16

Central tendency and dispersion

Statistical methods assist significantly in the analysis of data. For many researchers, analysis of data means statistical analysis and that any analysis cannot be successful without statistical operations. This point is contested by qualitative researchers; however, as we have already seen, even qualitative research lends itself to statistical analysis.

In this chapter examples of some elementary techniques will be introduced. These techniques relate to three groups of measures, namely to relational measures, measures of central tendency and measures of dispersion. Relational measures show how to relate parts of data to each other or to the whole; measures of central tendency show what the main trend is in the data; and measures of dispersion describe the spread of data. We begin with the relational measures.

1 Relational measures

Relational measures relate parts of a group of scores to each other or to the whole, for instance the relationship of males to females or the relationship of males to the whole group or to 100.

The measures we shall consider here are rate, ratio and percentage.

Rate

This is a measure used to compare values that are not a part of the same variable. It can measure the frequency with which a value occurs compared to the possible frequency with which this value could occur (Reid, 1987: 54). Rate is then expressed in the following relationship:

$$\text{Rate} = \frac{\text{number of actual occurrences}}{\text{number of possible occurrences}}$$

For instance, the graduation rate at Monash University is the number of graduations at this university divided by the number of graduations at all

universities in Victoria. Likewise, if the number of people killed in car accidents in NSW is 1200 and the number of deaths in that state for the same period of time is 4800, the rate of deaths caused by car accidents is 0.25. Rates are useful when comparing variables in different populations over time.

Ratio

'Ratio' describes the relationship of parts of a group to each other, and is computed by the following formula:

$$\text{Ratio} = \frac{\text{number of members of group A}}{\text{number of members of group B}}$$

If for instance there are 37 female students and 26 male students at the introductory sociology lecture, the ratio of male students (group A) to female students (group B) will be:

$$\text{Ratio} = \frac{26}{27} = 0.7$$

This suggests that the ratio of males to females is 7:10 (7 to 10); multiplied by 100 this figure becomes 70, which means that there are 70 male students for every 100 female students. The ratio of female students to male students will similarly be:

$$\text{Ratio} = \frac{37}{26} \times 100 = 1.42 \times 100 = 142$$

This means that the ratio of females to males is 1.4:1, or 1.42 females for each male, or that there are 142 female students for every 100 male students.

Percentage

While ratio relates two subgroups to each other, percentage compares a subgroup (n) to the total group (N). This is computed using the following formula:

$$\text{Percentage of } n = \frac{n}{N} \times 100$$

Using the above example where the number of male students was 26, the number of female students was 37 and the total number of students 63, the percentage (%) of the male students is:

$$\text{Percentage of male students} = \frac{26}{63} \times 100 = 41.27\%$$

The percentage of female students is:

$$\text{Percentage of female students} = \frac{37}{63} \times 100 = 58.73\%$$

A way of checking the accuracy of the results is to add up the percentages of the subgroups. If they add up to (about) 100 the computations are correct.

2 Measures of central tendency

Measures of central tendency are very popular and are the most commonly used statistical measures, not only by social scientists but also by people in everyday life. These measures represent the average or typical value in a distribution; in this sense, they summarise the entire distribution, by providing information about the main trend of the units of the population in question.

There are many ways of determining central tendency, and there are also many relevant techniques that help to compute these measures. In this chapter we shall introduce the most common measures: the mean, the mode and the median. We begin with the mean.

a The mean

The mean is by far the most important measure of central tendency, and also the most popular one among social scientists. It describes the central trend of the results or the average of all observations. Such measures are very common in everyday life. Everyone uses means when saying, for instance, that the average Australian believes in justice, that the average woman rejects patriarchy, and that the average wage of the Australians born overseas is lower than the average wage of those born in Australia.

The computation of the mean varies according to the nature of the data. The mean of listed data (those corresponding to individual scores, for example those presented in Example A) is computed in a manner that is different from that of grouped data (those set in frequencies, for example those shown in Example B).

Listed data

The computation of the mean in *listed data* is quite simple; it can be accomplished by adding up all scores and dividing the sum so obtained by the number of scores. For instance, if we were to compute the mean of the number of books read by five students during the holidays, and the number of books for each student was 2, 2, 3, 4 and 4, the mean would have been 2 + 2 + 3 + 4 + 4 = 15, divided by the number of scores, 5, which equals 3. Consequently, the mean is 3. This means that the average number of books read by these five students is 3. We use the symbol $\bar{x}$ (read: ex bar) to indicate the mean.

