5

INTRODUCTION TO IBM SPSS STATISTICS AND THE DATA SET

I mentioned in Chapter 1 that, nowadays, rather than having to calculate the mathematical equations for our data analysis ourselves, we usually employ software packages to do this. There are a variety of software packages that do quantitative data analysis.

R, SAS and Stata are all general purpose statistical analysis software packages. Excel and other spreadsheet software packages allow quantitative data analysis (Excel has an 'analysis tool-pak' add-on module), although these are more limited in scope and often less user-friendly than the specialised packages. In this book, we are going to use IBM SPSS Statistics (SPSS). This is because SPSS is probably the most common statistical data analysis software package used in educational research and is available at most institutions of higher education. It is also quite user-friendly and does everything we need it to do. This does not mean that it is necessarily 'better' than any of the other packages. Other packages may be better in some areas, but SPSS is one of the most versatile and commonly used statistical data analysis software packages.

5.1 Introduction to SPSS

Let's look at what SPSS is like. When we open it, we get the screen shown in Figure 5.1.

We can see that the screen is dominated by a grid. This is where (once we have opened a file) we will find the variables and the units. Our units are the rows, numbered from 1 to however many people we have in our study. The variables are the columns. The names appear at the top of the grid (where it now says 'Name').

Opening a data file is pretty similar to doing so in a program such as Word or Excel. We simply need to go into the open folder icon and select our file, called 'Quants file' (remember, you can download that file from the website as instructed in the Preface to this book).

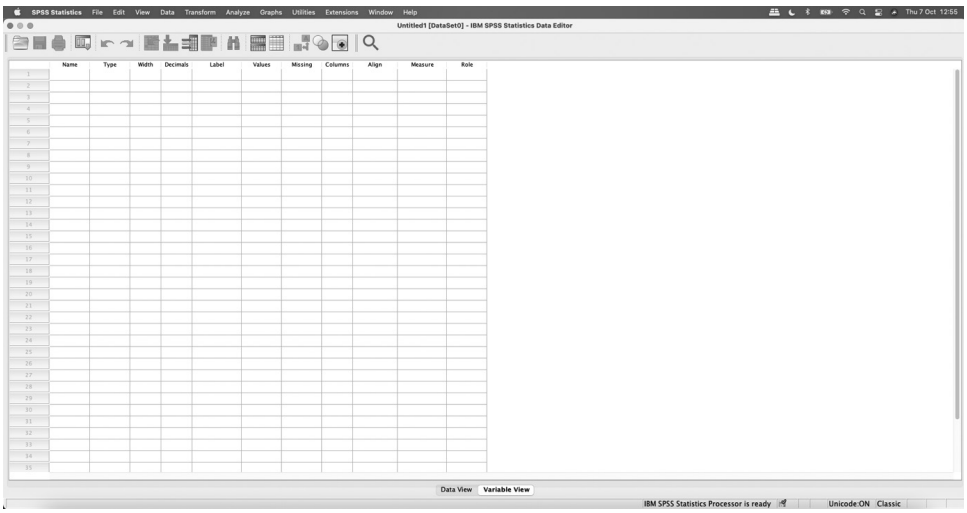


Figure 5.1 SPSS opening screen

Now we can see that the values for all variables for every unit have appeared in the grid (Figure 5.2). The value of the variable 'age' for respondent (unit) 1 is 143, for example (age is calculated in months here).

While we can see the names of all the variables along the top of the grid, these are not necessarily very clear. What does 'attsc1' mean? Well, the explanation can be easily found. To see the explanation, we need to go into a different screen, however.

How do we do that? Well, at the bottom of the screen you will see two tabs, one called 'Data View' and the other 'Variable View'. What we need to do is to go into 'Variable View', by clicking on that tab.

Once we have done this, we can see a new screen (Figure 5.3), which lists, in the first column, the names of all the variables in our file. The next columns list other variable characteristics, such as type (numeric [numbers], string [letters] and so on), width of the variable and number of decimal points. The next column gives us the labels. This is where we can find out what the variable actually means. So, for example, our variable 'attsc1' is the item 'school is boring' in our questionnaire. The next column gives us the values that variable can take on. For attsc 1, a value of 1 corresponds to 'agree strongly', 2 to 'agree', 3 to 'disagree' and 4 to 'disagree strongly'. This is an example of how we can convert answers to numbers, something we will always have to do if we want to analyse data quantitatively.

