

CS 1100
Fall 2009
Excel Lab Assignment 5

To complete this assignment you must submit an electronic copy to Blackboard by the due date. Download a copy of **LabE5.xls** to obtain the initial data for this assignment. The data includes the values of the Female and Male BMI extracted from a solution to Lab 1.

This assignment is about dividing the data into groups called **bins** in order to see how the data values are distributed. In addition, we will see how to visualize the distribution using assorted charts: a histogram, a frequency polygon, and a cumulative frequency polygon. We will use the **VLOOKUP** function to set up the distribution into bins.

The first step in defining the bins is to find the **minimum** and **maximum** of the data values. The VLOOKUP table will start with minimum. Therefore, if minimum is rounded, it must be rounded *down* so that no data value falls below this start point. For symmetrical reasons, if maximum is rounded, it must be rounded *up*.

The data **range** is defined as the difference **maximum** – **minimum**. We plan to divide the range into **N** bins of equal **size** via the formula:

$$\text{size} = \text{range} / \text{N}$$

Frequently, **size** will be rounded and it must be rounded *up* so that the full range of data values will be covered. The bins are then defined by the sequence of endpoints:

$$\begin{aligned} p_0 &= \text{minimum} \\ p_1 &= \text{minimum} + \text{size} \\ p_2 &= \text{minimum} + 2 * \text{size} \\ p_3 &= \text{minimum} + 3 * \text{size} \\ &\dots \end{aligned}$$

Now let us get to the specifics of the assignment.

Underneath the data area in the **LabE5.xls** spreadsheet, you will see an arrangement of cell labels forming 3 tables as illustrated in the snapshot on the next page:

- one table for the minimum, maximum, range, and bin size
- one table for associating bin labels to bin positions
- one table for computing and storing certain counts and frequencies

In the first table, the minimum and maximum should be computed from the *combined* Female and Male BMI data. No rounding is needed since the BMI data is already rounded to one decimal place. As stated above, the range is the difference of the maximum minus the minimum. We will fix the number of data bins at **7**. You must compute the bin size before rounding and then the bin size rounded *up* to one decimal.

A	B	C	D	E	F	G	H	I	J	K	L
	Minimum										
	Maximum										
	Range										
	Number of Bins	7									
	Bin Size										
	Bin Size Rounded										
	Bin Label Table		Female	Male	Female	Male	Female	Male	Cumulative		Cumulative
	Floor	Label	Counts	Counts	Subtotals	Subtotals	Frequency	Frequency	Frequency	Frequency	
	p0										
	p1										
	p2										
	p3										
	p4										
	p5										
	p6										
	p7										

In the second table called the Bin Label Table, you are going to compute the boundary points for the 7 bins as the 8 numbers **p0**, **p1**, **p2**, ..., **p7**. In the **Floor** column, the value next to **p0** should be set equal to the **minimum**. To get each cell value below that, add the **Bin Size Rounded** to the value in the cell above. It may help to use a named Excel range for **Bin Size Rounded**.

Let's be completely clear about what data values conceptually belong to each bin:

- Bin 1 consists of data values v such that $p_0 \leq v < p_1$
- Bin 2 consists of data values v such that $p_1 \leq v < p_2$
- Bin 3 consists of data values v such that $p_2 \leq v < p_3$
- Bin 4 consists of data values v such that $p_3 \leq v < p_4$
- Bin 5 consists of data values v such that $p_4 \leq v < p_5$
- Bin 6 consists of data values v such that $p_5 \leq v < p_6$
- Bin 7 consists of data values v such that $p_6 \leq v < p_7$

As you can see, to define 7 data bins we need 8 endpoints.

For the purposes of counting and computing frequencies, we want to associate a label with each bin and we want to use **VLOOKUP** to make this association. Since **VLOOKUP** uses the *lower endpoint* of a numerical range to make its match, we must actually put the labels in the **Label** column next to **p0**, ..., **p6** rather than next to **p1**, ..., **p7**.

The labels themselves are somewhat arbitrary. We could use names such as Bin 1, ..., Bin 7, or letters such as A, ..., G. However, to have more meaningful labels, we want each bin label to be the *average of its two bin endpoints rounded to two decimals*. Thus, for example, you must compute the average of **p0** and **p1** rounded to two decimals as the label to be placed in the **Label** column in the **p0** row. Notice that there should be no label next to **p7**.

The lookup table range will consist of the 2 columns **Floor** and **Label** and the 7 rows **p0**, ..., **p6**. You should name this lookup table as an Excel named range. You will use this lookup table in conjunction with **VLOOKUP** to assign a label to each of the Female and Male BMI data values.

