Descriptive Statistics

Descriptive statistics consist of methods for organizing and summarizing data. It includes the construction of graphs, charts and tables, as well various descriptive measures such as the mean, standard deviation and percentiles.

Begin by starting a new STATA session. Recall from the previous tutorial that the first thing to do before starting your analysis is to open a log file. If you don't open a log file no results will be saved from your STATA session. After finishing your analysis, the log file can subsequently be edited and printed.

To illustrate the commands we intend on covering in this tutorial, we will use a data set that can be found on the course webpage. In the command window type:

```
use http://www.stat.columbia.edu/~martin/W1111/Data/auto
```

This command will read in a data set auto, which is one of the many data sets available to be downloaded from the STATA web page. This particular data set contains information, from 1978, on a number of variables used to describe 74 different car models. To learn more about the data set we can type the command

```
describe
```

in the Command window. This will provide a non-statistical description of the data. In the Results window, the following output appears:
According to this output, the data set consists of 74 observations on 12 variables (make, price, mpg, rep78, headroom, trunk, weight, length, turn, displacement, gear_ratio and foreign).

Now, suppose we are interested in looking at the complete data set. To do this type

    list

in the Command window. This leads to a fair lengthy amount of output. Suppose we are only really interested in the variable mpg. To only list this variable type:

    list mpg

in the Command window. You should now find a list of the values for the variable mpg for the 74 observations.

The command *summarize* provides a simple numerical summary of all the variables in the data set. It calculates the mean, standard deviation, minimum and maximum values. Type the command

    summarize

in the Command window. This gives the following output:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>make</td>
<td>74</td>
<td>665.257</td>
<td>2949.496</td>
<td>3291</td>
<td>15906</td>
</tr>
<tr>
<td>price</td>
<td>74</td>
<td>21.2973</td>
<td>5.785503</td>
<td>12</td>
<td>41</td>
</tr>
<tr>
<td>mpg</td>
<td>69</td>
<td>2.438797</td>
<td>0.9899222</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>rep78</td>
<td>74</td>
<td>2.993243</td>
<td>8469948</td>
<td>1.5</td>
<td>5</td>
</tr>
<tr>
<td>headroom</td>
<td>74</td>
<td>13.75676</td>
<td>4.277404</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>trunk</td>
<td>74</td>
<td>3019.459</td>
<td>777.1936</td>
<td>1760</td>
<td>4840</td>
</tr>
<tr>
<td>weight</td>
<td>74</td>
<td>189.9334</td>
<td>22.26534</td>
<td>142</td>
<td>233</td>
</tr>
<tr>
<td>length</td>
<td>74</td>
<td>39.64865</td>
<td>4.379354</td>
<td>31</td>
<td>51</td>
</tr>
<tr>
<td>turn</td>
<td>74</td>
<td>197.2973</td>
<td>91.83722</td>
<td>79</td>
<td>425</td>
</tr>
<tr>
<td>displacement</td>
<td>74</td>
<td>3.044855</td>
<td>.4562821</td>
<td>2.19</td>
<td>3.89</td>
</tr>
<tr>
<td>gear_ratio</td>
<td>74</td>
<td>.2972973</td>
<td>.4601885</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
If we are only interested in the summary statistics for the variable mpg and weight, type

    summarize mpg weight

in the command window. This gives the following output:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>24</td>
<td>21.2397</td>
<td>5.80503</td>
<td>12</td>
<td>41</td>
</tr>
<tr>
<td>weight</td>
<td>24</td>
<td>3019.459</td>
<td>777.1936</td>
<td>1260</td>
<td>4840</td>
</tr>
</tbody>
</table>

Including the option detail produces additional statistics including median, and various percentiles.

The command

    summarize, detail

gives this information for all the variables. If we are only interested in this information for the variables mpg and weight, type

    summarize mpg weight, detail

You may have noticed that the variable foreign is a categorical variable. A car is given the value domestic if it is made in the USA and it is given the value foreign if it is made abroad. Suppose we want to study the mean of the variable mpg separately depending on whether or not the car is made in American. Including if in a command allows us to specify the subset of the data in which we want to apply the command. The commands given below give the summary statistics on mpg separately for domestic and foreign made cars. Adding the command foreign==0 specifies that we are only interested in the cars that are domestic (or not foreign), while the command foreign==1 specifies the cars that are foreign.

    summarize mpg if foreign == 0
    summarize mpg if foreign == 1
Correlation

Suppose we are interested in calculating the correlation between two variables in the data set. To find the correlation between mpg and weight type the command:

```
correlate mpg weight
```

This gives the following output:

We see here that the correlation between mpg and weight is -0.8072. We also see that the correlation between mpg and mpg is equal to 1, as is the correlation between weight and weight. Either of these results should be particularly surprising.

Graphics

Graphics can be made either by using the drop down menu Graphics in the STATA GUI or by using the command line.

