Basic Statistics for Environmental Studies: A Distance-Learning Course
Richard Cellarius, MAP Advisor, Prescott College

Texts: Triola, Mario F. (2001). *Elementary Statistics Using Excel.* Boston: Addison-Wesley Longman. [Triola]
Moore, David S. (2001). *Statistics: Concepts and Controversies* (5th Ed.). New York: W. H. Freeman. [Moore]

Introduction

This first installment introduces you to some of the initial concepts of data acquisition and description. Moore's text focuses to a great extent on problems of gathering and describing data of various types – best summarized as "research" and "descriptive statistics." – topics covered in his Parts I and II. In his 5th edition, he makes a strong attempt to present the ideas in a conceptual, nonmathematical framework. Making conclusions comparing different sets of data and extrapolating to larger populations represented by the data obtained from a subset of that population is done with "probability" analysis and "inferential statistics," summarily described in Moore's Parts III and IV. Triola's text introduces the basic tools of probability and inferential statistics in much greater depth, while giving shorter shrift, particularly, to problems of experimentation and data gathering. Thus, for a complete understanding of obtaining reliable data and then analyzing them, both texts must be consulted. Two reasons for the selection of Triola's text over other comparable and sometime more advanced "Introduction to Statistics" texts are (1) its use of real data – with a good collection of biological and ecologically relevant data – in its examples and problems, and (2) its emphasis on Excel as a tool for basic statistical analysis.

I caution you that my primary objective in this course is not to turn you into an expert practicing statistician, but to enable you to understand and critically evaluate data analyses encountered in your readings and professional activities. However, I would also hope that you will also be able to think critically about your own data gathering activities (research) and subsequent analysis, carry out appropriate basic statistical analyses, and talk knowledgably with statisticians when more elaborate analyses of your data are needed.

In the past, I often felt that a good text can be understood by an intelligent student without elaborate lectures that repeated the text. However, I have come to realize that repetition and reinforcement of the essential ideas helps students immensely in wending their way through a complex text. In statistics, seeing examples worked through step by step, as well as working through problems yourself are both critically important to understand the process. Additionally, critically evaluating the meaning of the numerical results is essential to developing comfort and competence with statistical procedures. I hope that this series of Guides and Assignments will be effective in helping you develop an understanding of statistical analysis. As you work through them, please consider them as incomplete alternatives to lectures, where the instructor emphasizes the critical points with figurative 'cannon shots.' My notes are compact, but consider many of the sentences to be the same type of cannon shots to help you focus on the major points. In a few cases, I might even point out a disagreement with the text or identify what I consider to be a significant omission. It's your job to realize that when there are such disagreements, it may be that neither I nor the text represent statistical gospel.

Procedure

As implied above, these Guides will consist of reading assignments with a guide to essential points to focus on. It is important that you read through each example in which statistical analysis is involved in your texts step-by-step, even repeating the steps on paper or in Excel to confirm the process and results. On the other hand, you may find Moore's descriptive example after example a bit tedious; these you may want to skim quickly unless you hit one that is particularly relevant to your interests. In addition, you should view the identified PowerPoint presentations for each section of the Triola text; these are available on the Web at

http://occ.awlonline.com/bookbind/pubbooks/ triolaexcel_awl/chapter98/custom2/deluxe-content.html

and I have also saved them onto a single CD-ROM for ease of access and use, which I am sending to you. I will also loan you a series of 12 CDs with "Digital Video Tutor" files to accompany sections of the Triola text. Moore also has a web site for his text

http://www.whfreeman.com/scc

where a variety of resources are also available. Finally, I will identify a series of problems from the texts for you to work either on paper or using Excel. These will be due as indicated, usually two-three weeks from when the assignment is made. I will try to review them and give you a response within a week. You should not struggle with any problem for more than 10-15 minutes, depending on its complexity, although some Excel exercises may take a bit longer in their manipulations. If you have questions on the material that you think might be helped with an on-line or telephone consultation, ask or communicate to set a mutually convenient phone consultation time. I will also try to meet with you during the November colloquium.

I strongly recommend that you print out these instructions for reference as you read the texts.

Topic 1: Overview, Sampling, the Nature of Data, Displaying Data

Comment: This is probably the most complex assignment, because you will be going back and forth between the two texts, and it covers a great many topics which might be review of previous learning, at least in part – in fact, I hope at least some of it is review. The fundamental issue here is to understand the different types of data and how data can be obtained correctly and displayed in a way that does not mislead or hide critical information – and to recognize when the presentation distorts as well.