The mean for ungrouped data is computed by means of the following formula:

$$\bar{x} = \frac{\Sigma x}{n} \qquad \textbf{(16.1)}$$

This formula displays exactly what we stated above, namely that the mean is the sum of the values of observations (Σx), divided by the number of observations (n). Σ (sigma) is a Greek letter that means 'the sum of'. Let us discuss an example.

Example A: In a sociology test the following scores have been recorded:

27, 17, 47, 52, 42, 37, 32, 27, 22, 17, 27, 37, 17, 32, 37, 27, 17,

22, 42, 57, 47, 27, 37, 32, 27, 22, 17, 12, 32, 27, 22, 17, 12, 27

Since the data are ungrouped, we add up all the scores (Σx) and divide by the number of observations (n). Since $\Sigma x = 993$ and $n = 34$:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{993}{34} = 29.2$$

This shows that the average score in the sociology test was 29.2.

Grouped data I (frequencies)

Quite often the data are available in a grouped form, where observations or scores appear in frequencies. In such cases the computation of the mean proceeds in a different manner. To demonstrate this let us consider the distribution given in Example B, which is a grouped presentation of the scores introduced in Example A.

Example B: Frequency of scores of sociology students

Score x	*Frequency* f	*Product of x,f* fx
12	2	24
17	6	102
22	4	88
27	8	216
32	4	128
37	4	148
42	2	84
47	2	94
52	1	52
57	1	57
	$n = 34$	$\Sigma(fx) = 993$

In order to compute the mean in this distribution, a slightly different formula is employed. This is as follows:

$$\bar{x} = \frac{\Sigma fx}{n} \qquad \textbf{(16.2)}$$

According to Formula (16.2), in order to compute the mean we need n as well as the product of x and f (fx), which is required in order to calculate *the sum of products*. For this reason a new column is required in the frequency table describing the factor. The value of n will be computed from the second column, namely the f column. The frequency table (shown in Example B) is complete, and provides the information needed to compute the mean.

Employing Formula (16.2), we find:

$$\bar{x} = \frac{\Sigma fx}{n} = \frac{993}{34} = 29.2$$

Grouped data II (class intervals)

It is quite common for distributions to include class intervals instead of single numbers, as shown in Example C. In order for the mean to be computed the same procedure as above is employed. However, this might cause some confusion to the beginner since there are two values of x to consider, the lower and the upper limits. In such cases, in order to compute the mean we take the midpoint instead. This way the intervals are transformed into single numbers. When this is accomplished, the computation follows the procedure employed for grouped data, as shown in Example C.

Example C: Age groups of unmarried mothers

Age group x	*Midpoint* x	*Frequency* f	*Product* xf
16–18	17	5	85
19–21	20	9	180
22–24	23	16	368
25–27	26	11	286
28–30	29	14	406
		$n = 55$	$\Sigma(xf) = 1325$

The mean age of the unmarried mothers of this particular group is:

$$\bar{x} = \frac{\Sigma fx}{n} = \frac{1325}{55} = 24.09$$

b The mode

The mode is the category with the largest number of observations. If, for instance, in a sociology test 6 students received an A, 9 a B, 16 a C, 7 a P and 4 an F, C is the mode because it is the most frequent category. As shown here, the mode is not computed mathematically but is identified logically on the basis of its relationship with other values. It is a measure you can *see*, rather than one you need to calculate.

While there can only be one mean in a distribution, with regard to the mode the situation is different. Distributions can have one mode (unimodal distributions), two modes (bimodal distributions), more than two modes (multimodal distributions) or even no mode at all (non-modal distributions), when, for instance, all observations in the distribution are the same.

The mode is a useful measure but it is not used very widely in the social sciences, except for nominal data and when a quick description of the trend in the data is needed.

c The median

The median is the point on a distribution that divides the observations (not their values) into two equal parts, so that half of the observations are above and half below this point. For example, in the distribution 36, 33, 30, 28, 26, 23, 18, 12, 11, 8, 4, showing the reading hours per week reported by the 10 students of a history class, the median is 23 because it divides the distribution into two parts with equal numbers of scores; there are five scores each side of 23. Similarly, in the distribution 150, 125, 110, 75, 68, 40, 23, 20, 18, 15, 12, the median is 40. Note that when computing the median, it is the number of observations rather than the actual values of the observations that counts.

