Block 11: Getting started with SPSS

Important! The following block is **not** examinable, however to impress potential employers with your computing skills, you are strongly advised to familiarise yourselves with SPSS, replicating many of the analyses seen later in the course. You can then claim a degree of SPSS proficiency on your CV.

SPSS is among the most widely-used programmes for statistical analysis in the social sciences. It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organisations and others. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored in the datafile) are features of the base software.

Learning Objectives

In this block we look at getting started with SPSS, familiarising ourselves with the SPSS environment and basic data management issues.

Reading List

Malhotra, N.K., D. Nunan and D.F. Birks. Marketing Research: An Applied Approach. (Pearson, 2017) 5th edition [ISBN 9781292103129] Chapter 15.

11.1 Getting started with SPSS

For each section of *Getting started with SPSS*, use the LSE ELearning resources to test your knowledge with the Key terms and concepts flip cards.

Obtaining SPSS Statistics software

While you will not be examined on your ability to *use* SPSS, to get the most out of SPSS it is advisable to access a 14-day free trial version of SPSS Statistics through IBM available here.

For students who really want to gain proficiency in using SPSS, you may wish to invest in a student licence version, accessible here. **Note that purchasing a student licence version of the software is optional!**

Note also that SPSS comes in various editions: Base, Standard, Premium etc. The **Base edition** is sufficient for the techniques covered in this course.

Our example SPSS dataset

We illustrate the basic functionality of SPSS using the dataset *Benetton.sav*, which will be used to demonstrate several multivariate statistical techniques in subseqeuent blocks.

The data come from a survey pre-test, obtained from 45 participants on Benetton clothes. The variables are usage, gender, awareness, attitude, preference, intention and loyalty toward Benetton of a sample of Benetton users. Usage was coded as 1, 2 or 3, representing light, medium or heavy users,

respectively. Gender was coded as 1 for females and 2 for males. Awareness, attitude, preference, intention and loyalty were measured on a 7-point Likert-type scale (1 = Very unfavourable, 7 = Very favourable). Note that five participants have missing values which are denoted by 9.

An Excel version of the dataset is *Benetton.xlsx*.

11.2 SPSS Orientation

For each section of *SPSS Orientation*, use the LSE ELearning resources to test your knowledge with the Key terms and concepts flip cards.

Getting started with SPSS

https://emfssvideo.s3.amazonaws.com/MT%26ST/ST3188/Screencasts/Block_11_Getting_started_with_SPSS.mp4

Click on the link to watch the video, or read the transcript below.

[music]

Speaker: Hello and welcome to getting started with SPSS. Hopefully, you've had the opportunity to download this data file Benetton.sav from the VLE and saved it into an appropriate folder on your computer. This file extension named .sav, this indicates it's an SPSS statistics data document i.e. a data file within the SPSS format. If one double-clicks the file name then this will launch the data set within SPSS.

You can see it has now loaded and we begin with the data editor and its different features. Towards the top, this is where we have the menu. The menu is where you select the actions that you wish to perform e.g. opening a new data files and as we shall see in future blocks, performing various types of statistical analysis as well as perhaps producing simple chats.

If we navigate down to the bottom left of the screen, you will see two tabs, the data view and variable view. You can see that my data set has launched within the variable view, but I'm just going to turn to the data view for a moment. The data view is where the raw data appear. Indeed if you needed to enter data into SPSS you would do in the data view.

As far as data management is concerned, each row corresponds to one case or one unit of analysis. For example, in a market research setting, as with this Benetton data set, each row corresponds to one participant in a market research survey. Column-wise, these all the variables. Each column represents one variable and you can see the names listed there at the topmost row.

Turning to the variable view, this is where we see the attributes and characteristics of each of the variables. Within the variable view, each row corresponds to a variable. For example in the third row here, this corresponds to the gender variable. We're now going to consider some of the most important attributes within the variable view.

We begin with the variable name, each variable in the data set has to have a name with a length that does not exceed 64 bytes. The variable names need to begin with a letter and cannot have spaces or special characters. The names listed here, number, usage, gender, et cetera, if we now just go back to

the day to view briefly, these are the variable names which appear as the headers for each of these columns.

