

Getting started with S-PLUS

This handout will give you a brief look at S-PLUS. Later handouts will delve into more detail on S-PLUS's useful and interesting capabilities.

Start-up S-PLUS by double-clicking on the icon on your desktop. The first time you start S-PLUS, you will get a dialog box indicating that an `.Data` or `.Prefs` folder is missing and S-PLUS will create one by default. Click **OK**.

Click on the **Commands Window** button. The **Commands** window will open with the first line saying something like "Working data will be in H:\Program Files\Insightful\SPlus8\Users\JohnDoe". The prompt `>` is where you will type your commands.

For example, try:

```
> 4*9 <ENTER>
[1] 36
> dog <- 1:5
> dog
[1] 1 2 3 4 5
> dog + 10
[1] 11 12 13 14 15
```

The object `dog` is a *vector*.

If you strike the `<ENTER>` key before typing a complete expression, you see the *continuation prompt*, the plus sign (+). For example, suppose you wish to calculate $3 + 2 * (8 - 4)$, but you accidentally strike the `<ENTER>` key too early:

```
> 3+2*(8 <ENTER>
+
```

Finish the expression by typing `-4)` after the `+`

```
+ - 4) <ENTER>
[1] 11
```

To obtain help on any of the commands, you can either access the **Help** dialog box by going to the menu: **Help > S-PLUS Help**, or by typing the name of the command you wish help on:

```
> ?hist
```

S-PLUS for Windows has both a command line and a Graphical User Interface (GUI). In this class, we will focus mainly on the command line.

Reading in data

We will bring in data stored in an S-PLUS transport file format (extension `.sdd`). At the menu, select **File > Open**, navigate to the class folder on Collab: **I:\Depts \MATH \Statistics \Data_LabManual_SPLUS** and open the file **USData.sdd**. This file actually contains several data sets, S-PLUS objects called **data frames**. To see their names:

```
> objects()
```

To see the actual data, type the name of the object at the prompt:

```
> july.precip
```

To see the names of all the variables, type

```
> names(july.precip)
```

For a data frame with a large number of variables, it may be easier to view it in a **Data window** using the GUI. At the menu, select **Data > Select Data** and select **july.precip** from the dropdown list (**Existing Data: Names**), or type

```
> guiOpenView("DataFrame", "july.precip")
```

Now, take a look at the data frame **states98** (if it is not open in a data window, use the above command to bring it up).

The columns are the *variables*. There are two types of variables: *numeric*, for example, **Population** and **Pop.Under.18** and *factor* (also called *categorical*) (for example **Region** and **Death.Penalty**). The rows are called *observations* or *cases*.

Bar charts and histograms

The factor variable **Region** in the **states98** data set assigns each state to one of four *levels*: Midwest, Northeast, South or West. To obtain the number of states in each region, type:

```
> table(states98$Region)
```

Note The \$ is one way to access the variables of a data frame. We will introduce other methods as we progress through the course.

To visualize the distribution of a factor variable, we will create a *bar chart*:

```
> barplot(table(states98$Region))
```

To make this plot more informative, add names:

```
> barplot(table(states98$Region), names=levels(states98$Region))
```

Note The new plot is graphed on the same graph sheet and so the former plot is removed.

To obtain the distribution of a numeric variable, we create a histogram.

```
> hist(states98$Population)
```

The shape of the distribution of this variable is *right-skewed*.

```
> hist(states98$Births.To.Teens)
```

The shape of the distribution of this variable is *bimodal*.

The attach function

We introduce the **attach** function to simplify the syntax.

```
> Population
```

```
> attach(states98)
```

```
> Population
```

The **attach** function places **states98** into the S-PLUS *search path* so we can now access the variables in **states98** by just typing their names.

Numeric Summaries

To find the mean and median of a variable:

```
> mean(Population)
```

```
> median(Population)
```

In addition, we can find the range, or just the maximum and minimum of a variable:

```
> range(Population)
> max(Population)
> min(Population)
```

To find the variance or the standard deviation:

```
> var(Population)
> stdev(Population)
> sqrt(var(Population))
```

To find the *quartiles*,

```
> quantile(Population)
```

Boxplots

Boxplots give a visualization of the 5-number summary of a variable.

```
> boxplot(Population)
```

The white line indicates the median; the bottom and top of the burgundy box indicates the 25th and 75th percentile, respectively. The vertical lines drawn from the box are *whiskers*. The *caps* at the end of the whiskers are drawn by computing:

- The interquartile range (*IQR*) is the difference of the 75th and 25th percentiles (essentially the length of the box).
- The upper cap is drawn at the largest observation that is less than or equal to the 75th percentile + 1.5*IQR.
- The lower cap is drawn at the smallest observation that is greater than or equal to the 25th percentile - 1.5*IQR.

Any observation(s) beyond the cap(s) are drawn as individual lines. These lines indicate *outliers*.

Boxplots are created primarily to compare the distributions of two or more variables. Typically, the variables we wish to compare will come in one of two different formats: a) a numeric variable with a second factor variable used as a grouping variable; b) two or more numeric variables.

