month, two students attended between three and four classes per month and so on.

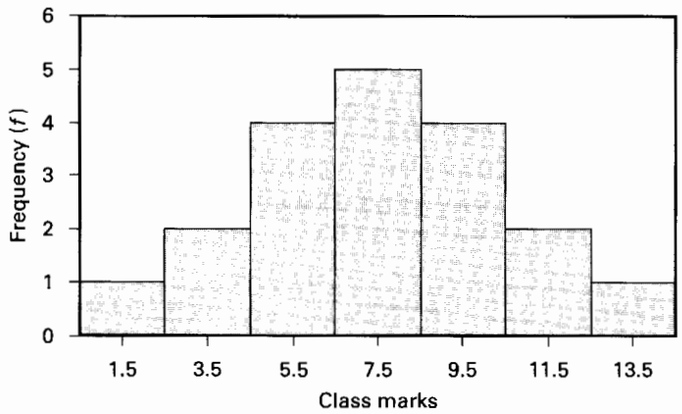

Figure 2.1 **Histogram of student participation in aerobic classes**

Cumulative Frequency Curves

Sometimes it is useful to illustrate the number of observations that are less than certain specified values. For example, it might be useful for the management of the student health club to know how many members attended less than 11 classes per month so that information on exercise adherence could be mailed to them, thereby encouraging them to continue with a regular exercise programme. The cumulative frequency values are the number of scores falling below the lower class limits, beginning with the second class. To see how this works let's use our previous array of data on attendance at aerobic classes shown in Table 2.2.

Table 2.4 shows that only one student attended fewer than three sessions per month, three students attended less than five times a month and so on.

Table 2.4 **Cumulative frequency distribution**

Lower class limit	Cumulative frequency (*F*)
<3	1
<5	3
<7	7
<9	12
<11	16
<13	18
<15	19

Just as a frequency distribution can be represented graphically by a histogram so the cumulative frequency distribution shown in Table 2.2 can be illustrated by a cumulative frequency curve (Figure 2.2). Note that the cumulative frequency (denoted by *F*) is on the vertical axis and the lower class limits are on the horizontal axis.

From this graph it can be seen that approximately 16 students attended less than 11 classes per month. This may indicate to the management of the student health club that advice on exercise behaviour and adherence should be mailed to 16 of the 19 students.

Cumulative frequency curves can also be used to illustrate rates of change. For example, imagine you were interested in comparing how effective two training schedules were at promoting skill development. If the frequency with which participants successfully displayed the activity was recorded and plotted against time, a cumulative record of their behaviour would be established. In this case, the slope of the two lines on the graph would show how rapidly the skill was developed by each of the techniques. This approach has often been adopted in learning in animals where the behaviour is automatically recorded on what is known as a cumulative frequency recorder.

A related approach is employed in leisure studies to create a graph that is similar in principle to what economists call a demand curve. For example, imagine you were interested in developing a health centre. After explaining the plan to members of the

public, you then ask them whether they would use the facility and how much they would be prepared to pay to use it. If you then construct a frequency distribution and plot a cumulative frequency curve based on the percentage values of each class frequency you will get something similar to the graph in Figure 2.3.

From this graph it can be seen how the percentage of the public who claim they will use the health centre decreases as the entry fee is increased. Multiplying the percentage who claim they will use the facility by the charge at various points would provide a calculation of the entrance fee that would maximise income for the facility.

