Geography 364—Assignment 1

Histograms and simple data manipulation using *Excel* and PASW (*SPSS)*

Due Date: September 13th 2010, Monday, 7pm
Total points available: 100 points

This assignment introduces you to some of the important simple ways of exploring data statistically. You will be using *Microsoft Excel* and PASW (formerly: *SPSS*). An important aim of the assignment is to provide you with an introduction to some of the capabilities of these programs. Although you may not have access to specialized statistical software in your career, it is very likely that you will have access to *Excel*, and/or *PASW*, so these are useful skills to acquire. It will also be assumed in later assignments and homework that you are familiar with the techniques you learned in this assignment.

Most of the questions follow up on something you are asked to do, asking you to comment on it, or explain something. Remember to ask for help if you are lost—you won't score any points for not answering a question. The points available for each question are clearly indicated. For some questions you are asked to provide printouts of results, and you should attach these to your answer. You will find a DropBox in ANGEL where you should submit your answers to as well (to make sure they do not get lost). You do not need to scan print outs but you are encourage to provide a completely digital solution. Please also hand in a print out.

Please type your answers. Handwritten answers will not be graded. You may hand-draw charts if you wish, but you must use a ruler or other straight-edge where necessary and will lose points for 'scrappy' diagrams. Finally, you are not obliged to use Excel ***or PASW to generate your answers. If you are already familiar with another program, then feel free to use it to answer questions where possible.***

1 Getting started

In this assignment the data is from the US Department of Justice, Bureau of Justice Statistics, recording violent and property crimes for all fifty states (and the District of Columbia) in 2008. You will find the file (named **`crime2008.xls`**) on the course website on the 'Labs' page.

Download a copy of the spreadsheet, and open it in *Excel*. ***Make sure to save your own copy of the spreadsheet***.

The spreadsheet contains a number of *worksheets*, called **`USDoJ-Data`**, **`POP-Histogram`**, and so on, as named in the tabs at the bottom left of the screen. Switch among the worksheets by clicking on the tabs. As you work on this assignment, you will fill in results in each of these worksheets in turn. The **`USDoJ-Data`** sheet is the one containing the data.

NOTE: The **`USDoJ-Data`** worksheet has been locked so that you don't lose or alter the data. This means that you <u>can't</u> do calculations in this worksheet.

Navigating around the spreadsheet

First, remind yourself of a couple of *Excel* features that help in making selections and navigating the spreadsheet (use the extra copy of the data in the **`Practice`** worksheet to practice on.):

- Use the **`SHIFT`** key to *extend* a selection. Click on a cell to select it. Now click on another cell or move to another cell using the arrow keys, ***with*** **`SHIFT`** ***held down*** so that the selection is extended from the original cell to a rectangular area whose opposite corner is your other selection.

- Try using the **CTRL** key in a similar way. Combined with the arrow keys this jumps to the *last contiguous filled cell* in the direction of the arrow key—to the top or bottom of the column you are in, or to one end of the row you are in.

- Now combine **CTRL** and **SHIFT** with arrow keys and mouse-clicks. Once you get used to it, these make it a lot easier to move around blocks of data and select multiple columns and rows of data. For example **CTRL-SHIFT-↓** can select all the values in a column, if the cursor is in the top cell of the column.

Other keyboard shortcuts are **CTRL-c** to ***copy*** the current selection, and **CTRL-v** to ***paste*** the last data you copied where you want it. You should get used to these operations before you proceed.

2 Interpreting and making histograms

I've already made a histogram of the state populations in the **POP-Histogram** worksheet. Look at this carefully and answer the following questions.

Q1. Describe the population distribution (Is it dominated by high or low values? What is the overall spread of values? What is a typical state population?) [5 points]

Q2. By inspection of the histogram what would you expect the *median* state population to be? (An approximate answer is OK.) [5 points]

Q3. Would you expect the *mean* state population to be higher or lower than the median? Explain your answer. [10 points]

Next, you are going to make a histogram of the number of recorded murders by state. Go to the **HistogramData** worksheet. Here the calculations used to make the state population histogram are shown, and space in columns E through G has been allocated to calculations for the new histogram.