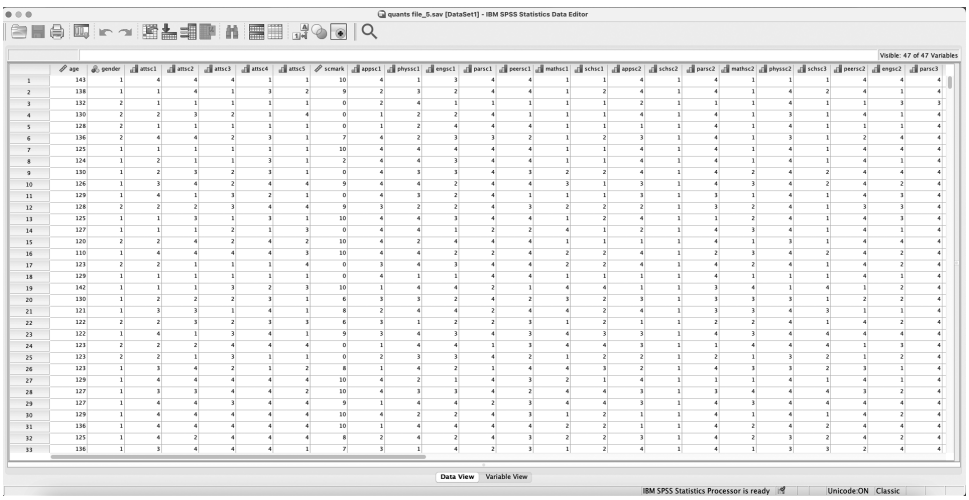
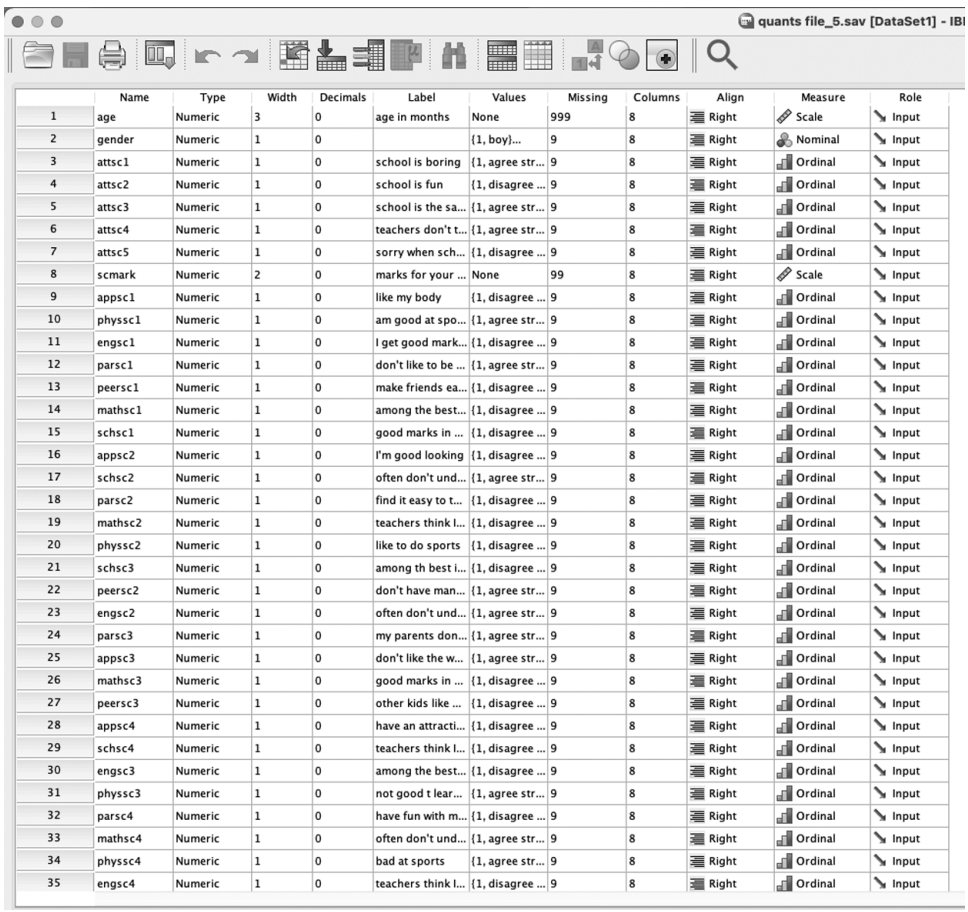


Figure 5.2 The 'data view' screen with our file opened

The next column gives us our missing value codes. Missing values commonly occur in quantitative research. This can happen because respondents don't fill in a particular question or because they fill it in wrongly. It is good practice to give a code to those missing values. That way, SPSS can recognise that they are missing and exclude them from analyses. Conventionally, values of 9, 99, 999 and so on are used for missing values. Obviously, our missing value code has to be one that is not a code for that variable (so, if we had a scale from 1 to 10, our missing value code would have to be 99 and not 9). The next two columns (number of columns and alignment) are purely concerned with layout. The penultimate column, labelled 'Measure', gives us the level of measurement for each variable (nominal, ordinal or 'scale' – what SPSS calls continuous variables). We will discuss levels of measurement in the next chapter. One thing to watch out for with this is that SPSS will assign a level of measurement to each variable based on its best guess of what type of variable it is. These are often wrong, so it is a good idea to check these and change them where necessary. The final column (input) is simply a description of how the numbers were entered.

Changing variables or their characteristics or adding a new variable is easy in SPSS. If you want to add a new variable, you can simply type the name in the first column of the bottom row. A number of default values (e.g., 'Numeric' in the type column) automatically appear. You can change these, using either pop-down menus or by typing in, for example, the label for your values. You can change the characteristics of existing variables in the same way.



quants file_5.sav [DataSet1] - IBI

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	age	Numeric	3	0	age in months	None	999	8	Right	Scale	Input
2	gender	Numeric	1	0		{1, boy}...	9	8	Right	Nominal	Input
3	attsc1	Numeric	1	0	school is boring	{1, agree str...	9	8	Right	Ordinal	Input
4	attsc2	Numeric	1	0	school is fun	{1, disagree ...	9	8	Right	Ordinal	Input
5	attsc3	Numeric	1	0	school is the sa...	{1, agree str...	9	8	Right	Ordinal	Input
6	attsc4	Numeric	1	0	teachers don't t...	{1, agree str...	9	8	Right	Ordinal	Input
7	attsc5	Numeric	1	0	sorry when sch...	{1, disagree ...	9	8	Right	Ordinal	Input
8	scmark	Numeric	2	0	marks for your ...	None	99	8	Right	Scale	Input
9	appsc1	Numeric	1	0	like my body	{1, disagree ...	9	8	Right	Ordinal	Input
10	physsc1	Numeric	1	0	am good at spo...	{1, agree str...	9	8	Right	Ordinal	Input
11	engsc1	Numeric	1	0	I get good mark...	{1, disagree ...	9	8	Right	Ordinal	Input
12	parsc1	Numeric	1	0	don't like to be ...	{1, agree str...	9	8	Right	Ordinal	Input
13	peersc1	Numeric	1	0	make friends ea...	{1, disagree ...	9	8	Right	Ordinal	Input
14	mathsc1	Numeric	1	0	among the best...	{1, disagree ...	9	8	Right	Ordinal	Input
15	schsc1	Numeric	1	0	good marks in ...	{1, disagree ...	9	8	Right	Ordinal	Input
16	appsc2	Numeric	1	0	I'm good looking	{1, disagree ...	9	8	Right	Ordinal	Input
17	schsc2	Numeric	1	0	often don't und...	{1, agree str...	9	8	Right	Ordinal	Input
18	parsc2	Numeric	1	0	find it easy to t...	{1, disagree ...	9	8	Right	Ordinal	Input
19	mathsc2	Numeric	1	0	teachers think I...	{1, disagree ...	9	8	Right	Ordinal	Input
20	physsc2	Numeric	1	0	like to do sports	{1, disagree ...	9	8	Right	Ordinal	Input
21	schsc3	Numeric	1	0	among th best i...	{1, disagree ...	9	8	Right	Ordinal	Input
22	peersc2	Numeric	1	0	don't have man...	{1, agree str...	9	8	Right	Ordinal	Input
23	engsc2	Numeric	1	0	often don't und...	{1, agree str...	9	8	Right	Ordinal	Input
24	parsc3	Numeric	1	0	my parents don...	{1, agree str...	9	8	Right	Ordinal	Input
25	appsc3	Numeric	1	0	don't like the w...	{1, agree str...	9	8	Right	Ordinal	Input
26	mathsc3	Numeric	1	0	good marks in ...	{1, disagree ...	9	8	Right	Ordinal	Input
27	peersc3	Numeric	1	0	other kids like ...	{1, disagree ...	9	8	Right	Ordinal	Input
28	appsc4	Numeric	1	0	have an attracti...	{1, disagree ...	9	8	Right	Ordinal	Input
29	schsc4	Numeric	1	0	teachers think I...	{1, disagree ...	9	8	Right	Ordinal	Input
30	engsc3	Numeric	1	0	among the best...	{1, disagree ...	9	8	Right	Ordinal	Input
31	physsc3	Numeric	1	0	not good t lear...	{1, agree str...	9	8	Right	Ordinal	Input
32	parsc4	Numeric	1	0	have fun with m...	{1, disagree ...	9	8	Right	Ordinal	Input
33	mathsc4	Numeric	1	0	often don't und...	{1, agree str...	9	8	Right	Ordinal	Input
34	physsc4	Numeric	1	0	bad at sports	{1, agree str...	9	8	Right	Ordinal	Input
35	engsc4	Numeric	1	0	teachers think I...	{1, disagree ...	9	8	Right	Ordinal	Input