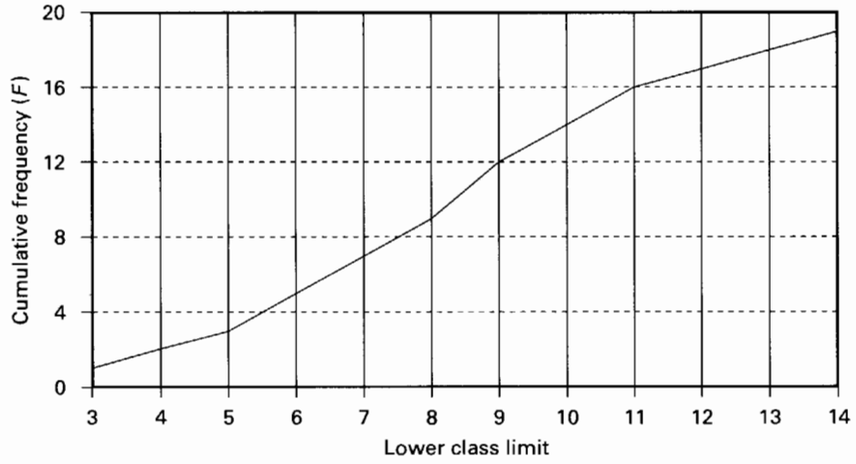

Figure 2.2 **Cumulative frequency curve**

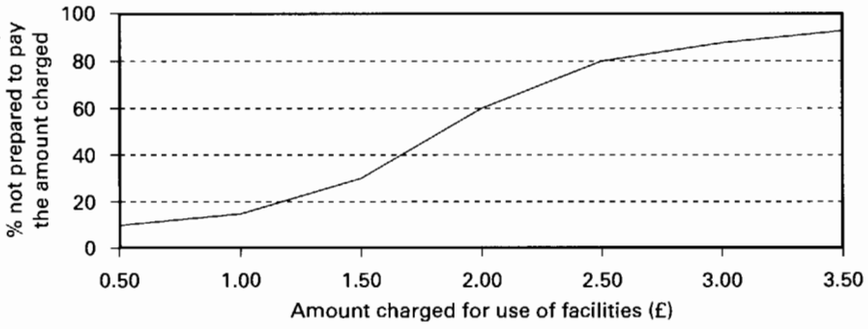

Figure 2.3 **Graph showing the percentage of a sample not prepared to pay various amounts for the use of health centre facilities**

Note: The area above the curved line represents the decreasing proportion of people who will use the health centre as the charge increases.

Measures of Central Tendency

In addition to knowing how the data is distributed, it is also extremely useful to be able to identify the typical score. Using the data in our student aerobic exercise example we might want to know the typical number of aerobic exercise sessions that students attend each month. To find such a value we refer to what are known as **measures of central tendency**, as these aim to provide a typical score around which the other scores are scattered.

Before discussing the student data in detail let's consider an additional data set that the researcher obtained from university staff attending aerobic exercise classes at the student health centre. The data in Table 2.5 represents part of the information that was collected and illustrates the number of times per month that a sample of university staff participated in aerobics classes.

Adopting the same procedure as before, the frequency distribution given in Table 2.6 can be constructed.

Table 2.5 **Array of data of the number of times staff participate in aerobic classes per month**

1	3	4	4	6	9	14
3	4	4	5	7	10	
3	4	4	5	8	12	

Table 2.6 **Frequency distribution of staff participation in aerobic exercises per month**

Class	Class limits	Class frequency (f)
1	1–2	1
2	3–4	9
3	5–6	3
4	7–8	2
5	9–10	2
6	11–12	1
7	13–14	1
		Total (Σ) = 19

The questions we now want to ask are first what is the typical number of times per month that students participate in aerobic classes and second what is the typical number of times per month that staff participate in aerobic classes?

To answer these questions we should note that there are three measures of central tendency that we might employ. These are:

1. the mode
2. the median
3. the mean.

The Mode

This is defined as the observation in the sample that occurs most frequently. If there are two scores that occur most frequently, then the sample is said to be bimodal. If more than two modes exist, then the data set is considered to be multi-modal. In the student data in Table 2.2 there is only one mode and this is the number 8 which occurs three times. Therefore, according to this measure, typically students attend eight classes per month.

For staff, on the other hand, the most frequent observation (from Table 2.5) is 4, which also occurs three times. This suggests that staff typically attend four sessions per month.

The Median

When the data is arranged from the lowest to the highest value, the median is the middle observation in the array. In both our student and staff data we have an odd number of observations ($n = 19$) so the median observation is the 10th observation.

$$\text{Median for student data} = 8$$
$$\text{Median for staff data} = 4$$

Thus for both staff and students, the corresponding modes and medians are the same. That is, both measures suggest that students typically attend eight classes per month and staff four classes per month.