Reading (in the following order):

NOTE: I suggest you do the problems for each section as you encounter them. In particular, you should do the "Additional Excel Exercise #1, Part 1" (Number 3 below) before starting into Triola's Chapter 2. Also, note that Triola has a ***Vocabulary List*** at the end of each chapter. I don't believe in verbatim memorizing of definitions, but one way to review your learning of the material in the chapter is to read through the list of terms and be sure you at least understand them.

Moore: Preliminary materials—To the Teacher; Prelude: Making Sense of Statistics; Statistics and You: What Lies Ahead in this Book.

I think the section "Statistics as a Liberal Discipline" is particularly worth thinking about.

Triola: Preliminary materials—Preface.

Triola: Chapter 1 – Introduction to Statistics, Sections 1-1 – 1-4

Moore: Part I, Chapters 1-9 – "Producing Data"

Again, the critical issue in the above reading is the nature of data. It is important that you understand the difference between ***nominal***, ***ordinal***, and true numerical (***interval*** and ***ratio***) data types and can identify them when you encounter them or use them in your research. Moore has removed their descriptions from this edition of his text, but does distinguish between ***categorical*** and ***quantitative variable*****s**. One of your first tasks is to determine for yourself how these two systems fit together. It is also important to distinguish between information gathered from a full ***population*** and from a ***sample*** of that population. Triola's Chapter 1, Sections 1 & 2 PowerPoint file <C01s0197.ppt> summarizes the critical definitions that you should become familiar with. The other two PowerPoint presentations for this chapter are less important, but do identify useful concepts. Moore's discussion of sampling and sampling errors is more extensive and deserves careful reading; it should be more useful when evaluating your own and others' experiments. An important concept here is ***confounding***. One topic we'll investigate much more thoroughly in a subsequent Topic (Triola's Ch. 5) is Moore's discussion of Sampling Variability in Chapter 3. I wouldn't worry too much about trying to understand it at this point, just accept the fact that the distribution of sample means is narrower than the distribution in the population, and the larger the samples, the smaller the variation (Moore, Figures 3.1 and 3.2, p. 33).

Triola: Chapter 1, Section 1-5 – Introduction to Excel; including checking that you have the **Analysis Tool Pak**s checked and installing **Data Desk/XL** (**DDXL**) from the CD-ROM as described on pp. 30-31. You might also want to copy the data files to your hard drive's default Excel file directory for easier access. If you are not experienced with Excel's mathematical tools, you may want to go through this section in detail, following through with actually replicating the examples. Otherwise, a relatively quick, refresher skim might be all that's needed.

Triola: Chapter 2 – Describing, Exploring, and Comparing Data **NOTE**: Complete Problem Assignment 3 and read through Problem Assignment 5 below before reading this chapter.

Moore: Part II: Chapters 10-13, chapter 14, initial section on "Scatter Plots – Organizing Data

These chapters contain the critical essential tools for ***descriptive statistics***. The things to focus on are of two sorts: *displaying or plotting data* and *measures of center and variance*. I think some of the types of plots are less useful than others, and you should think about which are most useful. Think about Moore's definition of ***line plots***, and compare it with ***scatterplots***, and Figure 10.15 (p. 190—see also his problem 10.12 on that page). One of the failings of Excel is its default chart format for a line plot rather than a scatterplot. Of course, it will depend more than a bit on the type of data being displayed. Also don't neglect the discussions of bad displays; one of the exercises that I have designed below is a term-long survey of data displays in the public media. With respect to the quantitative measures of center and variance, it is important to understand the meanings of the various measures and their applicability: ***mean*** (also known as ***average***), ***median***, ***mode***, ***standard deviation*** and ***variance***. For many types of data the ***five-number summary*** and ***boxplot*** are useful tools during the investigator's analysis of her data, but I think you rarely see them in published papers.

Several topics in these sections will reappear in more detail in the next study section or two, so I wouldn't spend too much time trying to figure them out at this point other than accepting the descriptions. In particular, we will revisit the ***normal distribution*** in glorious detail in later chapters in Triola. However, the various "rules of thumb" regarding the distribution in a population based on the descriptive statistics parameters are important enough that you should commit them to memory: ***Range rule of thumb***, ***Empirical (or 68-95-99.7) Rule***, and ***Chebyshev's Theorem***.

Here you might want to go back and forth between Triola and Moore as follows, with the caution that you might find reading Moore before Triola would be more useful: (1) Displaying Data – Triola Sections 2-1 to 2-3 and Moore, Chapters 10 and 11; (2) Measures of Center, Variation and Position – Triola Sections 2-4 to 2-7 and Moore Chapter 12.