Figure 5.3 The variable view

Our data set

In this book, we will use a data set collected as part of a study of children in Year 5 of primary school. They were around 10 to 11 years old when the data were collected. Due to grade retention, some may be older.

The aim of the study was to look at the relationship among pupils' achievement at school, their self-concept and their attitudes to school. Data on parental background, gender and some school variables were also collected.

The data were collected by means of a questionnaire given to each child. The researcher personally administered the questionnaire in all cases, usually with the class teacher present.

School achievement was collected from teachers in two subjects (English and maths) and globally (grade point average, or GPA) and was based on the results of teacher-made tests.

Self-concept was conceptualised as hierarchical and multidimensional. Construction of items was based on the seven-factor Shavelson model (see Chapter 4). Four items were constructed to measure each factor. A global self-esteem measure containing nine items was also used.

School attitudes were measured with six items, looking at students' attitudes to both school in general and their teachers. They were also asked to award marks to their school.

Two items measured parental background. One measured the education level achieved by the child's primary carers. The second variable measured their SES by classifying primary carers' occupations based on the International Labour Office's ISCO 88 classification. The data were collected by asking children to give a questionnaire to their parents. Some school data were collected – namely, school type and school environmental quality (a number of quality factors measured by the observer during school visits).

The sample was a random sample of 50 schools. Within these schools, all pupils in one class in Year 5 were surveyed.

The data set has been shortened (there were originally more variables) and cleaned for this book. The labels of a few variables were changed to better preserve the anonymity of schools and respondents.

5.2 Summary

In this chapter, we have introduced our statistics package, SPSS. This is one of the most widely used statistical software packages in the social sciences and is quite user-friendly.

The data set we will be using was based on a survey study of children in Year 5 of primary school. We collected data on their achievement, their attitudes to school and their self-concept and some parental background data.

5.3 Exercises

1 Open the data set. Have a look at the variables, and see if you can add a new variable.
2 Try to change the value labels of one of the variables.

6

UNIVARIATE STATISTICS

Now that we have explored the process of designing quantitative research studies, it is time for us to do some data analysis.

6.1 Introduction

While it may be tempting to start looking at relationships between variables straight away, it is a good idea to look at our individual variables first. We usually need to know how our respondents have replied to particular questions, or how many times a teacher has asked a particular question, for example, before we can look at relationships with other variables. We might often just want to know about individual variables – for example, how many boys and girls there are in our sample. This kind of descriptive information can give us useful information on our variables and our research questions. Because we are looking at individual variables, this type of analysis is called *univariate analysis*. As well as providing important information, univariate analysis can help us to look out for mistakes that may have been made during data input, for example. We can spot some (but obviously not all) errors by seeing whether there are values which are outside the range of possible values (e.g., if we have coded boys as 1 and girls as 2, we wouldn't expect to find any 3s!).

6.2 Frequency distributions

As mentioned above, the first things we want to look at are often things like how many people have answered in a certain way, or how many respondents belong to different ethnic groups, for example. The best way to do this is by looking at what we call a *frequency distribution* of the variable. This is simply a list of all the values that variable has acquired in the sample (e.g., 451 boys and 449 girls). What I would like to know about in our sample data set is how many kids say they think they get good marks in English. Let's have a look at how we can do this in IBM SPSS Statistics.

1 Once we have opened the file, we will need to go to the button marked ‘Analyze’. This is where all the statistical analyses ‘live’.
2 When we click that box, a new one pops up. This lists a whole slew of statistical procedures. We want to choose ‘Descriptive Statistics’, because at this stage we are just going to describe our variable.
3 When we click this, a new box comes up, with a new list of choices. We choose ‘Frequencies’ as that is where we are going to find the frequency tables (Figure 6.1).

