1 Chapter 2: Descriptive Statistics: Tabular and Graphical Methods

1.1 Summarizing Qualitative Data

1.1.1 Frequency Distribution

Definition 1 ***Frequency Distribution:*** *A table used to summarize quantitative or qualitative data.*

We set up a frequency distribution by defining bins or classes which are simply little collections of possible data values. The bins should stretch above the top value and below the bottom value. The bins should be small enough so that the data is spread across several bin and big enough so that the summarized data has some meaning. We typical want to follow two rules when setting up bins:

1. (a) Every data point should be in one and only one bin. We do not want any overlap across the bins or holes between the bins.

   (b) The bins should be the same width. This makes summarizing the data easier.

   Example: Consider the following 10 test score. C,B,B,A,D,F,C,B,C,A. Let's construct a frequency distribution. Let's define 5 bins.

Bin	Number in the Bin
A	2
B	3
C	3
D	1
F	1

1.1.2 Using Excel's COUNTIF Function to Construct a Frequency Distribution

Here are the Steps to create a frequency distribution in Excel

1. Enter or import the data to summarize.
2. Enter the labels or the categories to be summarized.
3. Enter the countif function as follows next to the label "X" in cell C2 and data in Cells A2-A11:

   $$= \text{COUNTIF}(\$A\$2{:}\$A\$11,C2)$$

   Press enter, and the count should show up where you entered the formula. Let's do this with the test data above.

1.1.3 Relative and Percent Frequency Distributions

A frequency distribution displays the number of times items show up is particular bins. This is not the only type of frequency distribution. There are also relative and percent frequency distributions.

1. **Definition 2** ***Relative Frequency Distribution:*** *This is a frequency distribution which states the proportion in each bin.*

   Definition 3 ***Percent Frequency Distribution****: This is a frequency distribution which states the percentage in each bin.*

   The following is the relative and percent frequency distributions for the test scores.

Bin	Relative Frequency Distribution	Percent Frequency Distribution
A	0.20	20%
B	0.30	30%
C	0.30	30%
D	0.10	10%
F	0.10	10%

   This says that 30% of the class got a B on the test.

1.1.4 Using Excel to Construct Relative and Percent Frequency Distributions

Here are the steps to creating relative and percent frequency distributions

1. Construct the frequency distribution. (see previous section)

2. Summarize over the frequencies by using the SUM function. If the frequency are in cells D2 through D6. Then the formula in D7 would look like

   =SUM(D2:D6)

3. In the next column simply divide the frequency by the total count to get a relative measure. For example in cell E2 we would have

   =D2/D7

   In the cell E2, you should see the relative weight for the appropriate bin. Simply fill down the column to finish the relative frequency distribution.

4. To get the percent frequency distribution you simply need to multiply each element of the relative distribution by 100. So, in cell F2 we would find

   =E2*100

   In the cell F2, you should see the percent weight for the appropriate bin. Simply fill down the column to finish the percent frequency distribution.

5. As a final check, you should summarize over the elements of the new frequency distributions. The sum of the relative frequency distribution should be 1 and the sum of the percent frequency distribution should be 100.

1.1.5 Bar Graphs and Pie Charts

Definition 4 ***Bar Graph and Pie Chart:*** *Graphical summaries for depicting data that has been summarized in a frequency, relative frequency, or percent frequency distribution.*

For a bar graph, the horizontal axis is made up of the bins in the frequency distribution. The vertical axis is the count, relative weight, or percent in each bin. Each bin will be represents by a bar of the appropriate height.

The pie chart, is a circular graph where each piece of the pie corresponds to a bin. The importance of the bin (amount of data) is represented by the size of the piece corresponding to that bin.

1.1.6 Using Excel to Construct Bar Graphs and Pie Charts

We will be using the Chart Wizard to construct bar graphs and pie charts. Here are the steps to construct a graph in Excel

1. Enter the data
2. Highlight the data you wish to graph. The first column you highlight will be the horizontal axis for the bar graph or the bins for the pie chart. The second column is the actual data.
3. Click the Chart Wizard button or go to the INSERT menu and Select the CHART option. Either way will send you into the chart wizard
4. (Chart Wizard Step 1 of 4) Choose COLUMN in the Chart Type list and CLUSTERED COLUMN from the sub-type list. Click Next
5. (Chart Wizard Step 2 of 4) Just Click Next
6. (Chart Wizard Step 3 of 4) Select the Title tab. Type in labels for the Chart title and axes where displays. The x axis should be a summary label for the bins. The y axis should be a label for the type of frequency distribution you are graphing. Click Next
7. (Chart Wizard Step 4 of 4) Just click the finish button.