The way of computing the median depends on the nature of the distribution. In listed distributions, such as those presented above, the computation is very simple. It only involves rank ordering of the scores and identifying the score that divides the distribution into halves. This process becomes more involved when there is an even number of scores and the median falls between two scores, when there is a large number of scores and frequencies are required, or when there are tied scores.

When there is an even number of scores, the median is the mean of the two adjacent middle scores. For example, in the distribution 18, 16, 15, 13, 11, 9, 7, 5, the point that divides it into two equal parts lies between 13 and 11. Consequently, the median is the mean of these two scores, namely 12.

In the following we shall explain how the median is computed in ungrouped and grouped distributions.

Grouped data (frequencies)

Let us first look at a distribution without intervals. Consider the following example: in a small study of 102 same-sex couples in Sydney's Western Suburbs it was found that the number of children per unit was as shown in Example D.

Example D: Number of children per unit

Number of children x	*Frequency* f
1	12
2	46
3	23
4	11
5	8
6	2

What is the median age of the children? To compute the median we follow the steps shown below:

1 The observations are set in a rank ordered form. This is already done in Example D.
2 The *midpoint observation* is defined. This is half of the sum of the frequencies (the sum of 12, 46, 23, 11, 8 and 2, which is 102, divided by 2). The midpoint observation is then half of 102, which is 51.
3 The frequency at which the midpoint observation occurs is located by adding up the frequencies from one end of the distribution. The frequency that contains the midpoint observation, that is, the frequency that is added last and gives a sum that exceeds the midpoint observation, in Example D is 46 (12 + 46 = 58; and 2 + 8 + 11 + 23 + 46 = 88) because 51 is reached only after 46 is added to 23 (if one counts from below) or to 12 (if one adds from above).
4 The category to which the observation containing the midpoint observation corresponds is located. In our example this category is 2 (i.e. 2 children).
5 Consequently the median number of children is 2.

It is quite possible that, when counting from above and from below, two different midpoints are identified and for this reason two medians are computed. In such cases the average score of the two medians is the true median.

Grouped data (with intervals)

Let us now consider an example with intervals. The obvious problem with computing the median in such a case is that after the midpoint observation is identified, it will point to a category that includes several figures. If, for instance, the category that corresponds to the observation containing the midpoint is 25–34, a specific median cannot be determined because it points to all numbers between 25 and 34. In order to compute the median in such cases, the following formula is employed:

$$\text{Median} = l + \frac{\left(\frac{N}{2}\right) - cn}{n} \times w \qquad \textbf{(16.3)}$$

where the letter l stands for the lower limit of the category that contains the midpoint observation (if the interval is 20–25, l is 20); N represents the number of observations included in the study, which is the sum of the frequencies; and cn stands for the cumulative number of observations of the category that immediately precedes the category containing the midpoint (this means that first the cumulative frequency is computed, the category containing the midpoint defined, and then the category above it is taken); n stands for the number of observations contained in the midpoint category (here we refer to the frequency column); and w stands for the width of the category.

This formula offers one way of computing the median; there is another way of computation which is, however, beyond the limits of this book.

Let us now compute the median in a distribution with intervals, by using the example of the age of single fathers (see Example E).

Example E: Age groups of single fathers

Age group	*Frequency f*	*Cumulative frequency cn*
16–18	5	5
19–21	9	14
22–24	16	30
25–27	11	41
28–30	14	55

To compute the median in this distribution we operate as follows:

1 The cumulative frequency (*cn*) is calculated by adding each frequency to frequencies above it; for example, the first frequency is 5, the second frequency is 14 (i.e. 5 + 9), the third frequency is 30 (i.e. 16 + 9 + 5), the fourth frequency is 41 (11 + 16 + 9 + 5), and the last frequency is 55 (i.e. 14 + 11 + 16 + 9 + 5).
2 The midpoint observation and the category that contains it are computed. In Example E it is 55 divided by 2, which is 27.5. The midpoint is in 30 and the corresponding interval is 22–24 (in exact terms it is 21.5–24.5).
3 Formula (16.3) is used. Following the definition of the symbols discussed above, $l = 22$, $N = 55$, $cn = 14$, $n = 16$ and $w = 2$ (if w represents the *exact interval* it will be computed by subtracting the lower limit, namely 21.5, from the upper limit, 24.5, and will be 3).

