

### Using SPSS

SPSS was created in 1968 as the Statistical Package for the Social Sciences. The modern Windows version of SPSS is well suited to storing and analyzing small to medium datasets. SPSS offers an extensive series of drop-down menus that include the ability to describe, graph, transform, and analyze data. These menus make SPSS relatively easy to learn and use. The primary shortcomings of SPSS are that it has fewer advanced statistical functions than packages such as SAS, R, or Stata, and also that merging multiple datasets in SPSS can be difficult.

Although it is primarily menu-driven, SPSS provides the option of displaying the programming commands that are behind the scenes in menu-based operations, so that these commands can be saved and a particular task can be replicated at a later time. In SPSS terminology, these commands are called syntax. Our class will not use syntax, but the Center for Social Science Computation and Research (CSSCR) at the UW offers classes on SPSS syntax.

### SPSS FILE TYPES

.SAV Data file  
.SPO Output file  
.SPS Syntax file  
.POR Portable data file

### STARTING SPSS and INPUTTING DATA

1. Open SPSS by either double-clicking on the SPSS icon on the desktop, or by going to the Start Menu > Program Files > SPSS.
2. When the program launches, a dialog box will appear and ask you whether you want to type in data, open an existing data source, or run a tutorial, among other options. If you choose Cancel, the box will disappear and you will face an empty spreadsheet.
3. Since you will most often be working with a dataset that already exists, we'll now open a data file by clicking on the menu **File > Open > Data**. The file is in C:\temp\mathcamp and is called **school.sav**. If on another occasion you want to input your own data, you can do that by typing directly into the spreadsheet.
4. A note about both SPSS and Stata: life will be much easier if you open the program first and then open the dataset from within the program. If you try to open a data file by double-clicking on it from a window, you will open another copy of the statistical program. I can tell you from experience that having multiple copies of the program open is not a good idea. In addition to being confusing, the programs take up considerable memory and slow down everything else.

### DESCRIBING DATA

The SPSS data editor has two viewing options: Data View and Variable View. Get acquainted with your variables by clicking on the Variable View tab near the lower left corner of the window. The

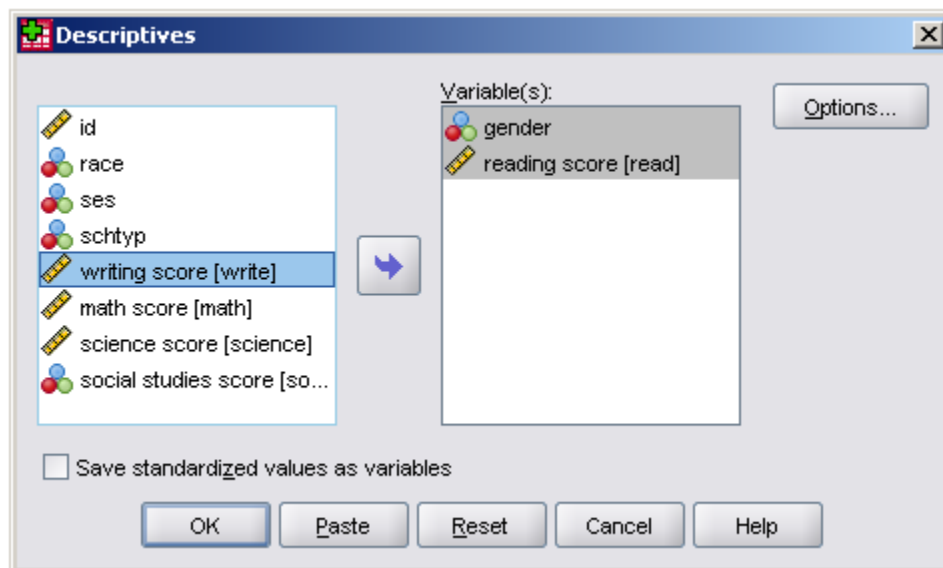
## Introduction to SPSS and Stata

Variable View shows the name of the variable, whether it is a numeric or a string variable (string variables contain letters), the width of the variable's field, the number of decimal places displayed for numeric variables, any label that exists to tell you what the variable measures, the values of a categorical variable including labels for each category, whether any values of that variable are missing (this can happen if someone doesn't answer a question on a survey), the column width that displays the variable, and the alignment of the variable within the column.

The school.sav dataset includes observations for 200 students. Gender is a binary variable, where 0 indicates male and 1 indicates female. For each student, we have an identification number, race, socio-economic status, school type (0=public, 1=private), program type, and test scores in reading, writing, math, science, and social studies.

*What percentage of students are women? What was the average reading score of all students?*

1. In the menus, choose **Analyze > Descriptive Statistics > Descriptives...**
2. A dialog box will appear. Select only the variables **gender** and **reading score**. Hold the control key to make these selections. Click the right arrow to move these to the Variable(s) box. Then click **OK**.



