

**Lab 1:  
Organizing and Describing Data**

**Objectives:**

1. Introduce Excel as a tool for organizing and analyzing data.
2. Practice manipulating data.
3. Review “Organizing Data” from basic stats – use Excel to illustrate *frequency* and *relative frequency distributions*.
4. Use descriptive measures: *mean*, *standard deviation*, *quartiles* and *interquartile range* to describe distributions.

**Key Terms:**

1. Grouped Data Table.
2. Frequencies and Relative Frequencies.
3. Histograms (Bar Graphs).
4. Cross sectional data
5. Mean and Standard Deviation
6. Quartiles, Interquartile Range, 5-Number Summary.

**Data: 2004 Accord Prices.xls**

Data were collected for **all** used Honda Accords within 100 miles of Amherst listed on January 17, 2006 from *AutoTrader.com*. We can consider these *population* data, a rather narrowly defined population. These are actual prices for 100 used 2004 Honda Accords. The spreadsheet *2004 Accord Prices.xls* is posted on the course website under **Lab Exercises** (all lab data will be posted on the website). They also represent an example of a *cross-sectional data set*, a data set that contains values collected for different cars at a specific point in time for several variables that describe those cars (price, mileage, transmission, etc.).

**Exercises:**

◆ **Save Your Excel File.**

1. Everyone should save the same basic Excel file saved – we’ll be using that file for several exercises. First, create your own folder within the **ResEc 312** folder. Next, open the Excel file for today’s lab. Before doing anything else, rename the file by saving it in your folder.
2. Within the “**Submit**” folder, save a copy of your file in the **Lab 1** folder as: *your name.xls*. This will give us a record of your progress on lab work. You should get in the habit of clicking the **Save** icon often while you work. We’ll review your progress as we move along.
3. Each lab session, be sure that you have the same current copy saved in our Submit folder and your personal folder. Remember, none of these folders are protected. They are for your convenience. You really must keep a back-up of your work on a thumb-drive, on your *UDrive*, etc.

◆ **Grouped Data Table for Accord prices.**

1. The data are in no particular order. You should find 100 observations for Honda Accord prices and other variables that characterize each car. Today we’ll work with the price data. **Highlight the variables Year, Model and Price. Right click and copy.**
2. On the **Home** ribbon, find **Insert** and the option for **Insert Sheet**. Then right-click the tab at the bottom of the new sheet that says *Sheet 1*, choose **Rename** and type in the new name *Accord Prices*. Now, right click in a cell and **Paste** the data you copied. (Or, click a cell and type “Ctrl v.”)
3. **Sort the data.** To sort the data, first click on one data cell and only one (you could also highlight all the data). Then select the **Data** tab and **Sort**. Make sure the box in the upper right that says “*My data has headers*” is checked. Next, select the variable by which you wish to sort the data - *Price* and choose “*smallest to largest*.” **Warning!** When sorting data, you must always keep all information about each observation together. If you were to highlight just one column, say price, and ask Excel to sort, it will give you a warning that there are other data around that column. ***Don’t ignore those warnings!!!!!!!!!!!!!!!!!!!!***

4. You can now easily determine the minimum price and the maximum price. These will be useful in creating a **grouped data table**. You should find that the **range** is \$11,000.
5. Create a set of **classes** – categories that contain prices in a specific range. Use the classes available in the worksheet: **Grouped Data Table & Histogram**. These classes have a consistent **class width** and there are a relatively small number of classes. As a rule of thumb, you should have between 5 and 15 classes.
6. Determine the **frequencies** – count the number of observations (cars) in each class and enter the results in the column labeled “Frequencies.” I usually left click and highlight all prices that are in a class. The “Name Box” in the upper left corner of the tool bar will tell you how many rows you’ve highlighted. (You could use the **Countif** function in Excel to do this, but each application requires a bit of typing.) At the bottom of the frequency column in your grouped data table, use the  $\Sigma$  button (on the **Home** ribbon) or the **Sum** function to determine the total number of cars (it should be 100).
7. Create a third column for relative frequencies (proportions of the total). The relative frequency is the frequency divided by the total number of observations (100). To do this in Excel, type:  