Employing Formula (16.3):

$$\text{Median} = l + \frac{\left(\frac{N}{2}\right) - cn}{n} \times w = 22 + \frac{\left(\frac{55}{2}\right) - 14}{16} \times 2$$

$$= 22 + \frac{27}{16} = 22 + 1.68 = 23.68$$

d Mean, mode and median

Each of the measures discussed above, mean, mode and median, provides specific information about the trend demonstrated in the data and is used when this specific information is required and when conditions allow it. Nevertheless, which measure should be chosen in each case? In order to answer this question we should carefully study the distribution and examine two major factors, namely the type of measurement and the shape of the distribution. Table 16.1 summarises the suitability of these measures for the various levels of measurement.

Table 16.1 Levels of measurement and central tendency

Level of measurement	*Measure*
Nominal	Mode
Ordinal	Mode, median
Interval	Mode, median, mean
Ratio	Mode, median, mean

Overall, there is agreement among researchers (e.g. Foddy, 1988; Sofos, 1990) that the following points may serve as a guide when deciding which measure is the most suitable:

- The mode is chosen if the variable is nominally scaled, although it can be used for all types of data.
- The mean is chosen if the variable is ordinal, interval or ratio.
- If the distribution shows a central tendency, the mean or median is a better measure; if there is no central tendency the mode is preferable.
- If the distribution is skewed, the median is a better measure. This is particularly so for distributions of interval data. When the skewness is extreme and if the distribution contains ordinal data the mode may be a better choice (Foddy, 1988: 74).
- If further measures are to be considered (e.g. standard deviation) the mean should be preferred.
- If a quick but rough measure is acceptable, the mode can be helpful.
- If information about the central trend is wanted, the mean is the best choice.
- If information is needed about the location of cases in the two halves of the distribution, the median is a better measure.

Despite the advantages and disadvantages of these measures, or perhaps because of them, social scientists seem primarily to employ the arithmetic mean as the measure of central tendency. The other methods are only marginally used. The mean has many mathematical properties that are very important, not only for providing a guide for central tendency but also for being necessary for computing other measures. It is also a stable measure, not being easily affected by shifts in a few data, and it is a clear and direct method obtained from raw data, irrespective of their order.

3 Measures of dispersion

a Introduction

The measures we presented above are employed to demonstrate central tendency, that is, the general trend that is evident in the findings of the study. These are useful measures, which help to define the direction demonstrated by the data, and present a summary impression of some major traits of the

population. Knowing, for instance, that the average IQ in a group of secondary school students is 120 gives us a very good indication of the level of intelligence of the members of that group and provides useful information to those teaching these students.

Measures of dispersion are equally useful and informative; they inform, however, of a different quality of the data, namely the degree to which the data are spread around the mean. Measures of dispersion show how close to or far away from the main stream of the data the observations are. If, for example, the average IQ is 120, how low is the lowest and how high is the highest score? How many low or high scores are in the distribution? And what is the average of the individual deviation of the scores from the mean? Such information is provided by means of the *measures of dispersion*. In this section we shall introduce the most common methods of computing such measures.

b Variance and standard deviation

Variance and *standard deviation* are the two most popular measures of dispersion in the social sciences. The standard deviation is the square root of the variance; and hence the variance equals the square of the standard deviation.

In simple terms, *the variance is the average of the distances of the individual scores from the mean.* The procedure for computing the variance is the same as that of the mean, namely adding the distances of the individual observations from the mean and dividing them by their number. Due to the fact that half of the scores lie above and the other half below the mean, half of the differences from the mean are positive and half are negative; the calculation of the mean of the differences from the mean is thus impossible, for the sum of the distances is inevitably zero. For this reason, the variance is calculated by using the squared deviations from the mean. Thus, *variance is the mean of the squared deviations of the observations from the mean.* Let us explain this in an example (see Example F). Assume that we are interested in the spending habits of a group of 10 primary school students. We asked them to state the amount of money they spend weekly at school; the results are presented in Example F.

Example F: Distribution of amount of money spent by primary school students at school

Students	*Amount ($)*
A	5
B	7
C	3
D	9
E	16
F	12
G	8
H	4
I	2
J	14

By adding up the observations and dividing by 10 (the number of students) we can calculate the average amount of money spent at school, which is $8. If we plot the observations on the coordinates, we obtain a figure which, as well as showing the place of the observations in relation to each other, also shows the distances of each score from the mean (see Figure 16.1).