3. Now an output window will appear in order to display the results. The descriptive statistics it provides are the number of observations, the minimum and maximum values attained by the variable, and the mean and standard deviation of the variable. What are the answers to the questions posed above? Hint: the mean of the variable **gender** is also the fraction of students who are women, because the variable is 0 for men.

*What percentage of low socioeconomic status students are in a private school?*

4. In the menus, choose **Data > Select Cases...** Then select **"If condition is satisfied"** and click on **If...** This is the same idea as filters in Excel.

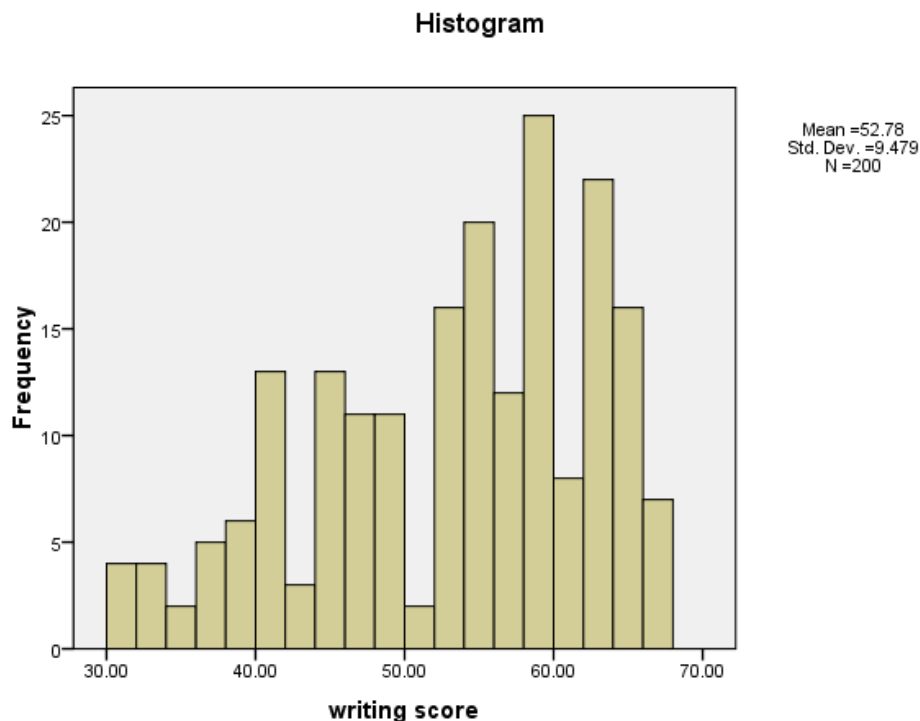
## Introduction to SPSS and Stata

5. In the select cases dialog box, highlight the variable **ses** and click the arrow to move it to the box on the top. We know from the Variable View that the values of the variable ses are 1=low, 2=medium, and 3=high. So type in =1 after ses, so that the statement reads **ses=1**. Click **Continue**.
6. Back in the data editor, choose **Analyze > Descriptive Statistics > Descriptives...**
7. In the dialog box, move **gender** and **reading score** back to the left, and move **schtyp** to the right. Click OK. The results will display in the output window. What's the answer?
8. Before we proceed, let's remove the filter so we can deal with all of the cases again. Go back to the menu **Data > Select Cases...** and click the radio button for **All Cases**.

## GRAPHING DATA

A simple graph that can be very useful in understanding your data is called a histogram. Create one in SPSS as follows.

9. In the data editor menus, choose **Analyze > Descriptive Statistics > Frequencies...**
10. Select the variable **write** and move it to the box on the right by itself. Then click on the **Charts...** button and select **Histograms**. Click **Continue**. Then click **OK**. You should get something like the following.



### Using Stata

Stata is primarily command-driven software. In the most recent version of Stata, menus have been added as an alternative to command-based operations. Stata deals with large datasets relatively well and has the ability to merge datasets together elegantly. It bears more similarities to a programming language than does SPSS, and it allows users to write their own statistical functions and commands. It is especially popular among economists and epidemiologists. Among the shortcomings of Stata are that it can only have one dataset in memory at a time (like SPSS but unlike SAS, which can have several open at once). Stata comes in different “flavors,” which allow different maximum file sizes. The Center for Studies in Demography and Ecology (CSDE) at the UW offers good Stata courses.