$$[= \text{cell address for the frequency} / \text{cell address for the total} ]$$
 Each cell address is comprised of a letter (for the column) and a number (for the row). When entering the cell address for the total in your formula, place a \$ before the number. Or, while in the formula bar, click on the cell and hit the F4 key. This will fix the row address for the denominator. Once the equation is complete, just hit **Enter** and the relative frequency will appear. Right click on that cell and choose **Format** then **Number** and choose 3 decimal places.
8. Next, place the cursor on this cell again and click. Move the cursor to the lower right corner of the highlighted cell. When the cursor changes from a fat plus sign to a thin plus sign, **click and hold down the left mouse button**, drag down over the remainder of the relative frequency cells and release. Bingo! Excel copies the equation to each cell, changes the numerator address, but not the denominator address (because you fixed it with a \$ sign) and calculates all the relative frequencies. If you sum the column of relative frequencies, you should get 1.000.

◆ **Relative Frequency Histogram**

Create a relative frequency histogram using the chart tools. You’ll find these tools under the **Insert** tab. Select the **Column chart**. Once you do that, a bunch of additional chart tools will pop up. The third set of tools over allows you to choose a **Chart Layout** – Layout 1 looks pretty good. You can always change these things later. After selecting that Layout, go back to **Select Data**. You’ll need to choose the **Legend Series** (the vertical axis – the Accord Price relative frequencies) and the **Horizontal Axis Labels** (your price classes or category labels from the grouped data table). Once you’ve entered the data, you’ll want to make sure the chart has excellent labels and a great descriptive title. I’ve got examples pasted below of charts.

◆ **Save early - save often.**

Click the **Save** icon. (You should also save your spreadsheet on your U-drive or a flash drive.) If you wish to print, put your name at the top of your grouped data table and then **print only your grouped data table, your frequency histogram and your relative frequency histogram**. Highlight those three items. Click the “**Office Button**,” choose **Print**, and **Print** again. **Stop!** Near the bottom of that window where it says “Print what,” choose **Selection**. Don’t print anything else – conserve paper and don’t tie up the printer. Your printout should look like the one on the final page of this document.

◆ **Descriptive Measures.**

Complete the worksheet titled **Descriptive Measures**. I’ve copied the accord price data into the worksheet. Use the formulas from basic statistics to calculate the **mean** and the **standard deviation** for these Accord Price data. Place the results in the shaded cells on that sheet.

The mean is easy:  $\mu = \frac{\sum X}{N}$ . First, sum the price data and divide by the number of cars,  $N = 100$ .

The standard deviation requires that you create a couple of extra columns, just as you did during your Intro to Stats course. Use the formula that was referred to as the “defining formula:”

$$\sigma_x = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

To complete this you need: (1) the mean you calculated; (2) a column of “deviations” – calculate  $(X_i - \mu)$  for each observation (if you sum this column, what should you get?); (3) a new column of squared deviations  $(X_i - \mu)^2$ ; and (4) the sum of the final column (the sum of the squared deviations). Then divide the column sum by N and take the square root. In excel, the function for the square root is “=SQRT(cell address).”

Find the **median** price of Accords. Recall that the median is the center of an ordered list. Then find the first and third quartiles. The difference is the **interquartile range**. Formulas for finding the **locations** are:

$$Q_1 @ \frac{(n+1)}{4}; \quad Q_2 @ \frac{(n+1)}{2}; \quad \text{and} \quad Q_3 @ \frac{3(n+1)}{4}.$$

Determine the quartiles by interpolating:  $Q_j = x_i + .dd \cdot (x_{i+1} - x_i)$ , where  $i$  is the integer portion of the quartile location and  $dd$  is the decimal portion (eg., if  $Q_3 @ 47.75$ , then  $i = 47$  and  $.dd = 0.75$ ).

