Descriptive Statistics

In this exercise, we'll be learning how to compute simple descriptive statistics for a dataset. In the examples below, we will use the cholesterol_edit dataset that is posted on Blackboard.

1 SAS

1.1 Read In The Data

Before we can calculate summary statistics, we must first read the dataset into SAS. We do this below using the `INFILE` statement and the text file containing our data.

```
DATA cholest ;
  INFILE 'H:\Bios600\Data\cholesterol_edit.txt' FIRSTOBS=2 ;
  INPUT Chol Gender $ Group ;
RUN ;
```

Suppose that in addition to cholesterol levels, we are also interested in the natural log of cholesterol levels. To analyze the log of cholesterol, we create a new dataset named `cholest2` based on the old dataset `cholest`. In this new dataset, we create a new variable named `logChol` using the log function, applying it to the variable `Chol`.

```
DATA cholest2 ;
  SET cholest ;
  logChol=log(chol) ;
RUN ;
```

1.2 Basic Statistics (Dataset: `Cholest`)

Now we would like to calculate summary statistics for the variable `chol` in dataset `Cholest`. To do so, we will use `PROC MEANS` – for now, we will ignore the variables `gender` and `group`.

We open the `MEANS` procedure by typing `PROC MEANS` and specifying the dataset of interest. We then use the `VAR` statement to declare the variables for which we want to compute statistics. Finally, we close the procedure with the `RUN` statement. Highlight this code in the Editor and run it. Check the Log for errors and finally view the results in the Output window.

```
PROC MEANS DATA=cholest ;
  VAR chol ; *List variables of interest in dataset ;
RUN ;
```

Note that the output includes certain descriptive statistics by default (N , MEAN , STD , MIN , and MAX). To change this, use specific SAS Keywords to ask for what you want and delete the Keywords corresponding to statistics you do not want. Some examples of these SAS Keywords are listed here.

`MEDIAN MODE RANGE Q1 Q3 QRANGE STDERR NMISS SUM KURTOSIS SUM`

As an example, suppose we also want to find the median and the range, but not the sample standard deviation. Then we would submit the following code.

```
PROC MEANS DATA=cholest N MEAN MIN MEDIAN MAX RANGE ;
  VAR chol ;
RUN ;
```

Notice that the Output displays the results in terms of 7 decimal places. If we want to change this, we can use the `MAXDEC=` option in the `PROC MEANS` statement. This option tells SAS the maximum number of decimal places we would like to keep in the Results display. An example of its use is given below, where we request that only one decimal place be kept in the Output.

```
PROC MEANS DATA=cholest N NMISS MEAN STD MIN MEDIAN MAX RANGE MAXDEC=1;
  VAR chol ;
RUN ;
```

1.3 Statistics By 1 Classification Variable

Sometimes we want to calculate descriptive statistics separately for different groups. In the cholesterol dataset for instance, we may want to describe the data separately for males and females. To do so, we must first sort the data according to the classification variable `gender` (using `PROC SORT`). Afterwards, we compute statistics in `PROC MEANS` as before, with the inclusion of a `BY` statement, which is identical to the one used to sort the data.

```
PROC SORT DATA=cholest ;
  BY gender ;
RUN ;

PROC MEANS DATA=cholest N MEAN STD MIN MEDIAN MAX MAXDEC=1;
  VAR chol ;
  BY gender ;
RUN ;
```

Alternatively, we can use the `CLASS` statement within `PROC MEANS`. There are subtle differences between the `CLASS` and `BY` statements, but we will not get into the details in this class. For our purposes, they can be thought of as performing exactly the same task in certain `PROC`'s, with the main difference being the format in which Output is produced. Highlight the code below and run it, and compare the display with what we have previously seen with the `BY` statement.

```
PROC SORT DATA=cholest ;
  BY gender ;
RUN ;

PROC MEANS DATA=cholest N MEAN STD MIN MEDIAN MAX MAXDEC=1;
  VAR chol ;
  CLASS gender ;
RUN ;
```

1.4 Statistics By 2+ Classification Variables

We can compute descriptive statistics according to more than one classification variable. So far, we have analyzed cholesterol as a whole, and separately by gender, but have ignored the variable `group`. Now we would like to separate descriptive statistics on cholesterol by group and gender simultaneously.