First you need to define a set of intervals for the histogram. The choice of intervals is up to you, but bear in mind they should be equally spaced, and neither too wide, nor too narrow to give a good picture of the data. You should examine the murder data in the **USDoJ-Data** worksheet before choosing the interval. Note that your interval boundaries should run from a value *lower than the lowest number of murders* to a value *higher than the highest number of murders.*

Note: Remember the rules you encountered in class provided by Burt and Barber (2009):

1. Use intervals with simple bounds.
2. Respect natural breakpoints
3. Intervals should not overlap and must include all observations.
4. All intervals should be the same width (equally spaced).
5. Select an appropriate number of classes.

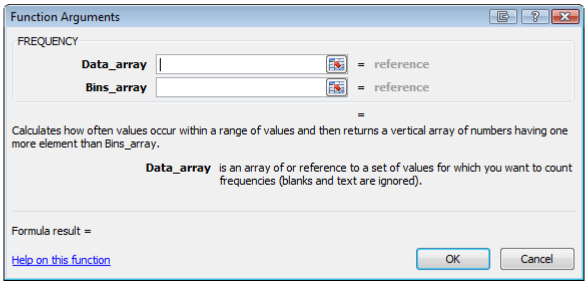
Enter your chosen interval boundaries starting in cell E3 and working down the sheet (just as in cells A3 through A7 for the state population histogram).

Having determined the intervals, you can use the Excel function FREQUENCY() to count the data values that fall in each interval. Carry out the following steps:

- Select cells in column G, starting from G3 and going down one cell more than your list of interval boundaries in column E (so if your interval boundaries are E3 through E10, select cells G3 through G11).

- Select Formulas – More Functions – Statistical - FREQUENCY

- The dialog below should appear:



Calculates how often values occur within a range of values and then returns a vertical array of numbers having one more element than Bins_array.

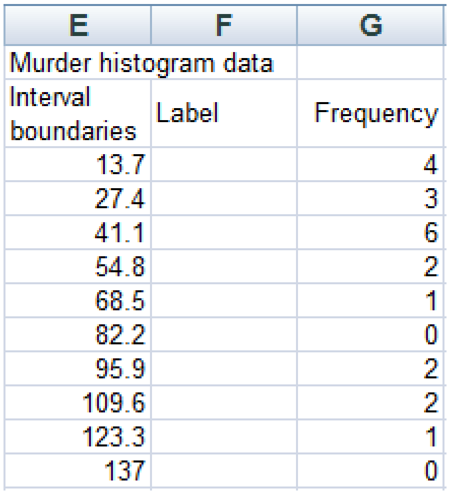
Data_array is an array of or reference to a set of values for which you want to count frequencies (blanks and text are ignored).

Formula result =

Here, you enter the cells containing the data (Data_array (from the column labeled *Murder* *and non-negligent manslaughter* in the *Practice* worksheet)), and references to the cells containing your interval boundaries (Bins_array (from the *Interval boundaries* column in the *HistogramData* worksheet)). To enter the cells, click on the button at the right-hand end of the relevant box (where it says "=reference"), and select the required information from the spreadsheet.

- When you have entered the required information, a slightly bizarre maneuver is required: with **CTRL-SHIFT** pressed click OK to confirm entry of the function [this is because FREQUENCY() is an *array function* that operates on sets of numbers. **CTRL-SHIFT** ensures that the function is entered in all the cells you selected, not just one.]

After you've done this, the spreadsheet should look something like the following, with counts for each of the intervals in column G. Note that this is not a good example of the choice of intervals.



E	F	G
Murder histogram data		
Interval boundaries	Label	Frequency
13.7		4
27.4		3
41.1		6
54.8		2
68.5		1
82.2		0
95.9		2
109.6		2
123.3		1
137		0

At this point you may want to change your choice of intervals. To do this, you should select all the cells in columns E through G from row 3 down, delete the data, and then repeat all the steps above.

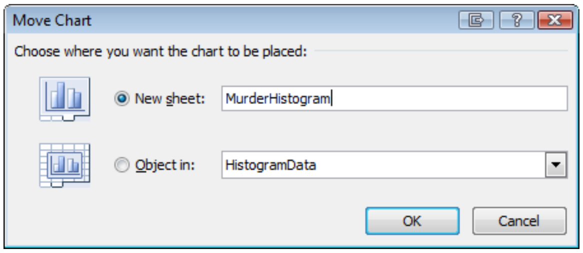
To create the histogram, proceed as follows:

- Put appropriate text in the cells starting from cell F3 and working down the column, to create labels for your intervals (refer to the population histogram to see what is required).