If we decide to use the command line, we need to type a command such as

```
graphtype varname(s)
```
The type of graph is specified by `graphtype`. There are several types of graphs: histograms, box plots and scatter plots. We will discuss how to make each of these types of plots below.

**Box plots**

We can make a box plot of the variable `mpg` using the command

```
graph box mpg
```

This creates the following graph in a specific graphics window:

![Box plot](image)

Alternatively, we can use the drop down menu `Graphics` and choose `Easy graphs` and thereafter `Box plot`; see Figure 1 for an illustration.

A window will appear that asks you to state the name of the variable(s) you want to plot, see Figure 2. Type in the variable name `mpg` and click OK. An equivalent box plot to the one shown above will be created.
Histograms

To make a histogram of the variable mpg type

```
histogram mpg
```

in the command window. The STATA Graph window will appear with a histogram of the relative frequency distribution of the variable mpg.

You can specify the number of bins for your histogram. Try typing

```
histogram mpg, bin(10)
```

This command allows you to have a histogram with 10 bins. You can experiment with several values and notice how the shape of the histogram is affected by the choice of number of bins.

Alternatively, we can use the drop down menu Graphics and choose Easy graphs and thereafter Histogram.
Scatter plots

We use scatter plots for assessing the relationship between two variables. To make a scatter plot in STATA type

    scatter mpg weight

in the command window. This produces a scatterplot with weight on the x-axis (horizontal) and mpg on the y-axis (vertical).

Alternatively, we can use the drop down menu Graphics and choose Easy graphs and thereafter Scatter plot.

Saving a graph

You can save a graph by clicking on the right button on your mouse and choosing the option Save graph....

Another option is to go to the drop down menu File in upper left hand corner of the STATA window and press Save graph....

It is important to note that your graphs will not automatically be saved to your log file, so you will have to save them separately.
Linear Regression

To create a scatter plot of \textit{mpg} and \textit{weight} type

\begin{verbatim}
scatter mpg weight
\end{verbatim}

Studying the plot, there appears to be a relatively strong negative association between the variables, so we decide to find the least-squares regression line. We can do this by typing the following two commands:

\begin{verbatim}
regress  mpg weight
predict yhat, xb
\end{verbatim}

These commands generate a variable \textit{yhat} that contains the predicted observations corresponding to the data. The first command gives rise to a fair amount of output which we will discuss more in detail later in the course.

For now it is enough to look at the output which is circled. These are the values for $a$, $b$ and $r^2$. Here

\begin{align*}
  a &= 39.44 \\
  b &= -0.006 \\
  r^2 &= 0.6515
\end{align*}

Hence, the least-square regression line is given by \( \hat{mpg} = 39.44 - 0.006 \text{weight} \).

The fraction of the variability in \textit{mpg} that is explained by the least squares line of \textit{mpg} on \textit{weight} is equal to 0.6515.
To plot the regression line together with the data type:

```
scatter mpg weight || line yhat weight
```

This command essentially tells STATA to create a scatter plot of mpg against weight and superimpose the line given by yhat. This command gives the following output:

![Graph showing scatter plot with linear prediction line](image)

We will discuss regression more in depth later in the course.
A guided example

(a) Start STATA. If you are continuing a previous session it is a good idea to clear all the variables. You can do this by writing *clear* in the command window.

(b) Create a log file named *Assignment2.log*. You can do this by writing

```
log using "a:/Assignment2.log"
```

in the command window.

(c) In this example we will use the same data set as described above. In the command window write,

```
use http://www.stat.columbia.edu/~martin/W1111/Data/auto
```

(d) Calculate the mean price of the automobiles in the data set. You can do this by writing

```
summarize price
```

What was the mean price of automobiles in 1978?

(e) Calculate the median price of the automobiles in the data set. You can do this by writing

```
summarize price, detail
```

What was the median price of automobiles in 1978? What does the difference between the mean and median price indicate about the shape of the distribution for the price?

(f) Make a histogram over the price of cars. You can do this by using the following command,

```
histogram price
```

What shape does the histogram take? Is it symmetric? Skewed? Does the shape of the histogram coincide with what you guessed in (e)? Save the histogram.
(g) Make a scatter plot of the variables weight and length. You can use the command,

```
scatter weight length
```

Study the plot. Does there appear to be any association between the variables? Save the scatter plot.

(h) What is the correlation between the weight and length. Use the command,

```
correlate weight length
```

(i) Fit a regression line for predicting the weight of a car using the length. Use the commands

```
regress weight length
predict yhat, xb
```

What is the least –squares regression line? What fraction of the variability in weight is explained by the least square regression on length?

Plot the regression line together with the scatter plot of weight and length. To do this type:

```
scatter weight length || line yhat length
```

Save the plot.

(j) Close the log file using the command

```
log close
```

Edit and print your log file and graphs. Hand them in together with your answers to the questions above.