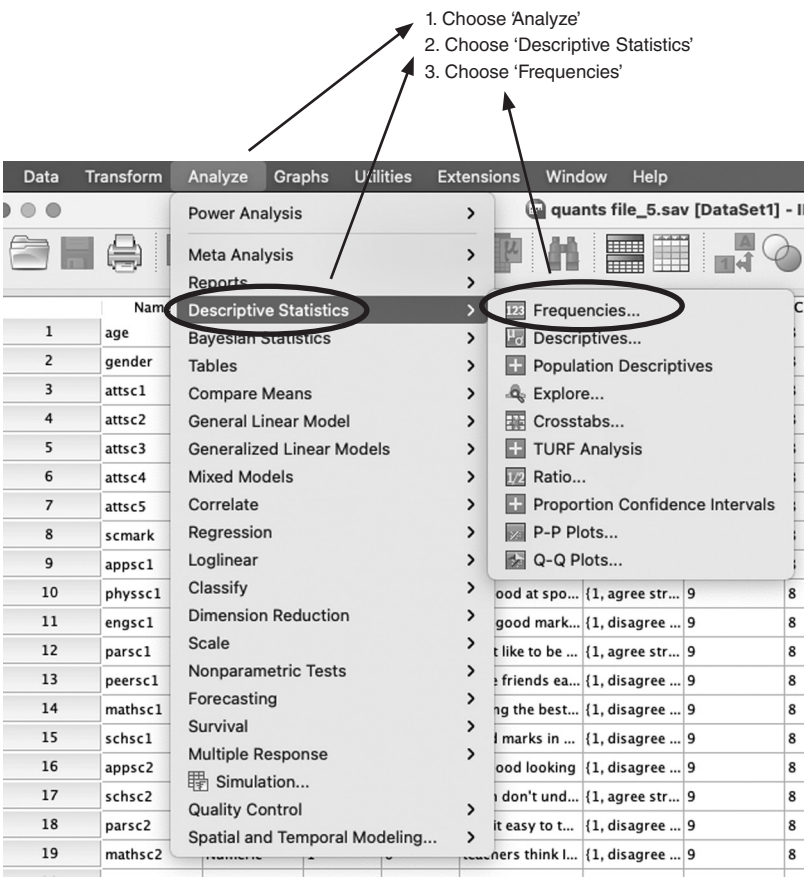


Figure 6.1 Producing a frequency table: steps 1–3

4 We now see a box pop up (Figure 6.2). We will see similar style boxes pop up in most of the analyses we do with SPSS. On the left side within the box, we can see a menu giving us the list of all the variables in our data set. In order to do the analyses,

we will have to highlight the variable(s) we want to look at (we can put up to 100 variables in the right-hand box at once). We do that by clicking on the variable, in this case 'I get good marks in English (engsc1)'.

5 We then need to put that variable in the now empty box on the right. This box needs to contain all the variables we want to include in the analysis. We do this by clicking the arrow in the middle. The variable then jumps to the right-hand box.

6 We click on the button marked 'OK' (Figure 6.2) (you will have seen that here are a number of other buttons, with names like 'Statistics'. We will look at some of those later on).

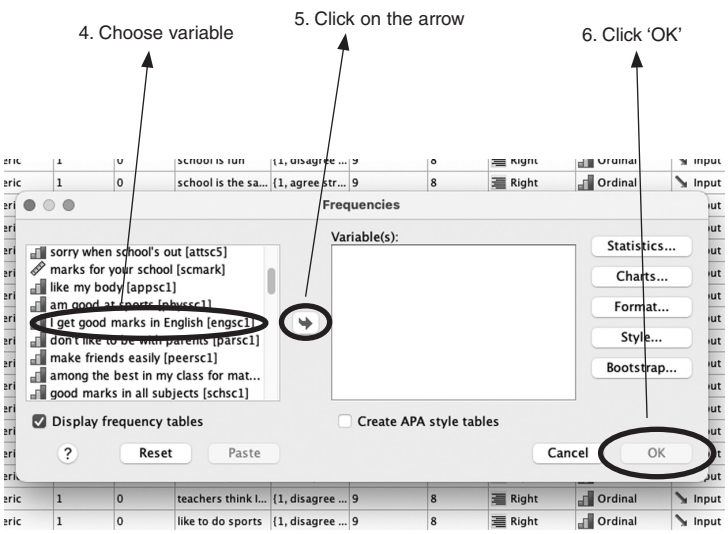


Figure 6.2 Producing a frequency table: steps 4–6

Now SPSS will open a new window, in which the output will appear. There are two main sections to this output (this will usually be the case in SPSS). The first box gives us some general information: the name of the variable, how many respondents actually answered the question (886), and how many did not (those people coded as missing, three in this case) (Figure 6.3).

The second box gives us our actual frequencies. This box has a number of columns:

1 Column 1 gives us the value labels for the variable we are looking at, in this case 'agree strongly', 'agree', 'disagree' and 'disagree strongly'. As you can see, frequencies will also give us the number of missing values (kids who didn't answer) and the total number of people in the sample.

2 The second column gives us our actual frequencies, the number of kids that have responded 'agree strongly' (237 in this case), and so on.

3 The third column expresses that as a percentage.

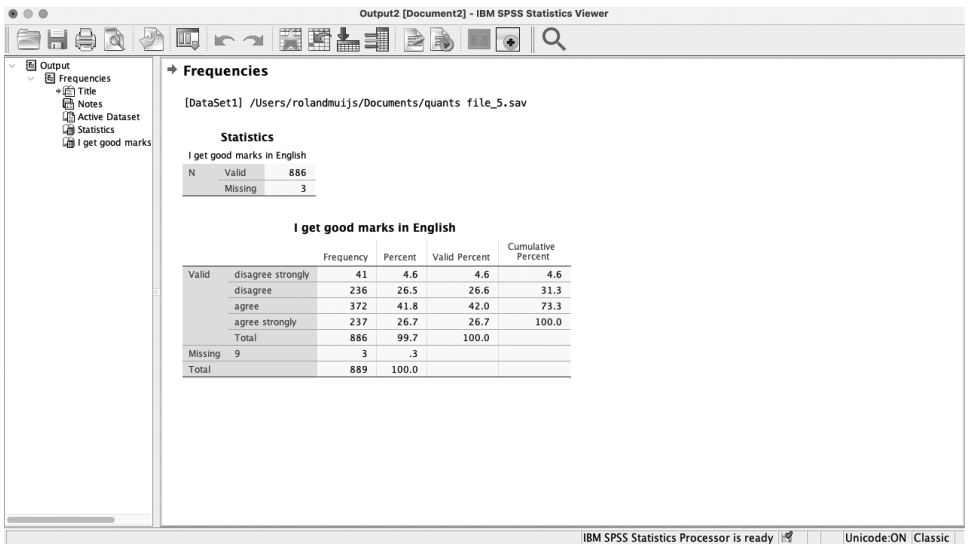


Figure 6.3 Frequencies output

4 The fourth column, labelled 'Valid Percent', gives us the percentage of kids who are not missing (i.e., those who did actually answer the question) for each of the four answer categories.
5 The final column gives us the 'Cumulative Percent'. This just means that the percentages are added up to 100.

This table in Figure 6.3 gives us some interesting information. Promisingly, no values lie outside those that we would expect (the four answer categories and 'Missing'). We can also see that the majority of kids (68.7% – we get this figure by adding the 'agree' and 'agree strongly' totals in the 'Valid Percent' column) think they get good marks in English, but that there is a significant minority who don't think that they get good marks in English. This is obviously a group whom we might want to single out for particular attention or support.

We can also depict the frequencies in graphical form. To do this, we need to go through steps 1 to 5 in the same way as before, but before we go on to click 'OK', we click on the 'Charts' button. A new pop-up screen will appear, which gives us a number of choices, such as pie chart, histogram, and bar chart (Figure 6.4). One of the most useful for us is the bar chart, as this will give us a good indication of the distribution of the variable (is it normal?). We can choose this option by ticking the relevant box. Then we click on 'Continue'. We then get back to the original frequencies box, and click 'OK' (Figure 6.4). As we can see from Figure 6.5, in this case, our variable is *skewed positively*; that is, most of the values are positive, rather than lying in the middle.