- Select all the cells in columns F and G from row 3 to the end of your frequency data.

- Select menu option Insert – Column and select the first option (Clustered Column)

- A chart gets inserted into your worksheet, but we want it in a new worksheet. With the chart still selected, click Design – Move Chart



Move Chart

Choose where you want the chart to be placed:

● New sheet: MurderHistogram

○ Object in: HistogramData

- Now you can tidy up the appearance of the histogram by setting various options under the different tabs. (There are three tabs under the Chart Tools category - Design, Layout and Format – each with several options)

You can also select any element in the histogram by double clicking on it, right-clicking on it and clicking *Format Data Point*.

Once you are happy with your histogram:

Q4. Print out a copy of the histogram and attach it to your submission. The histogram will be graded on completeness of labeling and overall appearance, so try to resist the temptation to go overboard with colors etc. [10 points]

Q5. Explain your choice of interval width. [5 points]

Q6. Compare your Murder histogram with the Population histogram in the Histogram worksheet. They should appear very similar. Why are they so similar? What is the major determinant of the number of murders in a state? [5 points]

3 Working with rates

From your answer to **Q6** above, you should realize that simple counts are often not very useful in human geography (in other areas of geography, too). To address this problem it is often necessary to work with *rates* of occurrence of a phenomenon. These are most often expressed as percentages (say the percentage of Hispanics living in a city), but for less commonly occurring events (such as, we hope, murder) it may be necessary to use rates per 1,000 or 10,000 or 100,000 population.

State crime rates

In this section we convert the raw crime data to rates before going any further.

The formula you need to express the number of murders in a state as a rate per 100,000 population is

$$rate = \frac{\text{number of murders}}{\text{population}} \times 100000$$

In *Excel* this looks like this:

`=X/Y*100000`

where X refers to a cell containing the number of reported cases in a state, and Y refers to a cell containing the state population. I'll talk you through the calculation for one cell, and then you can copy and paste to get the other results:

- Go to worksheet **`RatesPer100000`** cell C4.
- Enter the formula:

  `='USDoJ-Data'!C4/$B4*100000`

- **You have to enter this exactly as written.** You can do the usual thing of navigating around the spreadsheet to enter the cell references. **The $ sign is important (isn't it always?). It tells *Excel* to treat the B part of the reference as fixed.** This ensures that when you copy the function to other cells in the worksheet, it always uses the cell containing the state population as the divisor.

You should copy this function to other cells in worksheet **`RatePer100000`** before answering the next questions.

Q7. The District of Columbia has the highest rate for many of the crimes listed (not all). Why might this be considered an unfair comparison? [5 points]

Q8. Discuss reasons why even these rate statistics may not be a fair way of comparing crime in different states (*Hint.* Think about how data like these are collected.) [10 points]

One problem with comparing between different categories of crime in this way is that the rates are so different (car theft is about 60 times as common as murder nationwide). To resolve this issue we can ***index*** all the rates and express them as a percentage of the national rate. The necessary calculations have been done for you in the worksheet **`RatesComparedToNation`**.

Examine these data and answer the following questions:

Q9. Explain in words how you would do these calculations (not the *Excel* commands, the math, that is). It may help to think of two steps: (i) determine the national rates, (ii) index a particular state's rate for a particular crime. Examine the *Excel* functions in the worksheet `RatesComparedToNation` for further hints. [10 points]

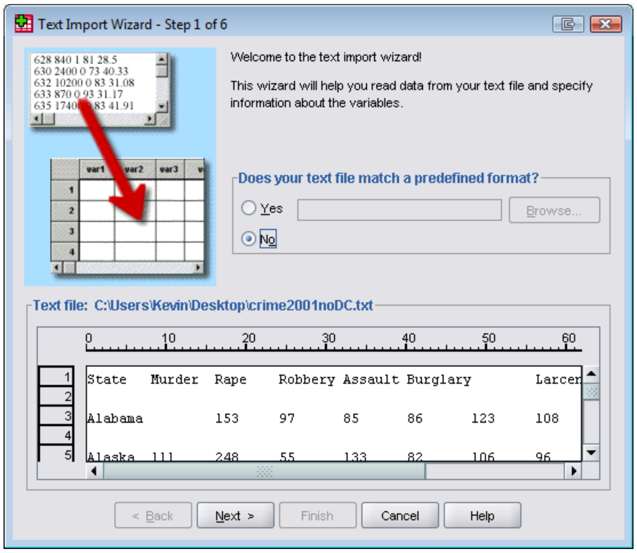
Q10. Make a histogram of the state index data for murder. Make a second histogram with the data for DC removed. Attach both histograms to your answer. Compare the two histograms with each other, and with the histogram for the raw murder data: how are these histograms similar or different, and why? [15 points]