#### STATA FILE TYPES

.dta    Data file  
.do    Do file, which contains commands  
.log    Log file, which contains commands and output

#### STARTING STATA and INPUTTING DATA

1. Open Stata by either double-clicking on the Stata icon on the desktop, or by going to the Start Menu > Program Files > Stata.
2. You will see these windows: Results, Command, Variables, and Review. Results are the output and messages about your commands, including error messages. The command window is where you can type in code. Review shows the code that you’ve typed previously. Variables are the variables in the dataset in memory.
3. There are several ways to open a dataset. The first way is with the drop-down menus, by clicking on File > Open and then navigating to your dataset. The second way is to type the command **use C:\temp\mathcamp\school.dta** into the command window. If there are spaces in a filename, you need to put quotes around the path, as in “C:\My Docs and Settings\Example.dta” Stata is case sensitive, so make sure that you type the name exactly.
4. My favorite way to open files and to use Stata in general is through the do-file editor. This is an additional window that you can open by clicking on **Window > Do-file Editor > New Do-file**. Save your (empty) do-file by clicking on **File > Save as**. Call it school.do and put it in the folder C:\temp\mathcamp.

#### ALL ABOUT DO-FILES

Do-files are helpful because they are a place to type commands such that you can use the commands now and again at another time, when you might need to repeat, change or share them. Also, do-files help you to be systematic and to remember the steps that you have tried previously.

Do-files are computer programs. If you have ever taken a computer programming class, many of the same ideas apply. You can type comments into your program. Comments are statements that are not executed as commands. A comment can also be entered directly into the command window, if

## Introduction to SPSS and Stata

desired. In that case, use an asterisk \* at the beginning of the line. Within do-files, you can use an asterisk, or you can instead use /\* to open a comment and \*/ to close a comment.

Here are some helpful commands that often go at the beginning of a do-file.

```
#delimit;           *the end of a line is now denoted with a semicolon;  
                    *allows for multi-line commands;  
                    *don't use semicolons directly in the command window;  
  
clear;             *clears any existing data from memory;  
  
capture log close; *verifies that any log file is closed;  
  
cd C:\Temp\mathcamp; *changes the directory to the one specified;  
  
set memory 100m;   *allocates more computer memory to Stata;  
  
set more off;      *lets the program keep running if output exceeds a page;  
  
log using school.log, replace; *creates a log and saves it in the directory;  
  
use school.dta;    *opens the specified file from the current directory;
```

At the end of a do-file, these commands are useful.

```
save newschool.dta, replace; *saves a new version of the dataset;  
  
log close;         *closes the log file;
```

## DESCRIBING DATA

You can look at your data by clicking on the Data Browser icon, which has a magnifying glass. Stata doesn't allow you to enter commands while the data browser is open. For our data, Stata displays the value labels for the variables **race** and **ses** instead of the numeric values. Those variables are numeric with labels, while the variable **prgtype** is a string variable.

Now, fill in the middle of the do-file with other commands. Type or copy and paste these into your own do-file. When the file is assembled, click the **do current file** button within the do-file editor to execute the commands and see results. If you run current file, it runs, but results are hidden.

*How many variables are in the dataset? What's the median writing score?*

```
describe;         *lists the variables and their types;  
  
codebook;        *provides summary statistics for each variable;
```

*Which student scored highest on reading? How many students were in a vocational program?*

```
sort read;       *sorts data in ascending order according to the variable;
```

## Introduction to SPSS and Stata

```
count if prgtype=="vocati";    *counts number of vocational program students;
```

Note that the double == is a logical operator to verify whether two items are equal to each other. The single = is an assignment operator, such as for creating a new variable.

**Exercise 1:** Suppose the teacher made a grading error and wants to increase everyone's science score by 10 points. We can create a new variable to reflect this fact.

```
summarize science;           *shows summary statistics;
generate newscience = science+10;    *creates a new variable;
summarize newscience;       *gives the new summary statistics;
```

## ASSORTED COMMANDS

Often you can guess what a command name will be, and then do a web search on Stata and that command name. Or, you can type **help** *command name* into the command window. The latter works especially well if you already know a command but need to remember the format and options.

```
display 5+3;                 *calculates 5+3;
mean read;                  *gives the average score (the command might be new);
sort prgtype;              *data must be sorted in order to work with groups;
by prgtype: summarize read; *for each program type, give mean reading score;
tabulate prgtype;          *another way to see the number of people per program;
insheet using filename;    *open a dataset that is not in Stata format;
keep varlist;              *retain only specified variables in your dataset;
drop varlist;              *delete specified variables from your dataset;
```

**Exercise 2:** Graph students' reading score as a function of their math score.

```
graph twoway scatter read math;
```

## Additional Resources

UCLA Statistical Computing (extensive video tutorials and written help files for a number of statistical packages) <http://www.ats.ucla.edu/stat>

Carolina Population Center <http://www.cpc.unc.edu/services/computer/presentations/statatutorial>

Stata NetCourses <http://www.stata.com/netcourse/>