Moving across we have type. The default here is numeric which means that only numbers can be entered into the cells. This does not necessarily mean that the variable is a numeric one. Rather, numeric means that numbers are used to define different values of the variable. In some cases like gender, we might want to use the number one for female and two for male. In this case, our variable is numeric even though the number is only used to label a category and has no arithmetic meaning in itself.

The default here as we see is a numeric. Other variable types you may come across in the future, in particular, might be string, which allows you to enter a text, potentially also things like dates or custom currencies.

Moving across, we now have the variable width and decimal columns. The width determines the maximum number of characters that will be displayed for the variable in all outputs. This is especially relevant when you are using string and enter text. Decimals indicate the number of decimal places that will be displayed and is relevant for numeric data. In all instances here the decimals entry is equal to zero and hence you will see that there are no decimal places in the data view.

However, if I increase decimals here to two and return to the data view, you can see now in this number column all of the entries now are to two decimal places and if I simply revert that back to zero you can see now that we have zero decimal places.

The next column says label and variable labels allow adding more information for each variable. In contrast to the variable name, variable labels have no restrictions on using symbols and spaces and the maximum length is considerably larger at 255 characters. In this instance, you can see that there is exact agreement between the name and the label really here because the names are fairly self-explanatory. In other data sets, you may come across, you may see a much longer descriptor in the label column.

Moving across to the values. It may be that you wish to assign labels to each or some of the values of the variable. In this instance, if we consider the usage variable, this has the numeric type and there are three distinct values 1, 2, and 3. We may wish to assign labels to each value particularly when working with nominal or ordinal variables where the numbers are used to represent categories of a variable. This allows us to perhaps clarify what the particular numbers mean. In this case, one corresponds to light, two corresponds to medium and three corresponds to heavy.

If for example, we had an additional value of four, which we actually don't, in our small data set here, but we could add a new label, for example, four might corresponds to very heavy. If I click add, then this adds very heavy as a label to the entry of four. I'll just cancel this because I don't want to include that particular label. If I expand across here, you can see we can now increase the width of this values column.

Turning to agenda, here we have the classifications of female and male where one has been coded to represent female and two has been coded to represent male. A very important column, as far as market research data is concerned, is that of missing. This is where you define the values that should be excluded from the analysis. For example, you might not want to consider don't know responses or it might be that some answers are not applicable.

I appreciate that non-response is a huge issue in market research data. It may well be that you need to indicate to SPSS when missing values or item non-responses have occurred. While SPSS considers an empty cell as a system missing value, user-defined missing values can provide more information for your research. For example, the distinction between those who do not know and those who refuse to answer. To define missing values, you type in the numbers that have been defined as missing values.

Just turning back to the data view for five variables of awareness, attitude, preference, intention, and loyalty, there are some instances of item non-response. These data entries of nine do not correspond to observations of nine as each of these five variables are on a one to seven Likert scale. If SPSS were to encounter these values of nine, I don't want them to be included in any statistical analysis, I simply want them to be excluded. This is why in our variable view, you can see that nine appears in this missing column for these five variables. This instructs SPSS that whenever it encounters the value of nine to exclude it from any calculations.

Moving across we then have columns and align, these are not that interesting. They refer to the display of the data in that data view. The default is eight characters for each column and to right align numeric variables and left align string variables. We will ignore the role column and just complete now with a discussion of the measure column. This allows us to define the measurement level of variables and SPSS works with three measurement levels. Nominal, ordinal, and scale.

For nominal variables, this means different values represent different categories of the variable, where the values are unordered. For example, gender being male and female, say. For ordinal variables, values are ordered in terms of some degree but they do not establish numeric differences between data points. For example, at level of agreement or disagreement with some statement in a market research survey, let's say from disagree to neither agree nor disagree to agree, would be an ordinal variable.

Finally, at scale, SBSS does not differentiate between interval and ratio levels of measurement as you have learned previously in this course. Rather, it lumps them together so both of these quantitative variable types are denoted as scale variables. Here, values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples might be age and weight.

Now in this instance, you can see that all variables have the default tier of scale. This is fine as a starting point but we may wish to override some of these, for example, our gender variable, we know gender is of a nominal level of measurement, so we can change this attribute to be as nominal.