The Mean

This is the sum of all the observations divided by the number of observations made:

$$\text{Mean} = \frac{\sum x}{n}$$

$$\text{Mean for students } (\bar{x}) = \frac{\text{sum of observations}}{n} = \frac{142}{19} = 7.47$$

$$\text{Mean for staff } (\bar{x}) = \frac{\text{sum of observations}}{n} = \frac{111}{19} = 5.84$$

While the mean for the students is very similar to the median and mode, that is 7.47 compared with 8, the corresponding figures for the staff are 5.84 and 4. Given that one measure suggests that staff typically attend nearly six classes per month and the other two measures suggest they attend four classes per month, which measure are we to believe? That is, which provides a truly typical score?

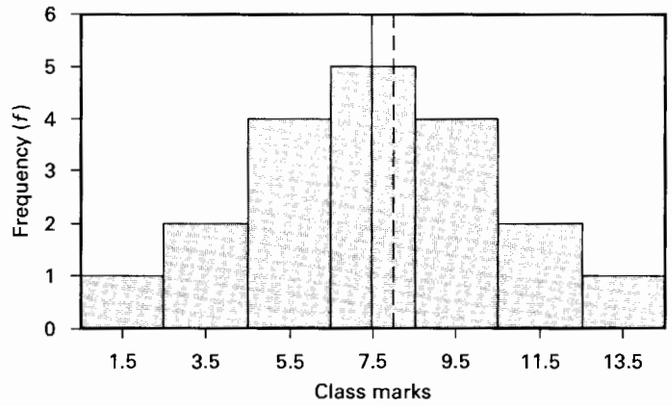

Figure 2.4 **Histogram of student participation displaying the mean, median and mode values**

To answer this question it is important to appreciate that the mean is sensitive to every score in the array, whereas the median and the mode are not. To illustrate this imagine that the score of 14 in the staff data was changed to 140; this would suggest that one faculty member was probably obsessive about exercise. It would also affect the mean of the array so that the mean would become 237/19 = 12.47. However, the median would remain at 4, because it is still the middle score. Therefore, it appears that while the mean becomes distorted by this extreme value, the median remains unaffected.

This example illustrates that how the data is distributed over the range of values will influence the three measures of central tendency in different ways. For this reason an understanding of a concept termed **skewness** is important.

Skewed Distributions

If the mode, median and mean for the student data are drawn on the histogram in Figure 2.1, it can be noted that they almost fall in the same place on the histogram, that is close to 8. If the distribution had been truly **symmetrical**, that is if the scores had been evenly distributed around a mid-point, then the mean, median and mode would all have had exactly the same value and would fall at the mid-point of the distribution.

Note that in Figure 2.4 the mean is represented by the solid vertical line and the median and mode by the dashed line. If this procedure is now repeated for the staff data it can be seen that the measures of central tendency do not fall at the same point because the mean is located to the right of the other two measures. It can also be seen that, whilst the figure representing the student data looks almost symmetrical, the figure representing the staff data (Figure 2.5) is clearly not symmetrical.

Distributions that are not symmetrical are termed **skewed**. They may be skewed to the left (negative) or to the right (positive). The direction of the skewness is determined

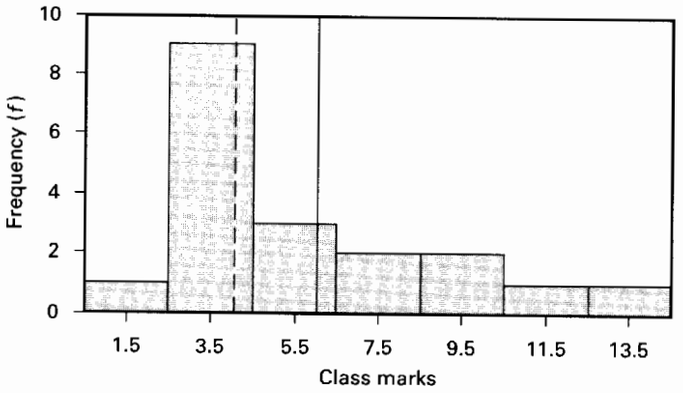

Figure 2.5 **Histogram of staff participation displaying the mean, median and mode values**

by the direction in which the tail of the distribution falls, that is the direction of the most extreme scores.