I think the PowerPoint files for Triola Chapter 2 are OK, but not critical. However, can you spot the error(s) in slide 22 of C01s0497.ppt? You'll get 5 bonus points if you get it right! ☺

Moore: Chapter 16 – Consumer Price Index and Government Statistics (Optional, good contribution to your liberal education)

Problem Assignments:

1. Moore, Part 1: For *Part I*, *Chapters 1-9*; , choose **2** exercises to do from the end of each chapter and for *Part I*, *Review*, choose a total of **3** Exercises, making the selections as follows:
   (1) for each chapter and the Review, determine the range of problem numbers, ignoring the Chapter number prefix.
   (2) Initially, select a line of Table A (pp. 545-6) at random, by closing your eyes and placing your finger on the page, then open your eyes, and select a line number. Write it down or mark on the page for future reference;
   (3) using the procedure described in Step 2 of Moore's Example 3 (p 23), find the first two numbers that represent exercises in that chapter, starting with the first column of the line. If the first acceptable number encountered is *odd*, select the next acceptable *even* number as the second, so that you end up with an odd and even exercise to do from each Chapter. For the Review section, use the same procedure for the first two exercises, and then select the next acceptable number
   (4) Note where you ended the problem identification for that section and use that point for the selections for the next section.

   For example, in Ch. 2, section 2 the range is 1-17. If we were to start at line 101, the first 2 digit number sequence corresponding to a valid problem number is 05, odd, and the second is 13, also odd, so we'll go on to the first two-digit even in the range 1-17; after several rows you finally find 02. So you would do problems 2.2 and 2.5; and start your selections for Chapter 3 in the middle of line 104. Once you get the of hang it, your selections should go quickly. Please indicate your starting line for each chapter's problem responses.

2. Triola, Ch. 1: Review Exercises 1, 4, 6 (pp. 34-35); Cumulative Review Exercises 4, 5, 7, 9* (p. 36). *For problem 9, show the result in both Scientific notation and normal notation in adjacent cells (use the cell reference tool to avoid repeating the calculation, i.e., if the original calculation is in B3, enter **=B3** in C3 and format the number differently.

Qwerty	Dvorak
2	2
2	0
5	3

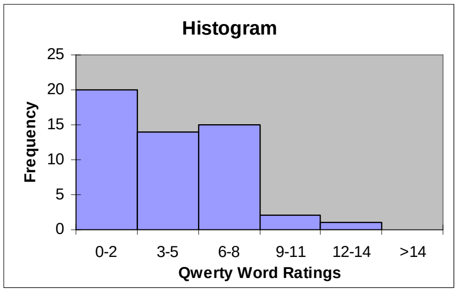
3. Additional Excel Exercise #1, Part 1: On a separate worksheet of the same Excel file you used for the calculations above (Cumulative Review Exercises), type in the data from Tables **2-1** and **2-2** on Triola's p. 41, putting each set in a parallel column. The first 3 numbers for each are shown on the right. In a third column, enter the differences as printed in the **Example** on p. 113. Assuming the data are in columns A, B, and C, with the

numbers starting in Row 2, in cell D2 enter the equation **=(A2-B2)-C2**, and copy it to all the rows with data in columns A, B & C. All numbers in column D should be 0. If they're not, check your data entries and correct accordingly. What happens if you change the sign on one of the numbers in the Difference column? How easy is it to identify the error in column D? Another trick is to type into cell E2 the formula **=if(d2=0,"","Oops")** and copy down the column as before. How visible is your "error" now? Remember these tricks for comparing columns as a way of checking your data entry accuracy. Correct any errors; normally you would then delete columns D and E, but leave them now to show your work. Right click on the sheet tab for this sheet and rename it "**Keyboards**." Rename the first sheet "**Triola Ch1**" Save the file to your disk for future use. Label it **Stats01.xls**.

4. Moore, Chs. 10-13: Again select **2** problems from each section as describe above, but do one on paper and the other using Excel, whichever is appropriate. Use additional worksheets in the Stats01.xls file for these problems, putting more than one problem on a worksheet if appropriate, and labeling them properly as well. BONUS: Using the numbers in Figure 10-15 (p. 190), make a vertical table in Excel, with the year dates in the first column and the corresponding Pound values in the second column. Then create two Excel charts and place them side-by-side on the same worksheet. (See the Appendix for instructions for doing this.) For the first, use Excel's Line Graph configuration and for the second use the XY (scatter) option, selecting the version with straight lines (lower left on the "Standard Types" page. Compare the two charts and consider Exercise 10-12 on the same page.