In this recording, we've just introduced how to launch an SBSS data set and considered both the data view and the variable view so you are clear about the different attributes of each variable in your data set.

[00:12:01] [END OF AUDIO]

Managing your dataset

https://emfssvideo.s3.amazonaws.com/MT%26ST/ST3188/Screencasts/Block_11_managing_your_dataset.mp4

Click on the link to watch the video, or read the transcript below.

[music]

Speaker 1: In this recording, we are going to consider some issues about managing your data set. SPSS provides a range of options to manage your data set under the option "data". Here we're just going to consider two particularly useful cases, which are split file and select cases. Of course, you are welcome to explore other options in your own time. The split file option allows you to divide the file in groups to run analyses by group. For example, it can be used to calculate descriptive statistics separately for males and females.

Now, once this split file option is activated, all analyses will be run separately for the specified groups. The default is to analyze all cases. But let's suppose we did wish to split the file by gender. Here we wish to compare groups, and we're going to do this using the gender variable. The variable by which we wish to split the file, we simply highlight it. If we click the arrow here, it will now look at arranging groups based on this variable of gender. If I now click "okay", it will now split the file by gender. Now that we've actually conducted something, i.e., we've instructed SPSS to perform some routine, we now see this output window.

This is simply going to be the window where the results of various analyses will be displayed. Don't worry too much about what's appearing here. This is, if you like, just an audit trail of what we've asked SPSS to do. Our data file has now been split by gender. If you look at the very bottom right-hand corner of your data view, it will say whether your file is actively being split. In this case, it is being split by gender. It's important to remember to look there because until we deactivate the split file by gender, it will continue to conduct analyses separately. Now, let's actually do a simple statistical analysis to showcase how the split file option is working.

If we click "analyze", this is our main navigation route to conduct various statistical routines. We're going to walk before we can run. We simply want to calculate some descriptive statistics on some particular variables. The "descriptives" dialog box has now appeared, and in the left-most box are listed all of the variables within our data set. You can start to see the easy-to-use interface of SPSS. It has a somewhat intuitive environment. Here we can simply select the variables upon which we wish to calculate some descriptive statistics. Now it is possible to highlight these one-byone, for example, "awareness and attitude" and manually move them into the rightmost box on an individual basis.

Alternatively, you could hold down "shift" and highlight multiple variables, and in one click move them all across. In this case, we're going to calculate some descriptive statistics for those five variables from awareness down to loyalty. If we click the "options" tab, you can see here that checked are the default descriptive statistics which are going to be returned. These tend to be the most common and useful ones, specifically the mean as our main measure of central tendency or location. You can also see some dispersion descriptive statistics have been selected by default, namely the standard deviation, also minimum and maximum.

If we wished, we could also ask it to report the variance range and the standard era of the mean, but I'm just going to stick to the default options here. Kurtosis and skewness, if we wish to investigate those attributes of a distribution, but they tend to be of far less importance compared to things like the mean and the standard deviation. If I click "continue" and click "okay", we now return to our output window. Here we can see the results of our first statistical analysis. Because the "split file by gender" option has been activated, you can see SPSS has calculated descriptive statistics separately for females and males.

The different variables are itemized. The number of observations N is reported, the minimum value of the variable, the maximum value of the variable also the mean and standard deviation to two and three

decimal places respectively. Now, note that if we return back to our data view, recall that for each of those variables, awareness to loyalty, we did have some missing values, some missing observations, specifically the code, for example, there of nine for attitude. In the "variable" view, we instructed SPSS to treat nine as a missing value. Indeed, that has occurred because in our output view, the maximum value reported for any of these variables across the males and females is indeed seven.

Let's deactivate the split file option to see all of the cases combined back together. We could go back to data and split file, but there is a bit of a shortcut. Here we can click "split file". This returns us split file dialog box. If we simply click "reset", this returns it to the default setting where all cases are considered. If I "okay", this is now turning the split file off and back in your data or variable view. The split file by gender has now disappeared. If we wish to conduct the same analysis of descriptive statistics, you can see those five variables have been retained. If I click "okay", you can see they've now all been combined together.