Your **Descriptive Measures** worksheet should look like the screen capture below. While numeric values can be seen in the screen capture below, **all your cells must contain formulas**.

The screenshot shows an Excel spreadsheet with the following data and statistics:

Year	Model	Price	Price Dev.	Price Dev.-sq				
2004	Honda Accord EX	13995	-\$7,033.96	49,476,593.28				
2004	Honda Accord DX	14495	-\$6,533.96	42,692,633.28				
2004	Honda Accord LX	16888	-\$4,140.96	17,147,549.72				
2004	Honda Accord DX	16900	-\$4,128.96	17,048,310.68				
2004	Honda Accord LX	17490	-\$3,538.96	12,524,237.88				
2004	Honda Accord EX	17495	-\$3,533.96	12,488,873.28				
2004	Honda Accord LX	17500	-\$3,528.96	12,453,558.68				
2004	Honda Accord LX	17575	-\$3,453.96	11,929,839.68				
2004	Honda Accord LX	17750	-\$3,278.96	10,751,578.68				
2004	Honda Accord LX	17988	-\$3,040.96	9,247,437.72				
2004	Honda Accord EX	17993	-\$3,035.96	9,217,053.12				
2004	Honda Accord LX	17995	-\$3,033.96	9,204,913.28				
2004	Honda Accord DX	17995	-\$3,033.96	9,204,913.28				
2004	Honda Accord EX	18499	-\$2,529.96	6,400,697.60				
2004	Honda Accord LX	18590	-\$2,438.96	5,948,525.88				
2004	Honda Accord EX	18688	-\$2,340.96	5,480,093.72				
2004	Honda Accord EX	18975	-\$2,053.96	4,218,751.68				
2004	Honda Accord LX	18990	-\$2,038.96	4,157,357.88				
2004	Honda Accord EX	18995	-\$2,033.96	4,136,993.28				
2004	Honda Accord EX	18995	-\$2,033.96	4,136,993.28				
2004	Honda Accord EX	18995	-\$2,033.96	4,136,993.28				
2004	Honda Accord EX	18995	-\$2,033.96	4,136,993.28				
2004	Honda Accord EX	18999	-\$2,029.96	4,120,737.60				
2004	Honda Accord EX	19200	-\$1,828.96	3,345,094.68				
2004	Honda Accord EX	19480	-\$1,548.96	2,399,277.08				
2004	Honda Accord EX	19485	-\$1,543.96	2,383,812.48				
2004	Honda Accord EX	19490	-\$1,538.96	2,368,397.88				
2004	Honda Accord EX	19795	-\$1,233.96	1,522,657.28				
2004	Honda Accord LX	19850	-\$1,178.96	1,389,946.68				
2004	Honda Accord EX	19888	-\$1,140.96	1,301,789.72				
2004	Honda Accord EX	19987	-\$1,041.96	1,085,680.64				
2004	Honda Accord EX	19990	-\$1,038.96	1,079,437.88				

  

Mean:	\$21,028.96
Standard Deviation:	\$ 2,296.69
<b>Five Number Summary:</b>	
Minimum:	\$13,995.00
Q1:	\$19,481.25
Median:	\$20,995.00
Q3:	\$22,995.00
Maximum:	\$24,995.00
<b>Quartile Locations:</b>	
Q1:	25.25
Median:	50.5
Q3:	75.75

Dan Lass:

**Distribution of Used 2004 Honda Accord Prices**

<b>Accord Prices</b>	<b>Frequency</b>	<b>Relative Frequency</b>
Less than \$15,000	You	You
\$15,000 < \$16,000	need	need
\$16,000 < \$17,000	counts	relative
\$17,000 < \$18,000	for	frequencies
\$18,000 < \$19,000	each	for
\$19,000 < \$20,000	class	each
\$20,000 < \$21,000	in	class
\$21,000 < \$22,000	these	in
\$22,000 < \$23,000	cells.	these
\$23,000 < \$24,000		cells.
\$24,000 and over		
	100	1.000