Obviously, the frequency table gives us important information about each individual variable. But often, we want to be able to 'summarise' our variable, using one number that

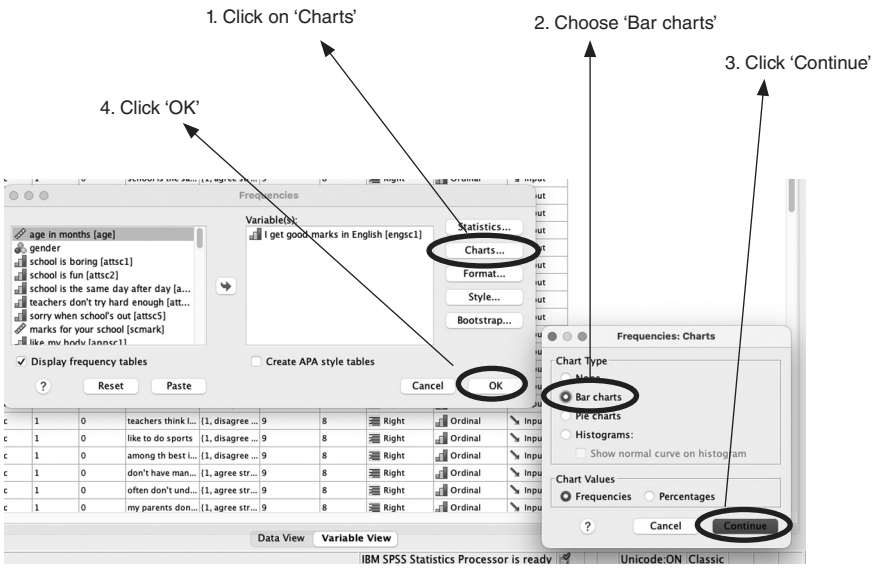


Figure 6.4 Getting charts in 'Frequencies'

represents the most 'typical' value. This is especially important where we are using variables with a whole load of possible answer categories (think of test scores, for example). To do this, we use a measure of *central tendency*, commonly known as an average. In a moment, we are going to look at how we can get SPSS to calculate an average for us. But before we do this, we need to take a look at something called 'levels of measurement'.

6.3 Levels of measurement

Levels of measurement are basically categories of variables. This categorisation is important, because it fundamentally affects the meaning of the variables and what we can do with them statistically, as we will see. There are three basic levels of measurement (some authors distinguish four, but for all practical purposes a distinction of three categories is sufficient): *nominal*, *ordinal* and *continuous*.

Nominal variables are measured at the lowest level. These are variables such as gender, ethnicity and place of birth, where any numbers we give to the values (e.g., 1 for boys and 2 for girls) only serve to replace a name. The values cannot be placed in order. We can't say 'a girl is more than a boy', for example, so in this case, we can't say 2 is more than 1. Nominal variables just have categories, which can't be ordered in any way. Any numbers given are merely a descriptor of that category (e.g., 1 = 'boy').

Ordinal variables do possess a natural ordering of categories. An example of an ordinal variable is the one we were looking at earlier, 'I get good marks in English'. Here, a code of

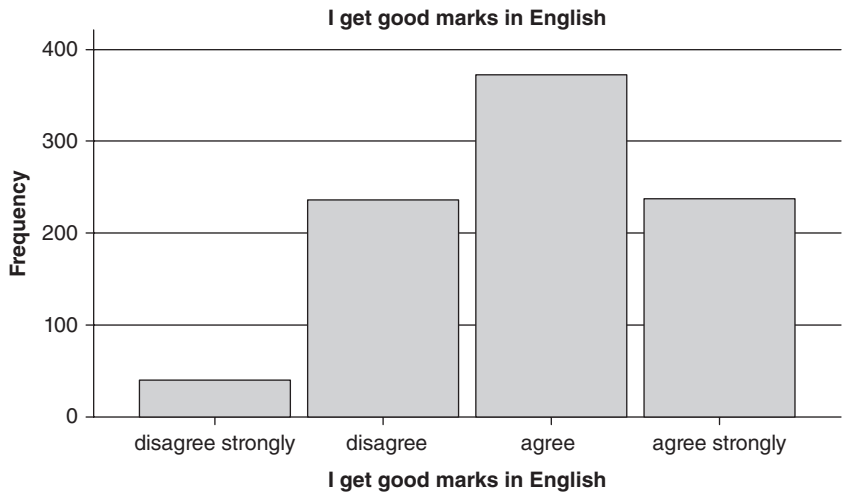


Figure 6.5 A bar chart

4 was given to 'agree strongly', 3 to 'agree', 2 to 'disagree' and 1 to 'disagree strongly'. These values can clearly be ordered, in that someone who agrees strongly 'agrees more' than someone who simply agrees, and so on. This is different from the situation with gender. Therefore, ordinal variables allow you to 'order' the values given. What you can't do is 'measure' exactly the distance between the scale points. Let me explain what I mean by that. When you have a ruler, you know that the distance between 23 and 24 centimetres is exactly the same as the distance between 10 and 11 centimetres; that is, 1 centimetre. This is not the case when we look at the variable, 'I get good marks in English'. Is the distance between 'agree strongly' and 'agree' the same as between 'agree' and 'disagree'? In order to know this, we would have to find out how people thought about these categories; that is, are these differences the same or different in respondents' minds? And does this differ between different respondents? As we cannot know this, we cannot assume that the distance between each scale point is exactly the same as it is on a ruler. All these agree–disagree type variables are therefore *ordinal*.

Continuous variables are those variables that do behave like a ruler. Not only can we order the categories, but the distance between each scale point is the same. They are measured on a continuous scale, like temperature, weight, or height. What variables in educational research are like that? Variables that are often considered to be continuous are scores on a standardised test, such as the SAT (US Scholastic Aptitude Test). Some authors would argue with that, though, saying that in some cases the distance between scores at the midpoint of the scale may not be quite the same as that on the high and low points, but conventionally, this type of variable will be considered continuous. We also sometimes want to look at variables such as age or birth weight, which are also continuous.

Whether a variable is nominal, ordinal, or continuous has important consequences for what types of analyses we can do with it, and how we can interpret the variable, as we will see below.

Do ordinal and nominal variables really constitute measurement?