The total number of observations is 44 because we had one missing value or a case of a missing value for each of those five variables. I said we would consider another option within the data menu option here. This is going to be "select cases". This function allows you to select part of the data set and to exclude the remaining cases from the analyses. For example, we might only be interested in female participants in our survey. To select some cases, we need to select "data" and select cases through the menu options. Then you can see the default is that all cases are selected.

Let's suppose we wanted to make this conditional, so we wish to select cases when a particular condition is satisfied. By clicking "if", we are now in a position to instruct SPSS what condition we wish to be satisfied. Here, if I only wish to consider females, that means I wish to take the gender variable. Whenever this is equal to one, and by our variable coding, one corresponded to the female participants, then we will only consider those cases in our analysis. If I click "continue" and then "okay", it's indeed going to filter to only consider the case when gender is equal to one.

Don't worry too much about the output appearing here in our output window. Think of this as your computer-speak for formalizing what we've instructed it to do using the graphical user interface. Turning back to our data view, you can see the active cases do not have a strike through. Whereas, which were all of the female participants, so those with the gender of value there of one. The cases where the gender value is two corresponds to the male participants, you can see the strike through here which indicates that they will be excluded from any analysis. Just to verify that, let's calculate the descriptive statistics for the selected cases.

If I click "okay" here, it's going to run the descriptive statistics only on those female participants. Indeed, if you look at the number of observations here, this clearly means it has excluded the males. Those are a couple of interesting features within our data tab. It is possible to weight cases by assigning differential weights to observations. However, this is generally not advisable to do because it can be very subjective about which weight to apply.

However, you may wish to apply weights for sampling reasons if it so happens, let's say, that one gender happened to be oversampled, i.e., suppose there was a larger proportion of females in your sample than in the population. Using weights allows you to give oversampled cases less weight and undersampled cases more weight to try to resemble the population proportions and get a more representatively weighted sample. This is just to give you a little illustration about how you can manage your data set.

[00:10:42] [END OF AUDIO]

Data operations

[music]

Speaker: In this recording, we're going to consider some of the data operations which exist within SPSS. SPSS provides many different functions which allow us to create new variables from existing variables. For example, you can use SPSS to recode variables into categories or perhaps create an average of variables. These functions can be found under the transform option. Again, there are many here. We won't consider them all. We'll simply consider computing variables as well as recoding into different variables. Let's begin with computing a new variable.

Let's suppose from our original variables in our Benetton data set, imagine we wished to create a new variable which corresponded to the average of these Likert scale variables of awareness, attitude, preference, intention, and loyalty. To compute a new variable, we need to assign a target variable so that the new variable we're going to create to be a transformation of existing variables. We need to give this sum a name. Let's suppose I will call this average score. Then we need to give SPSS a formula to know how to calculate this new variable.

Now, on the right-hand side, here we have the function groups. Now if you really want to become a pro at SPSS, do look at exploring all of the different function groups, but if I just click all here, is well despite all of the different inbuilt functions. Now clearly, many of these are for very advanced SPSS users. We'll just simply look at the simple case of calculating a mean. You know what a mean is. Is adding up observations and dividing by the number of observations. In the description box here, this tells us what the function achieves and in particular the syntax within SPSS for creating such a new variable.

If I double click mean under numerical expression, the function mean has appeared and we just need to provide it with the appropriate arguments. These must be the numeric variables upon which we wish the mean to be calculated. In this instance, I want SPSS to calculate the mean of awareness, attitude, preference, intention, and loyalty. From the variable list, if I now double-click awareness, you see awareness has appeared. As far as the syntax is concerned, each variable needs to be separated by a comma. We have awareness, and if I now double click attitude, this now adds this as our second argument.

Again a comma. We need the correct syntax preference, intention, and finally, loyalty. I'll just need to delete that question mark which was a default argument just to tidy things up. We now have the correct syntax. We're going to use the inbuilt SPSS function of mean, which will indeed calculate the mean of the five variables we have listed here. If I now click okay, it will run this analysis. It's given us an output window or audit trail telling us what it has done. If we now turn to our data view, you can see that this new average score variable has been created.