If a distribution is skewed, its mean will be changed in the direction of these extreme scores. The mean then may not provide us with the most appropriate representation of a typical score. In fact, it can be seen that this is the case for the staff data. Those one or two staff members who are rather more keen on aerobic exercise classes than the others are distorting the mean. However, the median is unaffected by these extreme scores. Therefore, with a skewed distribution, the median will provide a more appropriate indication of the typical score.

The direction of skewness is indicated by the location of the mean relative to the median. As the mean of the staff data is to the right of the median the distribution is skewed to the right or positively skewed. The magnitude of skewness will reflect the difference between the mean and median values.

Boxplots

A useful way of graphically representing the symmetry of data is the boxplot. This type of graph displays the median value by a horizontal bar surrounded by 50% of the scores shown within a box. This 50% of scores falls between the 25th and 75th percentile marks. The 25th percentile is at the bottom of the box and the 75th percentile is at the top. The whiskers extending from both ends of the box show the highest and lowest values that are not outliers. Outliers are scores in the distribution that are more than 1.5 box-lengths from the 25th or 75th percentile, and they are displayed by a circle; those that are more than 4 box-lengths away are shown by an asterisk. Look at the boxplots of the staff and student aerobic participation rates displayed in Figure 2.6.

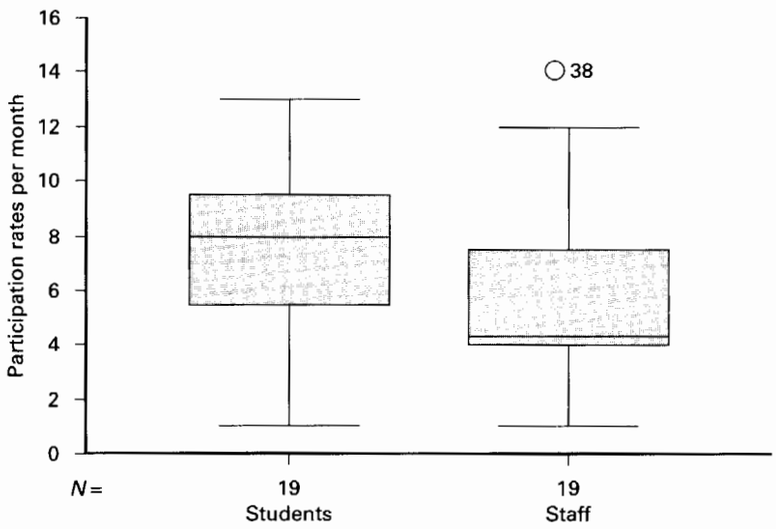

Figure 2.6 **Boxplot of staff and student aerobic class participation rates**

The boxplot for the students looks moderately symmetrical as the box is almost in the middle of the whiskers and the median is only slightly above the middle of the box. This suggests that this data is very slightly negatively skewed and therefore the mean may be employed as the appropriate measure of central tendency.

For the staff data the picture is quite different. The upper whisker is longer than the lower one and has one outlier. The number against this outlier is the case number. This case number identifies which participant had this extreme score. It can also be seen that the median score is at the bottom of the box. This suggests that the data is heavily positively skewed and hence the median may be employed as the appropriate measure of central tendency. Therefore the conclusion is that, on average, students attend 7.47 aerobic classes per month while staff attend only 4 classes per month. Later on you will be instructed in how to generate boxplots using SPSS.

The choice of which measure of central tendency to employ is not restricted to whether the data is skewed or not. In addition, the type of measurement that produced the data needs to be taken into account. Whenever anything is measured, numbers or values are assigned to observations. It is important to remember that how values are assigned determines the arithmetic operations that are permitted on the data. In other words, the type of data that is collected determines the statistical tests that can be used. Thus, different types of measurement result in the production of different types of data.

Levels of Data

Data may be measured in a variety of forms, but there are three specific forms of measurement to which we wish to draw your attention at this point.

First, data may be measured on a **nominal scale** whereby the units used bear no meaningful relationship to one another. For example, the researcher assigned the number 1 to someone who attended aerobic classes, 2 to someone who played field sports, 3 to someone who engaged in track events and so on. Here we can see that the researcher is only nominally (i.e. in name only) assigning these units 1, 2, 3, 4, etc., to participation in these activities and therefore they have no meaningful mathematical relationship to one another. As a consequence of this, it is meaningless to engage in any mathematical operation. For example, we cannot add the units together. Two people participating in aerobic classes who were given the number 1 do not equal one person who participated in field sports and had the number 2 assigned. All we can do is simply determine the frequency with which people participate in the various leisure activities. If the aim is to identify the typical leisure activity then the mode would be the most appropriate measure of central tendency. Because information collected in this way does not fall on any continuous scale it may be less confusing if the term nominal data rather than nominal scale is adopted.