Take a look again at **states98** and notice the numeric variable **Births.To.Teens** and the factor variable **Region**.

```
> boxplot(split(Births.To.Teens,Region))
```

What comparative statements can you make about birth rates of teenagers in the different regions of the US?

```
> boxplot(split(Violent.Crime,Death.Penalty))
```

The **split** command is used to break the numeric variables **Births.To.Teens** and **Violent.Crime** into groups determined by the levels of a factor variable (in the above examples, **Region** and **Death.Penalty**, resp.)

```
> detach()
```

The **detach** function removes **states98** from the *search path* so you can no longer access the variables in **states98** by just typing their name. Thus, for example, the last boxplot above would be obtained by using the rather cumbersome syntax of:

```
> boxplot(split(states98$Violent.Crime, states98$Death.Penalty))
```

The data set **jan.temp** gives average January temperatures of the 48 contiguous states and regional averages, from 1895 to 1999. To see the names of the variables,

```
> names(jan.temp)
```

Now, attach the data set to allow easy access to the variables.

```
> attach(jan.temp)
```

To create side-by-side boxplots of two or more numeric variables, just provide the variable names as argument to `boxplot`.

```
> boxplot(Washington, Oregon)
```

By default, S-PLUS does not label the two boxplots so you may wish to provide your own labeling:

```
> boxplot(Washington,Oregon, names=c("Wash.", "Oregon"))
```

```
> detach()
```

Remarks

- Functions in S-PLUS are called by typing their name followed by arguments surrounded by *parentheses*: ex. `hist(Population)`. Typing a function name without parentheses will give the code for the function.

```
> hist
```

- We saw earlier that we can assign names to data (we created a vector called `dog`.) Names can be any length, must start with a letter, and may contain letters or numbers:

```
> fish25<- 10:35
```

```
> fish25
```

Certain names are **reserved** so be careful to not use them: `cat`, `c`, `t`, `T`, `F`,...

To be safe, before making an assignment, type the name:

```
> whale
```

```
[1] Problem: Object "whale" not found Use traceback() to see the call stack
```

```
Safe to use whale!
```

- Unlike programs such as **Excel**, S-PLUS automatically saves newly created or modified data to the **working database**, here **H:\Math245**. In other words, once data is created or imported, they will persist from session to session unless you delete them.


```
> objects()
```

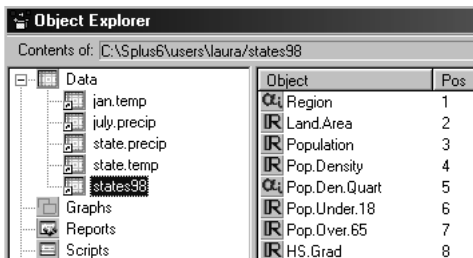
```
> rm(dog)
```

```
> objects()
```

- To save data to an external file/disk, use the **File > Export** command.
- Never save files to the **.Data** folder! This contains S-PLUS' own internal format of data.

The S-PLUS GUI


We will mainly use the command line in this course, but many of the commands are available via the GUI. For example, a handy tool for seeing a listing of data is the **Object Explorer**: click on this button  on the toolbar to open the **Object Explorer**.



In the left pane of the Object Explorer, you will see an icon labeled **Data**. If you click on this icon to select it, then on the right pane, you'll see a list of the data sets brought in. Double-clicking on any one of these icons will bring the data in view in a data window form.

Notice that if you click on the + sign next to the Data icon in the left pane, the data sets will be listed again below the Data icon. Now, if you select any of these data sets in the *left* pane, its variables will be listed in the right pane.

Graphs

First bring up the **2D Plots Palette** (the button  to the left of the word **Linear** on the toolbar). The basic idea for creating graphs is that you select the variables in the data frame (CTRL-click to select multiple variables), then click on the desired plot button on the 2D Plots palette.

For example, bring up the **states98** data set and select the variable **Population** (Click on the column name. The entire column should darken when selected). Now click on the histogram button on the **2D Plots** palette.



A graphsheet with a histogram will appear. The number of bins for the GUI histogram is not ideal so we will change it. Right-click on one of the histogram bins and select **Options** from the context menu. In the dialog box that comes up, change the **Number of Bars:** to 10 and click **OK**.

To create side-by-side boxplots of **Births.To.Teens** grouped by region, first click on the **Region** variable, then CTRL-click on **Births.To.Teens** (the order is important!). Then click on the **Boxplot** button:



To create a scatterplot of **Infant.Mortality** against **Births.To.Teens** click first on **Births.To.Teens** (the x-variable), then CTRL-click on **Infant.Mortality**. The scatterplot button is located in the row 1, column 1 position.

You can also add regional markers to a scatterplot: right-click on one of the data points of the scatterplot and select **Data to Plot**. In the resulting dialog box, select for **z Column** the variable **Region**. Then click on the tab **Vary Symbols**. For both **Vary Color by** and **Vary Style by**, select **z Column** from the drop-down list. Click **OK**.

Note You can make some changes to S-PLUS settings (fonts, startup windows, etc.) using the **Options** menu. For example, to have S-PLUS automatically open the Command Line window and the Object Explorer at start-up, go to **Options > General Settings...** and click on the **Startup** tab.

Project Folders

If you are working on several projects, you may find it useful to keep your data separate. One way to do this is to set up project folders. First, on your Home Drive, create a folder that will hold your project files. For example, suppose your project is your Senior Comps. Create a folder called **SeniorComps** on your home drive.

- Next, right-click on the S-PLUS icon that is on the desktop and choose **Create Shortcut** from the dropdown menu. Move this shortcut to your ITS Home drive (top level).
- Right-click on this shortcut and select **Properties** from the dropdown menu.
- Find the **Target** field; it will look something like
"C:\Program Files\...\cmd\SPLUS.exe" Leave a space after "SPLUS.exe", then type:
S_PROJ=H:\SeniorComps
(The exact path may differ depending on where you placed this folder).
- Close the **Properties** box by clicking **OK**.
- (Optional) Change the name of this shortcut to **SeniorComps**. If you use S-PLUS for other classes or projects, you can create more folders and have shortcuts pointing to each of these project folders. This separation of data is helpful if you have large projects.

Start-up S-PLUS by double-clicking on the icon. The first time you start S-PLUS, you will get a dialog box indicating that an .Data or .Prefs folder is missing and S-PLUS will create one by default. Click **OK**.