We may wish to tidy this up a little bit. You can see avg_score appears as our variable name. Turning to our variable view, we've now created this new variable and it's been added to our variable list. Looking at the attributes of this variable, you can see the defaults which have been applied and we may wish to overwrite these slightly. Avg_score is going to continue to be our name. It is a numeric variable. It has a width of eight and you can see the default is of two decimal places. Now we could increase this to three, reduce down to one or zero.

I think, though, two decimal places is a sensible level of a precision for this average score But you can see that the label itself is blank. We said the label attribute was for us to provide a descriptor. In here,

we may wish to type in the average of the Likert variables to be our label. Clearly, if we used average of Likert variables as our name, that's going to be a particularly lengthy name to appear in our data view. Better to have a shortened name to appear in the data view and then as a memory aid for ourselves, we can type a lengthier label.

If we had any missing values, we could assign values but that's not applicable here. If we wish to change the level of measurement, we could do so. The other data operation we are going to consider, again under transform, is the option to re-code into different variables. This could happen for a variety of reasons. If we go back to gender, say, we have the values 1 and 2 where, under values, we can see that 1 corresponded to female and 2 corresponded to male. Now let's imagine we wish to re-code this into a binary coding of let's say o and 1.

It is possible to re-code into the same variables. However, this is highly discouraged because, in this instance, you would be overwriting your original variable and that's a very unwise thing to do. What we're going to do is re-code into different variables. In this instance, we have an input variable which is going to be mapped to an output variable. We can choose from our variable list which the input variable we wish it to be, in this case gender, and click the arrow. This is telling us the original variable of gender is going to be now re-coded into a new output variable for which we will need to give it a name and also a label.

Let me call this gender_binary to indicate that it's going to be called gender_binary as the variable name. Then our descriptive label can be gender re-coded into a binary variable. It's very important to now click the change. Up here, it now tells us what our new destination or output variable is going to be. It's going to look for original values of gender and code them as we are going to define in a moment into a new variable called gender_binary. If we click old and new values, we have a fairly, I think, self-explanatory and intuitive dialog box.

Whereby, we simply state old values and map them to corresponding new values. Remember the original gender variable had 1 corresponding to females and 2 corresponding to males, but I suppose we wish to have a binary coding where the old value of 1, which was female, is still going to be coded to 1. If I click ADD, an original value of 1 is coded to a value of 1. Of course, that means no change to those values, but it's the 2, the original value of two for male, I now wish to re-code to a zero. You can see now the old values of 1 and 2 are going to be re-coded to our new binary values of 1 and 0 respectively.

If I click continue and okay, it's now going to do this re-coding for us. In the output window, this was simply the computer, audit trail of what we've done. If we turn now to our data view, you can see this gender_binary variable has been created. Whenever we create a new variable, it gets concatenated i.e. added as the next available free column in our data view.

Turning to our variable view, we may wish to add a few details here. I like the fact that it's called a gender_binary. It is indeed a numeric variable. Of course, though, having two decimal places is not particularly relevant here. We have the integer coding of zero and one. So we may wish to reduce the number of decimal places from two down to zero. Now we have the much tidier just zero and ones appearing.

Of course, an average score as this arithmetic mean, it made sense to show the variation of those average scores to display them to two decimal places. We have labels so that longer descriptor about what gender_binary means, so this is a memory A to ourselves because suppose we came back to this status set a few days or a week later, we have may have forgotten what it represented. We may now wish to assign labels to those values.

What did the 0 represent and what did the 1 represent? A value of 0 here should have the label of male. We can add that. And whenever there's a 1, that should correspond to the gender of a female, and adds that and do okay. This is a nominal variable, so even though it is numeric, remember it is a categorical variable, and we have of correct level of measurement. Reminder that the original gender was listed here as scale. Of course, we can now override that and more correctly stated as being of a nominal level of measurement.

However, the appearance of the level of measurement in the measure column, this is purely for our own interest and to make sure that we apply the correct statistical techniques on the different levels of measurement of variables. Here we've applied some simple data operations, one of which was to transform variables by creating an average of variables but also to do some re-coding of an existing variable.

[00:11:58] [END OF AUDIO]