A second form of data is measured on what is termed an **ordinal scale**. For example, children may be asked to rank a list of 10 sports from 1 to 10 in order of preference from most to least enjoyable. In this case the method of ordinal scaling provides no grounds for assuming that the difference in preference between sports listed 2 and 3 is of the same magnitude as that between sports listed 7 and 8. Furthermore, if a leisure researcher had students rank fitness clubs in terms of the facilities each provided for its members, it could not be assumed that the difference in facilities between the first and second ranked clubs was the same as that between the second and third ranked clubs. All that is known is the order in which the clubs were ranked for their facilities, not the magnitude of the distance between them. If a typical score of how these clubs' are ranked is required, then the median (the middle observation) should be employed.

The third distinct form of measurement is that of the **interval scale**. This is like an ordinal scale but with the additional feature that all of the intervals are equal. For example, the data the researcher gathered on participation in aerobic classes might be considered to be measured on an interval scale. The difference between attending three and four classes is the same as the difference between attending 10 and 11 classes, namely one aerobics class.

To obtain a typical measure with interval data the mean is normally employed. However, as described previously, this will only provide a typical score if the distribution is not heavily skewed, that is the distribution is symmetrical. If the data is heavily skewed then the median will render a more typical score. It was for this reason that the median was chosen as the appropriate measure of central tendency for the staff data even though it was measured on an interval scale.

Measures of Dispersion

By itself, a measure of central tendency does not provide an adequate shorthand summary of the data. Whilst it is very useful to know the typical score, it is also important to know how the rest of the data clusters around this mid-point. This requires a measure of the spread of the data. Measures of dispersion provide information on how the data is distributed around the typical score.

For example, if you are told that the mean exam score for a health and fitness class is 55, it tells you the average score that students achieved on the exam, but it does not tell you how the rest of the scores varied around that point. If there was no variability in the exam scores, everyone would have received a grade of 55. If there was a great deal of variability in the scores some students would have performed very poorly on the exam and some would have performed very well. In contrast, if the variability in exam scores had been low then most students would have performed close to the mean score of 55. One simple indication of the variability in exam scores might be provided by the range. Earlier in this chapter it was stated that the range equals the difference between the highest score and the lowest score. It was also noted that sometimes the value 1 is added to this equation to reflect the fact that the range is inclusive of the two extreme scores:

$$\text{Range} = (\text{highest} - \text{lowest}) + 1$$

For example, the range of scores in an exam where the lowest mark is 25 and the highest mark is 75, is 51, not 50. If the smallest value is subtracted from the largest, this would measure the difference between these values, but exclude the values themselves. Another way of expressing this is to recall that each of the scores in the distribution occupies a class interval equal to one unit. Therefore, the mid-points between class intervals extend 0.5 below the lowest score, and 0.5 above the highest score. So 24.5 should be subtracted from 75.5 to produce a range of 51.

Variance

The goal of a measure of dispersion is to produce some metric of how the scores are dispersed around the mean. One obvious solution would seem to be to subtract the mean from each score, and sum the differences to determine the total dispersion:

$$\sum(x - \bar{x})$$

Unfortunately, if the distribution is symmetrical, there will be just as many scores greater than the mean as there are scores smaller than the mean. Therefore, when all the difference scores (i.e. the deviations from the mean) are added together, the answer will always sum to zero. To overcome this problem the differences could be multiplied by themselves; that is, calculate the square of each of the deviation scores. This would have the effect of eliminating all the negative numbers, as a minus

multiplied by a minus equals a plus. This sum of the squared deviation from the mean is called the sum of squares:

$$\text{Sum of squares} = \sum(x - \bar{x})^2$$

This process would produce a measure that would increase as the range of the data increased. Unfortunately, the sum of squares would also increase as the number of the observations increased, that is as n increases. To control for the size of n the sum of the squares of the differences can be divided by n. This will calculate what is known as the **variance**, and the formula for this is

$$\text{Variance} = \frac{\text{sum of squares}}{n} = \frac{\sum(x - \bar{x})^2}{n}$$

This formula, translated into words, reads:

1. The mean is subtracted from each score.
2. The results (of step 1) are squared.
3. The squared deviations are summed.
4. The sum total of the squared deviations is divided by the number of observations (n).