Some researchers say that nominal and ordinal variables are not real measures, in the sense that measurement is understood in the natural sciences. They say real measurement means that variables must be continuous and conform to mathematical measurement models, and more particularly the Rasch model.

Measures that do not conform to this model are seen as impeding the progress of scientific advancement in the social and behavioural sciences. Obviously, these researchers are not talking about variables such as gender, but about measures of achievement, psychological constructs (e.g., self-concept) and attitudes. These researchers believe that if we were to use the Rasch model to develop our measurement instruments, instead of using existing ordinal variables, we would be able to improve the behavioural sciences to close to what they see as the high level of the natural sciences.

A good overview of these arguments and practical applications of the Rasch model are given in Bond and Fox (2002).

6.4 Measures of central tendency

6.4.1 The mean, the median and the mode

Now we have discussed levels of measurement, we can have a look at some measures of central tendency, or average.

Usually, when we speak about average in everyday terms, the value we are thinking of is the *mean*. The mean is simply *the sum of the values of all the cases divided by the total number of cases*. For example, if we had the following data set of the height in centimetres of people in a class (see Table 6.1), we would calculate the mean by adding all the heights (= 1,441) and dividing that by the number of people (8), giving us a mean height of 180.125 centimetres.

Although this is what we commonly mean when we talk about an average in daily life, this type of average actually only works with one type of variable, continuous variables. Let's think about this. Imagine that we were to take the mean of a nominal variable, let's say gender (see Table 6.1). If we calculate the mean gender, we get a value of 1.44. What does this mean? Is our average person a hermaphrodite with slightly more male than female features? We cannot have such an actual person in our data set. This value is essentially meaningless. This will be the same for all nominal variables. Imagine that we took birthplace. We could calculate a mean, say, 4.6, but what would that mean? Someone who came from between New Jersey and New York but was closer to New Jersey? The same problem occurs when we use ordinal variables. The final column in

Table 6.1 Height, gender and whether respondents like their job

Case number	Height (cm)	Gender (1 = female, 2 = male)	Likes their job (4 = agree strongly, 3 = agree, 2 = disagree, 1 = disagree strongly)
1	167	1	1
2	178	1	4
3	189	2	3
4	201	2	3
5	182	1	2
6	175	2	4
7	162	1	1
8	187	2	2
9	180	1	3
Total	1,441		

Table 6.1 is an ordinal variable, the answers to a scale asking them whether they liked their job. The mean value is 2.56. But that value does not correspond to any of the answers. It is somewhere between 'agree' and 'disagree', but a bit closer to 'agree'. This doesn't make too much sense either.

Therefore, we will also want to use other types of measures of central tendency.

One of these is the *median*. The median is essentially the middle category of a distribution. We can find that by ordering our values from low to high, and then seeing which one the middle one is.

Table 6.2 Height, gender and whether respondents like their job ordered by height

Case number	Height (cm)	Gender (1 = female, 2 = male)	Likes their job (4 = agree strongly, 3 = agree, 2 = disagree, 1 = disagree strongly)
1	162	1	1
2	167	1	1
3	175	2	4
4	178	1	4
5	180	1	3
6	182	1	2
7	187	2	2
8	189	2	3
9	201	2	3

In Table 6.2, we have ordered heights from low to high. To find the median, we have to look at which is the middle value. As we have nine observations, our middle category is number 5, or 180 centimetres, a value that is very similar to the mean. This type of average is most suitable for ordinal variables, as it is based on the principle of ordering that is typical of ordinal variables. Here, for example, when we try the same thing for the 'like my job' variable (try this if you want), we find that the median is 3, a far more sensible value that actually corresponds to a real value ('agree'). This one still doesn't work for the nominal variables, though, as we can't sensibly order those (think about ordering birthplaces).

How do we calculate the median if we have an even number of cases?

If we have an even number of cases, then the median will be a hypothetical value lying between the two middle cases in the distribution. For example, if we have the following set of data:

2 4 6 8 10 12

our median would be the value that lies between 6 and 8 (the mean of those two middle values, in fact) – that is, 7.

That is why there is a final type of average, the *mode*. The mode is simply the most common value. In our example of gender above, there are five females and four males, so the modal value is female.

Does that mean that whenever we have a continuous variable, we use the mean; whenever we have an ordinal variable, the median; and whenever we have a nominal variable, the mode? Almost, but it is not quite as simple as that. There is one situation where we might want to use the median for continuous variables as well. Table 6.3 gives the fictional distribution of wages in an organisation.

Wages is clearly a continuous variable, and therefore the mean would seem to be the sensible method to use if we want to calculate the most typical value or average. The mean here is 82,000 (902,000/11). When we look at this figure, though, something slightly peculiar seems to have happened. This mean wage is higher than that of 10 out of the 11 employees of this organisation. What is going on here is that one person (let's, for the sake of argument, refer to this individual as the vice chancellor) in the organisation is earning a whole lot more than anyone else. This is what we call an *outlier*. Because the wages of the vice chancellor lie so far outside the range of the other values, the mean is pulled towards that high value, and is no longer really representative. Where such outliers exist, it can be better to use the median, even with continuous variables (the median in this case is 32,600, a far more representative value).

Table 6.3 Wages in an organisation

Observation	Wages ($)
1	27,900
2	38,400
3	20,100
4	26,400
5	60,000
6	42,600
7	22,700
8	55,700
9	550,000
10	25,600
11	32,600
Total	902,000

6.4.2 Calculating measures of central tendency in SPSS

Let's look at how we can generate some measures of central tendency in SPSS. When we ask the program to give us a frequency table, we can also ask it to give us some measures of central tendency as well, so we will start by looking at the frequencies for the variable, 'I think I'm good at English'.

We want to start once again with steps 1 to 3 above (go into 'Analyze', choose 'Descriptives', and then choose 'Frequencies'). As we know, a box now appears in which we have to select the variable (step 4), and click the arrow (step 5 above) to add it to the list of variables we are going to analyse. Before pressing 'OK', we can now look at one of the other buttons on the bottom of the screen, which is labelled 'Statistics' (Figure 6.6).

When we press this button, a new screen appears that gives us a number of options (Figure 6.6). On the left it says 'Central Tendency'. There are a number of measures we can choose. We need to tick the boxes for each measure we want. As you can see, the mode, median and mean are all given as options. Let's tick all three. Once we have done that, we can click 'Continue', and then 'OK' in the main panel, and our output will appear (Figure 6.7).

We can now see that as well as the output we got last time we used frequencies, we now have a new set of items to look at.