The code below achieves this task, classifying cholesterol first by gender and then by group. Highlight this code and run it in SAS to view the breakdown of cholesterol for the four classifications.

```
PROC SORT DATA=cholest ;
  BY gender DESCENDING group;
RUN ;

PROC MEANS DATA=cholest N MEAN STD MIN MEDIAN MAX MAXDEC=1;
  VAR chol ;
  BY gender DESCENDING group;
RUN ;
```

There are a couple important things to notice about the code above.

- **Sort Direction** – SAS automatically sorts a variable in 'ascending' (alphabetical) order. To sort a variable in 'descending' (reverse alphabetical) order, insert the SAS Keyword `DESCENDING` in front of it.
- **Variable Order** – Matches exactly between `BY` statements in `PROC SORT` and `PROC MEANS`.
- **Nesting** – SAS nests statistics according to variable order. Here, SAS operates first on `gender` and then, within each level of `gender`, it operates on `group`.

***Note that the `BY` statement in `PROC MEANS` accepts the `DESCENDING` option, but the `CLASS` statement does not!

In order to get a feel for these rules, try changing your code in a couple of ways. When you change your code, make sure to **pay attention to variable order** in both `PROC SORT` and `PROC MEANS`! Compute statistics for

1. ascending `gender` and then ascending `group`
2. descending `gender` and then descending `group`
3. ascending `group` and then ascending `gender`
4. ascending `group` and then descending `gender`

1.5 Statistics For Multiple Variables (Dataset: `Cholest2`)

I can repeat all the above procedures for my variable named `logChol` in dataset `Cholest2`. Starting from the very first exercise we did, I only have to modify the dataset name and the variable name, as shown below.

```
PROC MEANS DATA=cholest2 ;
  VAR logChol ;
RUN ;
```

Additionally, I could have saved time by computing the descriptive statistics for the variables `chol` and `logchol` in one step, since both variables are available in the dataset `cholest2`. For instance,

```
PROC MEANS DATA=cholest2 ;
  VAR Chol logChol ;
RUN ;
```

The example above uses the simplest exercise we have completed so far. But this property holds for the more complicated exercises as well, as shown below. Note that I changed the dataset name in two places: first in `PROC SORT`, then in `PROC MEANS`. Try redoing the exercises with both variables at the same time!

```
PROC SORT DATA=cholest2 ;
  BY gender DESCENDING group;
RUN ;

PROC MEANS DATA=cholest2 N MEAN STD MIN MEDIAN MAX MAXDEC=1;
  VAR chol logChol;
  BY gender DESCENDING group;
RUN ;
```

2 Excel 2007

2.1 Analysis ToolPak Introduction

2.1.1 Activate the ToolPak

In order to start using Excel, just find the Excel version (.xls) of the cholesterol dataset on your computer and double-click the icon to open it. Before we can start computing summary statistics, we must check whether the Analysis ToolPak is active. To do so, you must

1. Open the **Data** Ribbon by clicking the word **Data** at the top of the screen (fifth item from left).
2. Along the **Data** Ribbon, look to the far right for a button that reads **Data Analysis**, in the Analysis group.
3. If the **Data Analysis** button is there, then the Analysis ToolPak is correctly activated.
4. If the **Data Analysis** button is NOT there, activate it by following the instructions from the Lab1 file Week1_GetSoftware.pdf posted on Blackboard.

2.1.2 Data Analysis Dialog Box

The Excel Analysis ToolPak works through a GUI, with interactive menus that allow you to select the analyses you want, specify your data values and request certain options and statistics. Due to the point-and-click nature of this system, many of the exercise instructions will be given in terms of enumerated lists.

Once the ToolPak is installed, open up the **Data Analysis** dialog box by clicking on the **Data Analysis** button, located at the far right of the **Data** Ribbon. Scroll through the Analysis Tools menu to view the analysis options available in the ToolPak (Correlation, Regression, *t*-tests, etc).