1.2 Summarizing Quantitative Data

1.2.1 Frequency Distribution

We can also use frequency distributions to summarize quantitative data. The main difference is that we have to manually setup the bins. There are three important things to do when setting up the bins for quantitative data.

1. Determine the number of bins we want to use and make sure that they do not overlap. Generally you want somewhere between 5 and 20 bins. This allow for variation in the data but also allows the data to be summarized.

2. Determine the width of the bins. The bins, in general, should have the same width. The easiest way to determine the width of the bins is to apply the following rule.

$$\text{Bin Width}=\frac{\text{Largest Data Value - Smallest Data Value}}{\text{Number of Bins}}$$

3. Determine the class limits. These are the upper and lower bounds for the data.

Let's Consider the following data.

Consider the following 10 test scores. 78,82,89,93,60,54,76,81,70,91. Let's construct a frequency distribution. Let's define 5 bins.

1. Bin A covers 90-99

   Bin B covers 80-89

   Bin C covers 70-79

   Bin D covers 60-69

   Bin F covers 50-59

   Using these bin we can construct a frequency distribution for the test scores.

Bin	Number in the Bin
A	2
B	3
C	3
D	1
F	1

1.2.2 Using Excel's FREQUENCY Function to Construct a Frequency Distribution

We can use Excel to construct quantitative frequency distributions. Suppose the test scores are in cell A2-A11 and I want to count the number of scores between 90 and 99. To do this, you simply apply the Frequency function. Here are the steps

1. Enter the data

2. Select the cells to be used for the frequency distribution. (in this case you should highlight five cells)

3. For this example type the following formula

=FREQUENCY(A2:A11,{59,69,79,89,99})

where the argument before the comma is the data and the list after the comma is all the upper limits of each bin. DO NOT HIT ENTER. This is an array function. You need to hit CTRL + SHIFT +ENTER to fill the formula down the selected cells.

The frequency distribution should show up in the highlighted cells.

1.2.3 Relative and Percent Frequency Distribution

The relative and percent frequency distribution are constructed in the same manner as with qualitative data. So for the following data you should get the same relative and percent frequency distributions we had before.

1.2.4 Histogram

Another common graphical presentation of quantitative data is a histogram. This graphs is very similar to the bar graph we constructed for qualitative data. The only difference is the horizontal axis is a range of possible data values and not the name of the bins. Let's use Excel to construct a histogram of our test data.

1.2.5 Using the Chart Wizard to Construct a Histogram

Were are going to use the chart wizard to construct a histogram. Here are the steps

1. Enter the data.
2. Construct your frequency, relative frequency, or percent frequency distribution,
3. Click the chart wizard button.
4. Construct a bar graph like we did for qualitative data.

   We now want to remove the gaps between the rectangles.
5. Right click on any rectangle in the graph and select format data series.
6. Select the Options tabs and enter 0 in the GAP WIDTH BOX
7. Click OK.

1.2.6 Cumulative Distributions

There is one more type of frequency distribution we need to talk about: the cumulative frequency distribution.

(a) **Definition 5** ***Cumulative Frequency Distribution:*** *This frequency distribution states the proportion or percentage of the data at or below a specific bin.*

The following is the cumulative frequency distribution for the test scores.

Bin	Percentage at or above this Bin
A	20%
B	50%
C	80%
D	90%
F	100%

This says that 50% of the classes got at least a B on the test.

We could also construct cumulative relative frequency and cumulative percent frequency distributions to track the proportion or percent of the data at of below certain values.

1.2.7 Graph of Cumulative Distributions: Ogive

A graph of a cumulative distribution is called a ogive. This graph has the cumulative frequency distribution on the vertical axis and the data value on the horizontal axis.

1.3 Crosstabulations and Scatter Diagrams

Crosstabulations and scatter diagrams are used when trying to uncover relationships between multiple variables.