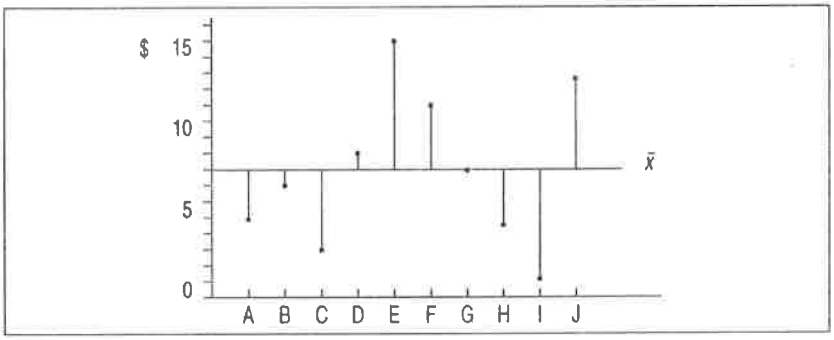

Figure 16.1 Amount of money spent at school

This figure shows that the distances of the scores from the mean vary from one case to another, some of them being below the mean and some above it. The variance demonstrates the average distance of these scores from the mean.

c Computation of variance and standard deviation

There are four basic formulae of variance employed by statisticians, two for ungrouped and two for grouped data.

Example G: Variance of the amounts of spending money from the mean

Students x	*Amount* $x-\bar{x}$	$(x-\bar{x})^2$	
A	5	–3	9
B	7	–1	1
C	3	–5	25
D	9	1	1
E	16	8	64
F	12	4	16
G	8	0	0
H	4	–4	16
I	2	–6	36
J	14	6	36
$\bar{x} = 8$		$\Sigma(x-\bar{x}) = 0$	$\Sigma(x-\bar{x})^2 = 204$

Listed data I

The first formula of variance is a translation of the definition of variance into mathematical symbols. This formula is:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{N - 1} \qquad \textbf{(16.4)}$$

This formula defines variance in mathematical terms and symbols. It tells us to subtract the mean from each score, square what we get, add up the results and then divide by the number of scores. In practice this can be a longwinded method of calculation. Formula (16.5) gives a more practical computational procedure, as we shall see soon.[1]

As you will recall, to calculate a mean we add up all the scores and divide the sum by the number of scores. Similarly, the variance is the sum of the squared deviations from the mean, divided by the number of scores. It is a measure of the average spread from the mean. To compute the variance we need x, $x - \bar{x}$ and $(x - \bar{x})^2$. Their value can be calculated as shown in Example G. Note that each column corresponds to each of the factors of the formula.

Returning to Formula (16.4) we find that:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{N - 1} = \frac{204}{9} = 22.67$$

Thus the variance is 22.6 ($s^2 = 22.6$). Given that, by definition, the standard deviation (s) is the square root of the variance (s^2), s is the square root of s^2 and therefore the square root of 22.6 (which is 4.76). Consequently, the standard deviation is 4.76 (s = 4.76).

The results suggest that the standard deviation of the amounts spent weekly by students at school is \$4.76; in simple terms this means that the average deviation from the mean amount of money (i.e. from \$8) spent by students is \$4.76.

Listed data II

This method is called the *raw-score method* and is widely used in the social sciences. The advantage of this method is that it requires neither computation of the mean nor calculations of the deviation from the mean and thus the computations are markedly easier than in the previous method.

The formula employed in the raw-score method looks more complicated than Formula (16.4); it is, however, much easier and less time consuming to compute. The formula is:

$$s^2 = \frac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{N}}{N - 1} \qquad \textbf{(16.5)}$$

[1]Social researchers distinguish between the standard deviation for a population (σ) and for a sample (s). In effect, the only difference between the computation employed in each case is that their formulae have different denominators: the denominator of the formula for s is $N - 1$, whereas the denominator for σ is N. In terms of value, the difference between the two formulae is significant for small samples only; the closer the sample is to 100, the smaller the difference and for samples larger than 100, there is no difference. Given that in quantitative studies (where it is more likely that statistical procedures are used) samples are usually large, in social research use of one or other formula makes no difference.

To compute s^2 it is necessary to calculate (1) the sum of the squared scores, Σx^2, and (2) the squared sum of the scores, $(\Sigma x)^2$. In other words it is necessary to square all the scores and add up all the products to get Σx^2, and to add up all the scores, thus obtaining Σx, and then square the sum to get $(\Sigma x)^2$. Let us compute the standard deviation in Example H.