2.1.3 Data Analysis Help Menu

To the right of **Analysis Tools** are three buttons, including **Help**. If you click **Help**, then the Help menu for Data Analysis pops up. If you forget how to use a specific tool, then open up **Help** and find the link for that tool. By clicking the link, Excel gives you a summary of what that tool does, as well as definitions and descriptions of all its input parameters.

Some computers automatically open the Online Help instead of the Data Analysis help menu. To change this, click ‘Connected to Office Online’ (located in bottom right corner of Help). Select ‘Show content only from this computer’ to view the Data Analysis Help menu.

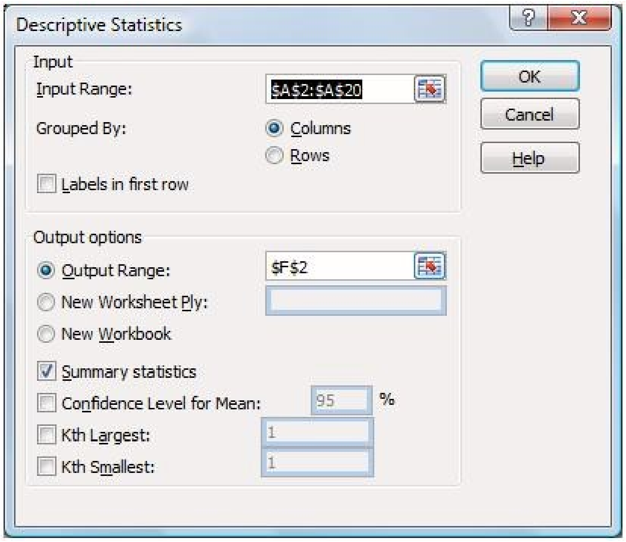
2.2 Basic Statistics

We will now compute summary statistics on the variable `chol`. To do so, first open the **Data Analysis** dialog box. Then select **Descriptive Statistics** and press OK. A new window pops up and requires you to fill in various parameters. Once you have filled these in, press OK and view your results.

The parameters to be filled in tell Excel about your data and what kind of results you would like to see. They are broadly categorized into 'Input', 'Output', and 'Options'. These parameters are listed below, along with a brief description of how to use each one.

- **Input** – Tell Excel about your data
  - **Input Range:** Cells containing the actual data we want to analyze. Simply press the button at right (diagonal red arrow) and then manually highlight your data on your spreadsheet with your mouse. Then press the button at right (downward red arrow) to get back to original window.
  - **Grouped By:** Whether the Input Data is organized in Rows or Columns.
  - **Labels in First Row/Labels in First Column:** Check this box if the first row (or column) of your Input Data, organized by columns (or rows), contains data labels. If it does NOT contain labels, then leave it unchecked.
- **Output** – Tell Excel where to display the results
  - **Output Range:** Cell in which you want the upper-left corner of your output to print. This option is for when you want the output to print in the same worksheet. Select a cell manually in the same way as you selected the Input Range.
  - **New Worksheet Ply:** Print output on new worksheet within this same workbook (file). Name the new worksheet by typing a name in the box at right.
  - **New Workbook:** Print output on new blank workbook (file).
- **Options** – Tell Excel what results to print in the output.
  - **Summary statistics:** Check this box to print an output table with descriptive statistics. By default, it includes (Mean, Standard Error (of the mean), Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, Count).
  - **Confidence Level for Mean:** Check this box to display the margin of error m of a 95% CI for the sample mean (using t-distribution). Change 95% level to whatever you want within the box.
  - K^{th} **Largest/Smallest:** Check this box to print the k^{th} largest/smallest value for each column (or row) of data. Value 1 prints the maximum/minimum.

To compute basic statistics on the variable `chol` and print the Output on the same worksheet, use the following parameter values.



Input
Input Range: $A$2:$A$20
Grouped By: ● Columns ○ Rows
☐ Labels in first row

Output options
● Output Range: $F$2
○ New Worksheet Ply:
○ New Workbook
☑ Summary statistics
☐ Confidence Level for Mean: 95 %
☐ Kth Largest: 1
☐ Kth Smallest: 1

Practice setting these parameters and viewing the results by changing the parameter values (especially Output and Options).