5. Additional Excel Exercise #1, Part 2: As you read through Triola Ch 2, try to replicate the "Using Excel…" examples using the QWERTY keyboard data you entered in Assignment (3) above. Show your working of these examples on the **Keyboards** worksheet, making sure all results are properly labeled. Note: I cannot replicate the third histogram on p. 54, but here are some other things to try with regard to the Frequency and Histogram functions.

(A) First enter a column of "bin" numbers: 2, 5, 8, 11, 14 in column F or G. Then, using Excel's Data Analysis Tool, Histogram, as described in the text, replicate the first histogram on p. 54. using the bin number range you just entered. Be sure the Output Range (it only needs to be a single cell) is in a different column from the Bin Range. Then format the chart as in the second diagram. Finally, replace the entries in the "Bin" column of the frequency table produced by the Excel tool by the actual ranges as follows: instead of the number 2, type '0-2 (include the single quote mark) and do the same for the other ranges, 3-5, 6-8, etc., remembering to include the single quote. The result should look like the figure to the right of this paragraph. This works because Excel considers that the x-values in a chart are labels (nominal data) and not numbers. The quote mark tells Excel the entry is text; experiment by taking them out: what happens? You'll find Excel has a number of annoying automatic behaviors!



(B) Again enter the same set of bin numbers in a column in a blank section of the worksheet. Then using your mouse, select all the cells in the adjacent column to the set of bin numbers; for example of the bin numbers are in cells E43:E47, select cells F43:F47. Then click on the function tool (f_x) main toolbar. Select "Statistical" as the **Function category:** and "Frequency" as the **Function name:** and click **OK**. For the **Data_Array** enter the range for the Qwerty data, presumably A2:A53. One simple way to do this is to click on the red & white checker board icon on the right end of the blank (note it reappears as a single line at the top of the sheet) and select the range of data with your mouse (the cell range appears in the blank), and again click on the red & white checked icon. Do the same for the **Bins_array**, entering the range of cells for the bin numbers, e.g., E43:E48. Click **OK**. The number **20** appears in the first selected cell and nothing happens to the rest. Oops. We forgot that the Frequency is an *Array Function* in Excel. With the same cells still selected, again click on the function tool (f_x). The function fill-in box appears with all the data still identified. Note that just below the two sets of data entries, there is a third line, with the full set of frequencies enclosed in curly brackets: {20, 14; 15; 2; 1; 0}. Note that the function description says it "returns a vertical array of numbers having one more element than the Bins_array." Now, hold down the **Ctrl** and **Shift** keys at the same time as you click on **OK** (or you could also just press **Enter** on the keyboard). Note that the full set of frequencies now appears in the selected column of cells. That column of selected cells is surrounded by a thick black line with a small black box in the lower right-hand corner.

Move the mouse pointer over that small box, and note that it becomes a small cross (+). Click on it and drag down one more cell. The number **1** appears in the cell. If you select any of the first 5 cells you had highlighted, you will note that they contain the same expression, **{=frequency(<data_array>,<bins_array>)}** with the formula enclosed in curly brackets, with identical ranges in each cell for the two arrays, but the ranges are different in the last cell you just added. Now again select all of the frequency array cells, including the new one, select the function tool (f_x) and again press **Ctrl** and **Shift** and **OK** (or **Enter**). Now all cells have exactly the expression you want and **0** appears in the last cell. What is the significance of the 0? (Hint: How many numbers in the Data_array are there that are larger than 14?) You have just been introduced to one of Excel's Array functions, which produce a set of data which fills several cells in an array. To complete the entering of an array function, you must press **Ctrl** and **Shift** and **Enter** in order to get the full array result.

6. Triola, Ch 2: Review Exercises 1-6. Cumulative Review Exercise 3. In addition, based on your answers to Review Exercises 1-6, write a few paragraphs about the results, such as you might find in a "Results and Discussion" section of professional journal paper.

7. Additional Excel Exercise #2:

(A) Enter the data below into an Excel worksheet, using a new file, and a separate worksheet for each set of temperature data, January, July, and Annual. The data should be entered as a single column (column B on each sheet), with the corresponding years in column A of each sheet. How could you check to assure that your data have been typed in correctly beyond comparing each value with the one printed here? I can think of at least three, two of which involve a coworker. Use one of them to check each column of data.