1.3.1 Crosstabulation

A crosstabulation is a table summarizing data for two variables. These are particularly useful when one of the variables is qualitative. For this section we are going to use a data sample which gathers monthly salary and education data. The education data is qualitative in that a person can have three possible values

(a) NOHS: did not graduate from High School

(b) HS: attain a High School degree and did not have any college.

(c) COLL: has some college experience

We want to study the relationship between education and salary. Using a cross tabulation we can begin to see the relationship between salary and education. The following is a crosstabulation for our data sample.

Count of Person	SALARY						
Schooling	3000-3999	4000-4999	5000-5999	6000-6999	7000-7999	8000-8999	Grand Total
NOHS	0	6	6	1	0	0	13
HS	1	15	23	10	0	0	49
COLL	0	2	14	14	0	1	31
GRAND TOTAL	1	23	43	25	0	1	93

The last row of the crosstabulation contains the frequency distribution for the salary data. The last column of the data contains the frequency distribution for the schooling data. The interior of the table contains the conditional frequencies of the data for given salary and schooling pairs. From this, we can quickly see that it appears that people with more education earn higher salaries.

1.3.2 Using Excel's Pivot Table Report to Construct a Crosstabulation

Let's construct this crosstabulation (called a Pivot Table) in Excel.

(a) Enter the data. This will only work if you have at least 3 columns of data.

(b) Before we can create the Pivot Table, we need to establish NOHS, HS, COLL as a custom list.

Under the Tool menu select Options.

(c) Click the custom lists tab then type NOHS press enter, HS press enter, COLL press enter, then click add.

(d) The new list will be add, Click OK.

(e) To create a Pivot Table highlight the data then go to the Data menu and select PivotTable and PivotChart Report.

(f) (Step 1 of 3) Be sure that the options for Microsoft Excel List or database and Pivottable are selected

(g) Click next.

(h) (Step 2 of 3) If you highlighted the data correctly, you simply need to click Next

(i) (Set 3 of 3) Click the Layout button.

You are now ready to construct the layout of your PivotTable.

You are going to drag the names of the column into the actual table.

(j) First, take the Salary label (the quantitative data) and click and drag it into the space labeled "drop column fields here".

(k) Second, take the School label (the qualitative data) and click and drag it into the space labelled ”drop row fields here”.

(l) Finally, take the person label and drop into the space for data. Click OK. Then Click Finish.

The Pivot Table should be in front of you now.

We now what to get the data in the form of a frequency distribution.

(m) Double click on the cell labelled ”Sum of person” and select the count option. Click OK.

Notice every individual salary shows up as an individual column in the Pivot Table. It would be useful to group the salary data into fewer columns. Here is how this is done.

(n) Right click on the ”Salary” label and select the Group and Outline option and choose Group.

(o) In the starting at row type in 3000 and make sure the box is not checked.

(p) In the ending at row type in 9000 and make sure the box is not checked.

(q) Make sure the by row should say 1000, and click OK.

You should now see a much smaller Pivot Table that is more summarized.

1.3.3 Scatter Diagram and Trendline

Definition 6 ***Scatter Diagram:*** *A graphical representation of the relationship between two variables.*

Definition 7 ***Trendline:*** *A line that provides an approximation of the relationship between two variables.*

For any scatter diagram, the horizontal axis represents one variable and the vertical axis represents the other variable. The points in a scatter diagram represent actual data.

The trendline allows us the see the obvious relationship house values increase with house size.

1.3.4 Using the Chart Wizard to Construct a Scatter Diagram and a Trendline

Now we are going to learn how to make scatter diagrams and trendlines in Excel. Here are the steps.

(a) Enter the data.

(b) Highlight the data you want graph then activate the Chart Wizard.

(c) (Step 1 of 4)Choose XY (scatter) as the chart type and scatter as the subtype. Click Next.

(d) (Step 2 of 4) Click Next.

(e) (Step 3 of 4) Under the titles tab enter the title of the scatter diagram and label the x and y axis appropriately.

(f) Under the Legend tab, uncheck the box labelled Show Legend. Click Next

(g) (Step 4 of 4) Click Finish.
You should now see the scatter diagram. The data point will be represented by points on the graph.
Now we want to add the trendline to the graph.

(h) Right click on any data point in the graph and select add trendline.

(i) In the popup window make sure the linear type is highlighted and click OK.
There should now be a trendline on your scatter diagram.