You will notice that there is an empty column in the data just to the right of the **Female BMI** column. Label this column **Female Bin**. Then use the lookup table in conjunction with **VLOOKUP** to compute the associated bin label for each Female BMI value. When this is done, make a Excel named range for the female bin data.

Do the same task for the Male BMI.

You are now ready to fill in the main table. You should first focus on the **Female Count** and **Male Count** columns. What you want to do is to count how many females and how many males have a given bin label. The Excel function **COUNTIF** is ready-made for this task. This function has two parameters: a *range to count* and a *criterion to test*. For the range, you will use the named range for the female bin data or the male bin data as appropriate.

The criterion may be a cell reference, a constant number or string, or a string snippet that begins with one of the six comparison symbols =, <>, <, <=, >, >= followed by the constant value that will be compared. In the case of a cell reference, there is no way to specify the comparison symbol and Excel assumes that the comparison is equality. Fortunately, this is exactly what we need to count the females and males with a given bin label. Indeed, *the key purpose of defining the bin labels is to obtain an entity that may be tested for equality to accomplish the counts*.

To summarize: In cell **E116** in the first row of the **Female Counts** column, you should have a formula that looks more or less like

=COUNTIF(FemaleBin,C116)

Note that cell **C116** is the cell reference of the bin label in the first row. This formula says to count all items in the **Female Bin** column whose value is exactly equal to the value in the label cell **C116**. You should be able to generalize this formula to obtain the rest of the female and male counts.

The rest of the main table will be calculated based on the data in the first two columns. You will not need to go back to the original BMI data.

In the columns **Female Subtotals** or **Male Subtotals**, you should compute the running subtotals of the corresponding **Female Counts** or **Male Counts** columns. In particular, the final values in the **p6** row should be the total of females and males respectively. You should make each of these totals an Excel named range to use in the remaining computations.

The final four frequency columns have already been formatted to show percentages. To compute the **Female Frequency** or **Male Frequency** columns simply divide the corresponding values in the **Female Counts** or **Male Counts** columns by the female or male totals.

To compute the **Cumulative Female Frequency** or **Cumulative Male Frequency** columns, you might in principle do a running total of the corresponding frequencies. However, this may introduce slight rounding errors during addition. Therefore, it is better to simply divide the

corresponding values in the **Female Subtotals** or **Male Subtotals** columns by the female or male totals to directly compute the cumulative frequencies.

The remainder of this exercise is to construct 3 charts of this data using the Excel chart tools. Each chart will show a sequence of 7 female values and a sequence of 7 male values together with the 7 bin labels as labels. Use the *Insert tab* in the Excel ribbon to get to the chart choices.

The 3 charts are:

- A **histogram** of the **Female Counts** and **Male Counts**. Construct this using a *Column chart* which is also known as a *Bar chart*. Traditionally, a *histogram* is defined as a bar chart that shows the counts of data values that have been distributed into bins.
- A **frequency polygon** of the **Female Frequency** and **Male Frequency**. Construct this using a *Line chart* that shows data values with markers.
- A **cumulative frequency polygon** of the **Cumulative Female Frequency** and **Cumulative Male Frequency**. Construct this using a *Line chart* that shows data values with markers.

In each case, you will follow the following pattern:

- Select the chart type in the *Insert tab*. This will cause a blank chart to be placed on the spreadsheet. By clicking in the interior, you can drag the chart and position the upper left corner. Then drag on the lower right corner to change the size of the chart region.
- Right click in the chart region and choose *Select Data*.
- Click in *Chart Data Range* and select what is there by default. Then drag in the spreadsheet over the cells that will be the correct data range. This will replace what was in the *Chart Data Range*.
- Now continue by clicking *Edit* on the *Horizontal (Category) Axis Labels*. In all cases, you will replace the default labels 1, 2, ... by selecting the column with the 7 bin labels.
- Next select each of the *Legend Entries (Series)*. You should replace the default series labels with **Female** and **Male**.
- Now click *OK* to dismiss the dialog box.

While you are still working on a chart, you may add a title. Notice that you will see an area in the ribbon that says *Chart Layout*. You may click on an icon that shows a title at the top. Once you do this, a title will appear in the chart. You may select the default title and change it to what is appropriate for the particular chart.

Warning: In tests, it appears that the order of setting ranges when making a chart is important. Therefore, we recommend that you follow the order described in the bulleted list above. Furthermore, you may use a named range for the *Chart Data Range* if you wish but not for the *Horizontal (Category) Axis Labels*. This appears to be a bug in Excel.

On the next page, we will show samples of the 3 charts.