In effect, when the variance is calculated, the calculation produces the average of the squared deviations from the mean, that is the mean of the squared deviations from the mean. To illustrate this let's work out the variance for the data concerning student participation in aerobic classes.

Example 2.1 Calculation of the variance of the data set presented in Table 2.2. (The mean of this data set was calculated subsequently to be 7.47.)

x	$x - \bar{x}$	$(x - \bar{x})^2$
1	−6.47	41.86
3	−4.47	19.98
4	−3.47	12.04
5	−2.47	6.10
5	−2.47	6.10
6	−1.47	2.16
6	−1.47	2.16
7	−0.47	0.22
7	−0.47	0.22
8	0.53	0.28
8	0.53	0.28
8	0.53	0.28
9	1.53	2.34
9	1.53	2.34
10	2.53	6.40
10	2.53	6.40
11	3.53	12.46
12	4.53	20.52
13	5.53	30.58
$\Sigma = 142$	0.0	172.72

$$\text{Variance} = \frac{\text{sum of squares}}{n} = \frac{\sum(x - \bar{x})^2}{n}$$

$$= \frac{172.72}{19}$$

$$\cong 9$$

Standard Deviation

In calculating the variance the differences between each observation and the mean were squared. Therefore, the variance reflects the dispersion of scores measured in units that are the square of the original units. For example, if the original units were the number of aerobic classes attended per month then the variance would be these units squared. To get back to our original units, the simplest solution is to take the square root of the variance. This will produce what is known as the **standard deviation**. This is a measure of the average variability in the distribution of scores:

$$\text{Standard deviation } (s) = \sqrt{\text{variance}}$$

If the variance of a data set was equal to 9 then

$$\text{Standard deviation } (s) = \sqrt{9} = 3$$

Standard deviation is a very important concept in statistics but this importance will only become apparent when we consider the normal distribution in the next chapter.

Computational Formulae for Calculating Standard Deviations

The computational formula used above for the variance can be used in calculating the standard deviation. All that needs to be done is to calculate the square root of the variance:

$$s = \sqrt{\frac{\text{SS}}{n}} = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

Usually this statistic is calculated for a sample of scores rather than for the whole population, so the value produced is only an estimate of the standard deviation in the population. Because it is only an estimate an error could have been made, and the statistic may not reflect the true variability in the whole population. In order to overcome this problem and offer a more conservative estimate of variability, $n-1$ is used in the denominator instead of n, thereby increasing the size of the standard deviation slightly.

The formula now reads

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

When there are whole numbers in the calculation, this formula is easy to use. However, when the mean involves decimal places, the calculation can become cumbersome. In this case an easier formula is available:

$$s = \sqrt{\frac{\sum x^2}{(n-1)} - \bar{x}^2}$$

A Simple Example

In case the previous discussion on variance seems a little obscure to you and the formulae a little abstract let's consider some simple data sets. In arrays of data that are based on small samples and whole numbers it is easy to visualise what the average variability in the data set might be. Consider the following example where there are three sets of data labelled A, B and C.

A	B	C
3	2	1
3	3	3
3	4	5

For each data set, the mean score is 3. As all the values in data set A are the same there is no variability at all in this data set. There is clearly greater variability in data set B, where the scores range from 2 to 4, and even greater variability in data set C where the scores range from 1 to 5. In data set B, the two scores that deviate from the mean appear to vary by an average of one unit. For data set C, however, the two scores that deviate from the mean appear to vary by an average of two units.

If we were to use the formula below to calculate the standard deviations for data sets B and C we would confirm our speculation:

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Consider data set B in Table 2.7.

Table 2.7 **Data set B**

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	−1	1
3	0	0
4	1	1
		$\Sigma = 2$

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{2}{2}} = \sqrt{1}$$

$$s = \sqrt{1} = 1$$

Now consider data set C in Table 2.8.