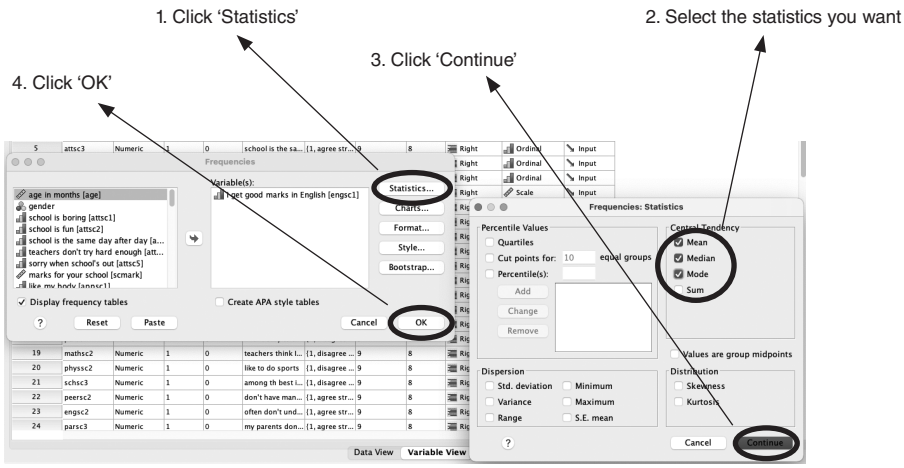


Figure 6.6 Measures of central tendency in SPSS

Statistics

I get good marks in English

N	Valid	886
	Missing	3
Mean		2.91
Median		3.00
Mode		3

Figure 6.7 Measures of central tendency output

The first box in the output gives us our measures of central tendency: the mode, median and mean. We can see here that our mode is 3. This is the value that is most common, in this case the answer that most respondents have chosen (agree). The median is also 3. This is the middle value of the distribution, once we have ordered all our answers from the lowest to the highest. Finally, our mean is 2.91. This value does not actually correspond with any real answers. This is because 'I get good marks in English' is an ordinal variable.

6.5 Measures of spread

6.5.1 Range, interquartile range and standard deviation

Measures of central tendency give us one set of important information when it comes to describing our variables. They don't tell us the whole story, though. Take for example the two examples of (fictional) test scores in two schools shown in Table 6.4.

Table 6.4 Test scores in two schools

Case number	School 1	School 2
1	45	60
2	50	65
3	55	65
4	60	70
5	65	70
6	70	70
7	70	70
8	75	70
9	80	70
10	85	75
11	90	75
12	95	80
Mean	70	70
Median	70	70

The median and mean scores for both schools are equal, at 70. This could lead us to conclude that both schools have equal patterns of attainment. However, if we look at the data more closely, we see that there is clearly more going on than that. While measures of central tendency are the same, they have been arrived at in rather different ways. In school 1 there is quite a spread of values, going from 45 to 95, while in school 2 all pupils seem to have scores that are closer together, the lowest being 60 and the highest being 80, with six pupils getting 70. If, on the basis of measures of central tendency, we conclude that attainment in both schools is similar, we would be missing some important distinctions here. The spreads of the values around the mean or median are clearly different.

That's why, as well as measures of central tendency, we also need measures of spread if we are to give a good description of our variables.

The first way of looking at spread seems obvious: Why not just subtract the lowest from the highest scores, giving us the *range* of values in our data set? If we do this with the example in Table 6.4, it gives us a spread of 50 for school 1 and a spread of 20 for school 2, capturing pretty well the distinction between the two. This measure doesn't always work that well, though. Think about our example in Table 6.3, where we looked at wages in an organisation. If we took the range there, subtracting the lowest from the highest value, we would end up with 529,500. This seems like a massive range, suggesting that the values lie spread out far from the mean. When we look at the data more closely, though, this is not really the case. Rather, it is once again the one outlier that is distorting this statistic by making the measure of spread seem larger than it should.

What can we do to solve this problem? One common method is to use a measure known as the *interquartile range*. The interquartile range is calculated by first ordering the sample from low to high, and then dividing it into four quarters (see Table 6.5). We then need to calculate the third and first quartiles. The first quartile is given by the first dotted line in Table 6.5. It lies between 55 and 60 in school 1 (we take the mean of those two values, as we did for the median where we have an even number of cases), and is 57.5. In the second school it lies between 65 and 70 (67.5). Then we calculate the third quartile. This is between 80 and 85 in school 1 (82.5) and between 70 and 75 (72.5) in school 2. We can then finally calculate the interquartile range by subtracting the first from the third quartile.

Table 6.5 Calculating the interquartile range

Case number	School 1	School 2
1	45	60
2	50	65
3	55	65
4	60	70
5	65	70
6	70	70
7	75	70
8	75	70
9	80	70
10	85	75
11	90	75
12	95	80
Mean	70	70
Median	70	70

School 1: Q3 – Q1 = 82.5 – 57.5 = 25

School 2: Q3 – Q1 = 72.5 – 67.5 = 5

Again we see that the spread in school 2 is far smaller than in school 1.

This measure is less likely to be distorted by outliers than the range, as it cuts out all extreme values at the top and bottom of the distribution. However, a disadvantage of this method is that it only uses a small amount of the information that could be used, as we are only looking at two values when calculating the range. A measure that does use all the information we have, because it takes all values into account rather than just two, is the *standard deviation*.

The standard deviation (*SD*) is a measure of the extent to which the values in a distribution cluster around the mean. It is related to a value called the *variance*, which you might also encounter. In fact, the standard deviation equals the square root of the variance. The variance, in turn, is the sum of squared deviations of the observations from their mean divided by the number of observations minus 1. You needn't worry too much about that, but what this basically means is that the variance is calculated by looking at the extent to which each observation differs from the mean. This implies that the standard deviation (and the variance, of course) can only be calculated where we can calculate a mean. Therefore, we can only calculate a standard deviation for continuous variables. With ordinal variables, it is better to use the range. If we have nominal variables, it doesn't make sense to calculate measures of spread.

In Table 6.5, we would find a standard deviation of 14.6 for school 1 and of 4.8 for school 2, again showing clearly the difference in patterns of responses between the two.

6.5.2 Calculating measures of spread in SPSS

How can we calculate measures of spread by using SPSS? Let's look again at our variable, 'I think I am good at English'. Again, we don't need to look further than our frequencies we used earlier. We can once again go through steps 1 to 5 ('Analyze', 'Descriptives', 'Frequencies', select 'Variables' and click arrow) and, as with the measures of central tendency, click on the 'Statistics' box (Figure 6.8). We can then see that in that same box, as well as there being the option to check a number of measures of central tendency, we can also tick boxes for a number of measures of spread. We will choose 'Range' and 'Std. deviation', and click 'Continue'. Once we have clicked OK in the main box, the output will appear (Figure 6.9).