Example H: Standard deviation of scores from the mean

x	x^2
5	25
7	49
3	9
9	81
16	256
12	144
8	64
4	16
2	4
14	196
$\Sigma x = 80$	$\Sigma x^2 = 844$

Using Formula (16.5) we find:

$$s^2 = \frac{844 - \frac{80^2}{10}}{10 - 1} = \frac{844 - 640}{9} = \frac{204}{9} = 22.67$$

$$s = \sqrt{22.67} = 4.76$$

Thus, the standard deviation score computed using the raw-score formula is the same as the score obtained using Formula (16.4).

Grouped data I

The formula employed when we deal with grouped data is similar to Formula (16.4) with the exception that the formula for the grouped data takes into account the frequency (f) of appearance of the individual scores. For this reason, the construction of the table and computation of the standard deviation are similar to those in Formula (16.4). The formula for grouped data is:

$$s^2 = \frac{\Sigma f(x - \bar{x})^2}{N - 1} \qquad \textbf{(16.6)}$$

According to this formula, in order to compute the standard deviation the following factors are required: x, $x - \bar{x}$, $(x - \bar{x})^2$, $f(x - \bar{x})^2$ and $\Sigma f(x - \bar{x})^2$. These factors indicate the type and number of columns we need to set up in the table to compute the standard deviation. Let us look at an example.

In a recent sociology test the scores of 30 students were distributed between 1 (high distinction) and 5 (failed) as shown in Example I.

Example I: Distribution of test scores

x	f	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
1	1	−2	4	4
2	5	−1	1	5
3	18	0	0	0
4	5	1	1	5
5	1	2	4	4
$\bar{x}=3$	$n=30$			$\Sigma f(x-\bar{x})^2=18$

Employing Formula (16.6) we find:

$$s^2=\frac{\Sigma f(x-\bar{x})^2}{N-1}=\frac{18}{29}=0.62; \quad \text{and } s=\sqrt{0.62}=0.79$$

The standard deviation of the test scores is 0.79. This means that the average deviation of the scores from the mean (3) is 0.79.

Grouped data II

This method has the same qualities as the raw-score method discussed above under 'Listed data II'. The standard deviation is computed without calculating the mean or the difference of the scores from the mean. The relevant formula is:

$$s^2=\frac{\Sigma fx^2-\dfrac{(\Sigma fx)^2}{N}}{N-1} \qquad \textbf{(16.7)}$$

As in the previous cases, in order to compute the standard deviation it is necessary to calculate x^2, fx^2 and fx. From these values Σfx^2 and $(\Sigma fx)^2$ will be computed. Let us look at the figures in Example I again, here shown in Example J.

Example J: Distribution of test scores

x	f	x^2	fx^2	fx
1	1	1	1	1
2	5	4	20	10
3	18	9	162	54
4	5	16	80	20
5	1	25	25	5
		$\Sigma fx^2=288$		$\Sigma fx=90$

Employing Formula (16.7) we find:

$$s^2 = \frac{288 - \frac{90^2}{30}}{30 - 1} = \frac{288 - 270}{29} = \frac{18}{29} = 0.62; \quad \text{and } s = \sqrt{0.62} = 0.79$$

Thus, the variance is 0.62, and the standard deviation is 0.79, which is the value obtained using Formula (16.6).

d The range

The range is another measure of variability. As the title indicates, this measure demonstrates the range that the distribution covers, from the lowest to the highest score. For this reason, its computation is quite simple; it describes the distance between the highest and the lowest score of a distribution, and is thus computed by subtracting the lowest score from the highest score.

Obviously, the range is quite different from the standard deviation. While the latter considers the spread of the data on the basis of their distance from the mean, the range does not refer to or depend on the value of the mean. Rather it refers to the continuum of the scores contained in the distribution, and shows how far apart its two extreme scores are. Let us study two examples.

Example K: The 10 students of class A in a Sydney primary school were found to spend the following weekly amounts of money on sweets (in dollars):

10, 9, 9, 8, 8, 8, 7, 7, 7, 7

In a grouped form the data show that one student spent \$10, two spent \$9, three \$8 and four \$7. The range here is 10 – 7 = 3 (\$3); this means that the students have similar spending patterns, since the difference between the person who spent the highest amount and the one who spent the lowest amount is just \$3.

Let us look at another example.