Table 2.8 **Data set C**

x	$x-\bar{x}$	$(x-\bar{x})^2$
1	−2	4
3	0	0
5	2	4
		$\Sigma = 8$

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{8}{2}} = \sqrt{4}$$

$$s = \sqrt{4} = 2$$

These calculations confirm the previous estimates of the average deviation based on a simple inspection of each of the two data sets, for they show that the average deviation for set B is 1 and for set C is 2. However, they also confirm the legitimacy of the formulae as they produce the same values that our quick visual inspection of the two data sets suggested.

Summary

Descriptive statistics aim to provide a shorthand description of large amounts of data by making summary statements about the sample as a whole, rather than making reference to each participant's scores. These statistics include the sample size (n), the typical score (mean, median and mode) and how the scores are dispersed around the typical score (standard deviation).

Which measure of central tendency will provide a typical score depends upon the type of scale on which the data was measured upon (interval, ordinal or nominal) and whether the distribution is symmetrical or skewed. If the data is symmetrical all three measures will have the same value but if the data is skewed the mean will be distorted in the direction of the extreme scores. Hence, the direction of skewness is indicated by the location of the mean relative to the median.

The distribution of the data can be graphically represented in the form of a histogram or a boxplot, or statistically represented through the concepts of variance and standard deviation. The variance is the average of the squared deviations from the mean whilst the standard deviaition is the square root of the variance.

EXERCISE 2.1 DESCRIPTIVE STATISTICS

A student researcher was interested in the importance of sound effects on computer games. He first of all noted the frequency with which a particular computer game was played. Then, using two identical computer games, he arranged for the sound effects on one computer game to be almost inaudible while the other was set at a high level. Observers then noted how often individuals would play the two computer games between midday and one o'clock on two consecutive days. The results of his observations were as follows:

Data list showing the number of times different individuals played the computer game with the high sound volume:

8, 7, 2, 3, 5, 3, 4, 6, 4, 4, 5, 3, 6, 7, 5, 6, 4, 5, 6, 2, 7, 5, 8, 1, 9

Data list showing the number of times different individuals played the computer game with the low sound volume:

1, 7, 2, 5, 3, 6, 4, 5, 3, 1, 4, 6, 5, 7

For each of the two groups of data:

1. Construct a frequency distribution.
2. Draw a histogram of the frequency distribution.
3. Compute:
   (a) the mode
   (b) the median
   (c) the mean
   (d) the variance
   (e) the standard deviation.

If you are not sure how to work out the variances then look at the example on page 22.

From your examination of both sets of descriptive statistics does it appear as if there is any difference in the frequency with which the computer game was played when the sound was high and when it was low?

EXERCISE 2.2 IDENTIFYING TYPES OF DATA USED IN RESEARCH

Below are listed a number of scenarios in which you are asked to identify the type of data collected *and* explain the reason for your choice. In carrying out this task you might find it helpful to consider how the researcher records the data and how it could be represented graphically.

1. A woman with a clipboard accosts you outside a shopping centre, and asks you which one of five laundry detergents you use most often.

   Data = Laundry detergents: .....

2. An exercise physiologist is looking at the effects of exercise on the core temperature of the body, and uses a rectal thermometer to measure body temperature.

   Data = Temperature: .....

3. A sports sociologist investigating the effects of socio-economic status (SES) on children's participation in leisure activities measures SES by recording the father's occupation.

Data = SES: .....

4. A researcher interested in comparing the attacking capabilities of each football team in the Premiership assigns the value 1 to the team which has scored the most goals, 2 to the next top scorers, 3 to the third top scorers and so on.

   Data = Attacking capability: .....

5. A researcher measuring how successful marathon runners felt after completing the London Marathon asked participants to rate themselves on a 1 to 10 scale, with 1 being unsuccessful, and 10 being successful.

   Data = Perceived success: .....

6. A physiotherapist, trying to investigate how different sports influence the type and severity of injuries, collects data from athletes involved in several types of sport, and determines the severity of injury by how many days of practice have been missed.

   Data = Type of sport: .....

   Data = Severity of injury: .....

7. During a drunken conversation in a local pub, a friend of yours asks you to list the 10 best athletes in the world in order of their ability.

   Data = Athletes: .....