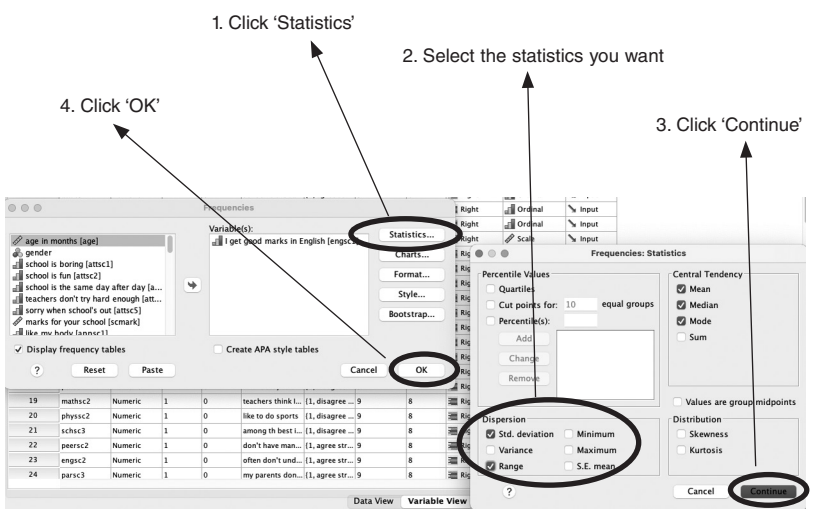


Figure 6.8 Measures of spread

Statistics

I get good marks in English

N	Valid	886
	Missing	3
Mean		2.91
Median		3.00
Mode		3
Std. Deviation		.843
Range		3

Figure 6.9 Output measures of spread

If we look at the output, we can see that along with the measures of central tendency, we now also have a number of measures of spread. The first measure given in the top box is the standard deviation. This is 0.843. In a large sample, approximately 68% of respondents will lie 1 *SD* from the mean. We know that the mean was 2.91. Therefore, 68% of observations are likely to lie between 2.91 – 0.843 (=2.067) and 2.91 + 0.843 (=3.753), and 95% of observations are likely to lie within 2 *SD* of the mean. The problem in this case, though, is that these values do not correspond to any actual responses that could be part of this agree strongly–disagree strongly type scale. This is because this is an ordinal variable. A better measure here is the range. This is 3, which corresponds with the difference between the highest and the lowest value.

Common misconceptions

1 *If a variable is measured in numbers, we can order it, can't we?* Not necessarily. When we use statistics, we have to assign numbers to our categories in order to do calculations. In some cases, these numbers are merely a replacement for an unorderable label, such as place of birth. We could assign 1 to France, 2 to Spain, 3 to Britain and so on, but that doesn't mean that we could order them in any way.

2 *'Average' and 'mean' are the same thing, aren't they?* In daily life, when we talk about average, we are usually referring to the mean. In statistics, however, the mean is actually only one possible average. The mode and the median are also averages.

3 *When we have continuous variables, we always use the mean as the measure of central tendency, don't we?* Not necessarily. The mean is not always the best measure of central tendency for continuous variables. Outliers (extreme cases) can distort the mean, as we

saw in Table 6.3. When we have such outliers, the median may be a more accurate representation of central tendency even for continuous variables.

4 *When we have nominal variables, we use the range as our measure of spread, don't we?* No. When we have nominal variables, the concept of spread is meaningless. As we can't order the categories, the concept of them being spread (around the mode) is not a useful one.

6.6 Summary

In this chapter, we have looked at describing single variables. This is called univariate analysis. One of the most obvious (and most important) things to do at the start of an analysis is to look at the frequency distribution of the variables. As well as looking at the frequency distribution, we will usually want to be able to describe the most typical or average case or response. To do this, we calculate measures of central tendency.

In order to be able to do this, we need to know at what level our variable is measured. There are three levels of measurement: nominal, ordinal and continuous. Nominal variables, like ethnicity, don't allow us to order categories. Any numbers or categories we assign are just labels. Ordinal variables allow us to order categories from low to high or from less to more (or disagreement to agreement), but we can't measure precisely what the distance is between scale points. A typical ordinal variable is an agree–disagree type scale. Continuous variables allow us both to order categories and to say that the distance between all categories is exactly the same (like measuring length with a tape measure).

There are three measures of central tendency that go along with these three levels of measurement. The *mode* is the most common value in a data set. It is the most suitable measure for nominal variables. The *median* is the middle value in a set of data ordered from low to high. It is the best measure of central tendency for ordinal variables. The *mean* is the sum of all values divided by the number of observations. This is the best measure for continuous variables (except where there are outliers, when it may be better to use the median).

As well as measures of central tendency, we often want to look at measures of spread of the values around the centre. The *range* is simply the difference between the highest and the lowest value. As it is sensitive to outliers, we often use the *interquartile range* instead. This is the difference between the third and first quartiles. Both are good measures when we are using ordinal variables. A measure that makes better use of all the information we have is the *standard deviation*. This is a measure of the spread of all the values around the mean, and is best suited for continuous variables. The usage of all these measures is summarised in Table 6.6.

Table 6.6 Summary of measures and central tendency and spread

Level of measurement	Central tendency	Spread
Nominal	Mode	~
Ordinal	Median	Range
		Interquartile range
Continuous	Mean	Variance
	Median (if outliers are a problem)	Standard deviation

6.7 Exercises

1 Have a look at the data file on the website. Can you find an example of a nominal, an ordinal and a continuous variable?
2 Can you have a look at the frequency distributions for the variables, 'I like going to school' and 'school is boring'? What can you say about these two variables?
3 Can you compare the central tendency and spread of the two variables, 'I like going to school' and 'school is boring'? Which measures do you use, and what do they tell you?
4 Can you compare central tendency for grades in maths and English? What measure do you use? What does this tell you?
5 Can you compare the spread of the variables for grades in maths and English? What measures do you use? What do they tell you?

6.8 Further reading

Any basic statistics text will contain a section on measures of central tendency and spread. For a more mathematical treatment than we have given here, the following text is a good one: Carslon, K., & Winquist, J. (2021), *An introduction to statistics: An active learning approach* (3rd ed.) (Sage).