2.3 Statistics By 1 Classification Variable

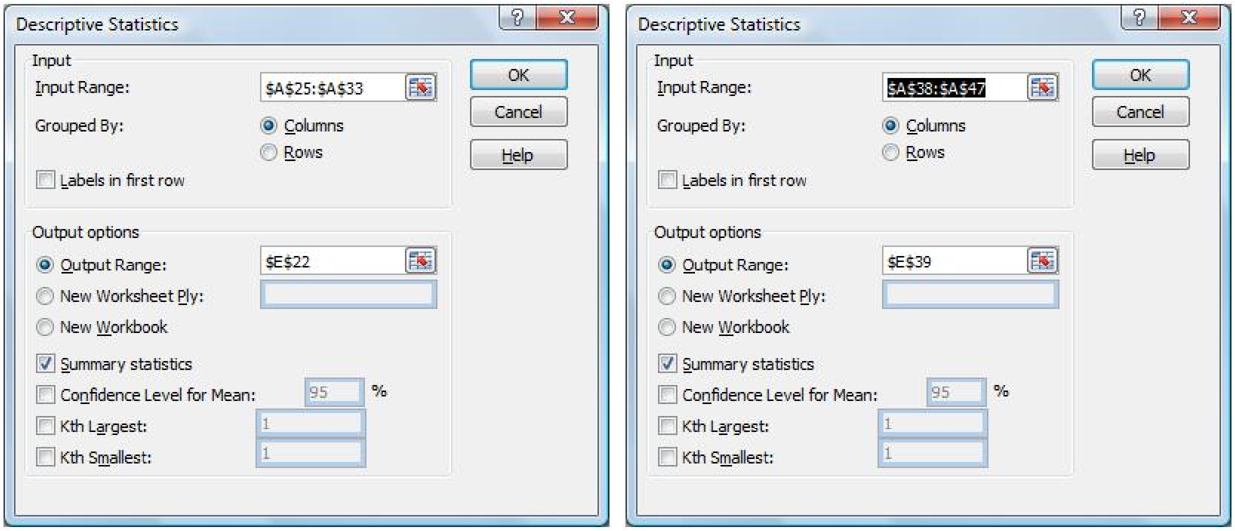
2.3.1 Rearrange Data

As we did in SAS, we would like to compute these summary statistics according to the variable `gender`. In Excel, we can do this manually, by selecting the appropriate data. That is, we use the **Descriptive Statistics** analysis tool once on the data for males, and again on the female data. The first step requires you to rearrange your data in your Excel worksheet.

For this example, we will copy-paste the male data (9 values) to one location and the female data (10 values) to another. Specifically, I will copy-paste the male data to the area from Row 25 to Row 33 (Rows 25:33) in Columns A to C (Columns A:C). Then I will copy-paste the female data to the area in Rows 38:47 and Columns A:C.

2.3.2 Use Descriptive Statistics Tool

Finally, we compute the descriptive statistics for both sets (i.e., genders) by applying the Descriptive Statistics analysis tool to each dataset. Figure 1 on page 8 displays the respective parameter values necessary to produce summary statistics for males and females, assuming the data rearrangement structure described above.



Descriptive Statistics

Input

Input Range: $A$25:$A$33

Grouped By: ● Columns ○ Rows

☐ Labels in first row

Output options

● Output Range: $E$22

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

☐ Confidence Level for Mean: 95 %

☐ Kth Largest: 1

☐ Kth Smallest: 1

Descriptive Statistics

Input

Input Range: $A$38:$A$47

Grouped By: ● Columns ○ Rows

☐ Labels in first row

Output options

● Output Range: $E$39

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

☐ Confidence Level for Mean: 95 %

☐ Kth Largest: 1

☐ Kth Smallest: 1

Figure 1: Parameter values to compute descriptive statistics for Males (left) and Females (right).

Clearly, this method can be tedious as the number of levels in your classification variable increases (e.g., states in the U.S.). If you know of an easier and/or quicker method of computing such descriptive statistics, please let me know so I can add it to this exercise!