Example L: In another class of the same grade and school the 10 students interviewed each spent the following amounts of money on sweets (in dollars):

5, 7, 3, 9, 16, 12, 8, 4, 2, 14

Following the same procedure, we find that the range is 16 – 2 = 14, i.e. \$14. The range demonstrates here that there is a large difference in the spending patterns of the members of the second group of students. This group is rather diverse.

A study of the central tendency of both distributions shows that *on average* both groups spent the same amount of money; the mean in both groups is $8. The range, however, indicates that a conclusion stating that both groups are therefore similar in their spending habits is misleading. The range shows that these groups are diverse in their spending patterns and not uniform as the mean suggests. The information offered by the range is not as specific as that offered by standard deviation; it is, however, very useful indeed.

e Interquartile range

This measure is a version of the range except that it excludes from the computation the two ends of the distribution. More precisely, it leaves out the lower and upper quarter of the distribution. In this way extreme cases (outliers) which can skew the range value are excluded. To compute the interquartile range we proceed as follows:

1 Rank order the scores from the highest to the lowest.
2 Devide the distribution into four equal parts (first, second, third and fourth quarters or, better, *quartiles*).
3 Subtract the lowest score of the second quartile from the highest score of the third quartile. The difference is the *interquartile range.*

For instance, if the income of 20 students ranged from $20 to $250 per week and we wished to compute the interquartile range, we would rank the students according to their income from the highest to the lowest, divide them into four groups (quartiles) and subtract the income of the student at the bottom of quartile 2 (e.g. $31.00) from that of the student at the top of quartile 3 (e.g. $69.00). The difference between these two amounts ($38.00) is the interquartile range. This measure is more realistic than the range (in our case $230.00) because it excludes outliers, such as $250.00 in our example.

4 Computing $\bar{x}$, the median, the mode and s^2 using SPSS

Although the computation of the measures of central tendency and dispersion is easy, it also is tedious and time consuming. In any case it is not as easy as when computers are used. As a result there is no researcher who will compute these measures manually. The ease and speed of computation by computer and the high degree of accuracy have made manual computation obsolete. To work out these measures using SPSS you first enter the data in the computer; then while at the data editor window you proceed as follows:

1 Click on **Statistics > Summarise > Frequencies**
2 Transfer the variable for which you need information to the **Variable(s)** box
3 Click on **Statistics** (at the bottom of the window)
4 Activate the desired statistic (mean, median, mode, standard deviation, variance and range)
5 Click on **Continue**
6 Click on **OK**

The computer will display all activated measures on the same screen. For Examble D the computer output will be as follows:

CHILDREN

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	1.00	12	11.8	11.8	11.8
	2.00	46	45.1	45.1	56.9
	3.00	13	22.5	22.5	79.4
	4.00	11	10.8	10.8	90.2
	5.00	8	7.8	7.8	98.0
	6.00	2	2.0	2.0	100.0
	Total	102	100.0	100.0	

Mean	2.637	Median	2.000	Mode	2.000
Std dev	1.184	Variance	1.402	Range	5.000

Valid cases 102 Missing cases 0

The results are self-explanatory; the mode is 2 because it is the category with the largest number of observations; and the value of the median is the same as that computed above manually.

5 Comparing scores and standard deviations

Scores and standard deviations offer useful information if interpreted inside but not outside their distribution. For instance a test score of 80 is larger than a score of 75, but if the mean of the distribution of the former is 78 and the mean of the distribution of the latter is 55, a score of 75 may have a higher value than a score of 80. In a similar fashion, a standard deviation of 6 is larger than a standard deviation of 3, but this is not necessarily so if the former refers to a distribution of 10 respondents and the latter to one of 100.

To have a realistic value and to allow comparisons these measures need to be brought down to a common denominator or, better, to be standardised. This is done for scores by means of the *standard scores* and for standard deviations by means of the *coefficient of variation.*

a Standard scores (z-scores)

As stated above, although raw scores offer specific information about the position of a respondent on a scale, they are of little use if comparisons are to be made between different distributions. Standard scores offer a handy and effective method for expressing relationships and allowing comparisons between raw scores. Let us study this in an example (Example M).

Example M: Two male students have applied for a scholarship which is granted on the basis of academic standards and achievement in the end-of-the-year examinations. The first student studies psychology and his score is 65; the second student studies history and his score is 70. On face value, the scores indicate that the history student should receive the scholarship, for his score is higher than the score of the other student. The question here is: Does the score 70 indicate that the history student has a higher achievement than the psychology student?