The index data (with the D.C. results removed) have also been saved to a text file called **`crime2008_noDC.txt`** available in ANGEL in the folder Lab1. Copy this file to your local hard disc (in the lab).

4 Using PASW

Start up *PASW* (⊞ button >All Programs>Spreadsheets & Statistics>SPSS Statistics 18> PASW Statistics 18). Close the opening window.

Select File – Open – Data. Go to the drop down menu and select Text (*.txt). Select and locate the file **`crime2008_noDC.txt`** (download this from ANGEL from the Lab 1 folder). Once you click on Open the PASW Text Import Wizard will appear. This is an important exercise as you will often be required in your career to use/convert data files into different programs. Click Next in the first screen. Click Yes in the second field in the second screen to indicate that variable names are included, then click Next. Click Next in the third screen. In the fourth screen, deactivate the check button Space as a delimiter then click Next. Ignore the Error Message, click Next and then Finish in the last screen. CONGRATS, you imported the data file into PASW.



Text Import Wizard - Step 1 of 6

```
628 840 1 81 28.5
630 2400 0 73 40.33
632 10200 0 83 31.08
633 870 0 93 31.17
635 1740 0 83 41.91
```

Welcome to the text import wizard!

This wizard will help you read data from your text file and specify information about the variables.

Does your text file match a predefined format?

○ Yes

● No

Text file: C:\Users\Kevin\Desktop\crime2001noDC.txt

```
1 State   Murder  Rape    Robbery Assault Burglary        Larcen
2
3 Alabama         153     97      85      86      123     108
4
5 Alaska  111     248     55      133     82      106     96
```

For now, we will simply use one of the graphic options that *PASW* offers to depict a boxplot. To create a boxplot:

- Click on Graphs >Legacy Dialogs> Boxplot
- Click Define in the first screen
- Select one of the crime variables (just choose one) and enter it in the Variable field.
- Select division as entry for the field Category Axis.
- Click OK.

The result should be the typical PASW analysis window. The first part contains a summary of the data, which is often a useful overview of the statistics applied. The second part contains the boxplots showing the relative rates the crime you specified, with a separate plot for each of nine *Census Divisions*. The census divisions are regions of the country often used to compare statistics like these at a broad scale. The census divisions are mapped at

`http://www.bls.gov/ncs/ocs/compub.htm`

and identified by the following abbreviations:

PAC	Pacific
MTN	Mountain
WNC	West North Central
WSC	West South Central
ENC	East North Central
ESC	East South Central
NE	New England
MA	Middle Atlantic
SA	South Atlantic

Double click in the chart and a chart editor will appear. You are welcome to explore different options.

Q11. Print out and attach the boxplot you made. Comment on what it shows. Which part of the country is best or worst for this crime? How much variability do you observe? [10 points]

There are ways to create boxplots in Excel too, although it tends to be a bit elaborate: http://support.microsoft.com/default.aspx?scid=kb;EN-US;155130 The Excel solution also requires that you have calculated 5 number summaries for all groups. This can be done e.g. by using the statistical functions QUARTILE(,1), MIN, MEDIAN, MAX and QUARTILE(,3)

5 One more chart…

You should already have a good idea that you can do all sorts of clever stuff with *Excel*'s (or *PASW's*) chart options. For the final part of this assignment, I want you to explore those capabilities… so:

Q12. Make one more diagram that is not a simple histogram (pick a chart… any chart…). You can do any other calculations you need to produce the chart. Attach a printout of your chart to your answer, and explain what it shows. The most important thing here is to get some familiarity with *Excel* (or *PASW*), to explain what the chart shows, and *to label the axes and chart meaningfully, so that we can tell what it is.* [10 points]