The logic behind this question is whether scores taken from two different scales can be compared. If both scores had come from the same distribution it would have been very simple to evaluate the actual value of the scores and to decide about who deserves the scholarship. But this is not so in our example.

To allow a valid comparison between these two scores and to make the decision about the scholarship easier, the z-score is employed.

Standard scores (or z-scores) transform raw scores from different distributions into a common distribution, which has the same mean (a mean of zero) and the same standard deviation (a standard deviation of 1). Standard scores convert raw scores into standard units of standard deviation. A z of 2 means that the raw score is two standard deviations above the mean. A negative z-score indicates that the score is below the mean. A z-score of –1.5 means that the score in question is one and a half standard deviations below the mean.

Raw scores are converted into standard scores (z-scores) by means of the following formula:

$$z = \frac{x - \bar{x}}{s} \qquad \textbf{(16.8)}$$

The transformation is fairly simple. How raw scores are converted into standard scores is shown in Example N, which uses the scores for the two students in Example M.

Example N: Raw scores converted to standard scores.
Student 1: Assume that this student's raw score of 70 came from a group of students in which the mean was 71 and the standard deviation 6. The conversion into a standard score proceeds as follows:

$$z = \frac{x - \bar{x}}{s} = \frac{70 - 71}{6} = -\frac{1}{6} = -0.17$$

This shows that the score 70 is slightly below the mean; more accurately, it is 0.17 standard deviations below the mean. In simple terms this shows that the student in question performed slightly below average in the examination.

Student 2: Assume that the score 65 came from a distribution of scores in which the mean was 58 and the standard deviation 3. The conversion of this score into a standard score follows the same procedure:

$$z = \frac{x - \bar{x}}{s} = \frac{65 - 58}{3} = \frac{7}{3} = +2.3$$

This shows that the score 65 is an excellent score, being 2.3 deviations above the mean. In simple terms student 2's performance in the examination was far above the average.

The information provided by the two standard scores offers a valid basis for comparison, because it employs the same framework: both standard scores have a common mean and standard deviation and they are evaluated on the basis of the same standards. Therefore, their difference is more meaningful than the difference between the raw scores.

b Computing z-scores using SPSS

1 Enter raw scores in the computer
2 Choose **Statistics > Summarise > Descriptives**
3 Transfer the variable you wish to standardise to the **Variable(s)** box
4 Click on the square in front of **Save Standardised values as variables** box
5 Click on **Continue**
6 Click on **OK**

To convert raw scores to z-scores using SPSS you proceed as follows:

The computer will display in the output window information which tells you that a new variable has been added to your data, showing also the name under which this variable has been saved. If you switch over to the Newdata screen, you will see that z-scores have been added alongside the original raw scores. These scores can be statistically treated as any other set of data.

c The coefficient of variation

The coefficient of variation (also known as coefficient of relative variation) serves a purpose that is similar to that of *z*-scores: it allows researchers to compare standard deviations and decide whether one is larger than another. This is done by relating standard deviation to the mean and converting it to a percentage. Consequently, to compute the coefficient of variation *we divide the standard deviation by the mean and multiply the result by 100.* The formula for the coefficient of variation (CV) is as follows:

$$CV = \frac{s}{\bar{x}} \times 100 \qquad \textbf{(16.9)}$$

Let us look at an example.

Example O: A researcher investigated the amount of money spent on Fridays after work by members of a large union, and recorded their educational status. The data collected produced a standard deviation of $25.00 for spending and of 6 for education (in years of education). Obviously, income seems to have a larger standard deviation than education, but does this mean that the variability of the former is larger than the variability of the latter? The mean for spending is $50.00 and the mean number of years of education is 10 years.

To answer the research question we substitute the values in the formula and obtain the following figures:

$$CV_{spend} = \frac{25}{50} \times 100 = 50\% \quad \text{and} \quad CV_{educ} = \frac{6}{10} \times 100 = 60\%$$

Given that the coefficient of variation for education is larger than the coefficient for money spent on drinking, one can conclude that, despite the fact that the standard deviation for spending is larger than the standard deviation for education, variability in years of education is larger than in spending on drinking.

Key concepts

Ratio
Mean
Mode
Variance
z-score
Rate
Listed data
Median
Range
Percentage
Group data
Standard deviation
Standard scores