Average surface temperatures, Northern Hemisphere, °F, based on 1961-90 average of 57.2 °F								
Year	Jan	July	Annual		Year	Jan	July	Annual
1981	58.23	57.34	57.63		1991	58.06	58.05	57.83
1982	56.93	57.33	57.27		1992	58.19	57.22	57.47
1983	58.21	57.67	57.63		1993	57.94	57.51	57.52
1984	57.54	57.38	57.11		1994	57.72	57.76	57.79
1985	56.97	57.04	57.04		1995	58.51	57.99	58.15
1986	57.65	57.31	57.36		1996	57.58	57.85	57.60
1987	57.40	57.74	57.56		1997	57.92	58.19	58.10
1988	57.99	57.87	57.70		1998	58.15	58.64	58.42
1989	57.40	57.85	57.60		1999	58.19	57.99	58.01
1990	57.76	57.85	58.03		2000	57.70	58.01	57.94

(B) For each set of data, January, July, and Annual, use the appropriate Excel function or tool to produce a complete set of descriptive statistics. Also use appropriate DDXL tools to analyze them (HINT: try copying all three data sets to a single worksheet and use DDXL to produce a side-by-side boxplot).

(C) Write a short narrative description of what conclusions you make from these descriptive statistics, particularly identifying any interesting differences in the three sets of data.

8. Term-long assignment: At least once each week, select an item from a daily newspaper or weekly newsmagazine (even a news-oriented web site is OK) that has a display of numerical or survey data. (See your texts for examples of the types of things to look for.) Cut out or copy the article and data display and critique it in terms of its accuracy in representing the results. This is a similar but longer-term assignment to that in Moore's "Writing Project" 4.1 (p. 282). Submit the published material, properly identified with source, page, and date, along with your critique. Send in these critiques with each submission of your statistics work.

Appendix

Instructions for formatting charts for the BONUS exercise under problem assignment group 4, p. 1-4:

Using the data in Figure 10-15 on Moore's p. 190, in a new Excel sheet, enter the years (1925-1975) in cells A2:A7 and the corresponding dollar values ($4.86-2.03) in cells B2:B7. Type "Year" in cell A1 (no quotes), and "$/£" in cell B1. (You can get the pound symbol (£) as follows: with your cursor still in the cell after "$/", click on "Insert" on your top menu, then select "Symbol", then the "Symbol" tab, and scroll up or down to find the £ symbol. Click on it, then click on "Insert" at the bottom of that window and then on "Cancel" to return to the Excel screen; the £ symbol should appear in cell. Press "Enter" to complete the text entry.)

To make the first chart, select all the cells A1:B7 with your mouse cursor, then click on the "Chart Wizard" icon, which looks like a vertical bar chart; it should be on the top "Formatting" tool bar. Select "Line" Chart type and the default "Line with markers at each data value" sub-type, and press "Next". On Step 2 of the Chart Wizard – "Chart Source Data," click on the "Series" tab. In the lower left-hand window, the word "Year" should be highlighted, if not click on it to select it. Select the text in the middle right-hand window labeled "Values:", which should be "=Sheet1!$A$2:$A$7"; copy it by pressing CTRL-C, and then paste it into the bottom window labeled "Category (X) axis labels" by pressing CTRL-V. Then click on the word "Year" and then the "Remove" button. The chart in the upper window should begin to look like the figure in the text. Press "Next" and the Step 3 –the "Chart Options" window should appear. (1) Title the chart axes as desired. (2) Click the "Gridlines" tab, then the "Major Gridlines" box under "Category (X) axis" to select it and the "Major Gridlines" box under "Category (Y) axis" to <u>un</u>select it. (3) Click on the Legend" tab, and click on the box "Show legend" to remove the legend (you don't really need it since you have only one line on the chart). (4) Then select the "Data Labels" tab and select the "Value" box. Finally click on "Next", accept the default "Chart Location – As object in:" <your active sheet ["Sheet 1"]> and press "Finish". Click on the chart and with your mouse, move the upper left corner of the chart to cover cell C1 and then click on the black square in the lower right hand corner of the chart and resize the chart so the black square is at lower right corner of cell F17. Now click with the <u>right</u> mouse button on the lower (x or time) axis and select "Format axis". On the "Scale" tab, click on the box "Value (Y) axis crosses between categories" to <u>un</u>select it and click OK. The chart should look very similar to the one in your text, although the first data point will be on the Y-axis.

To make the second, "XY (scatter)" chart, again select the data in cells A1:B7, click on the Chart Wizard icon as before, and this time choose the XY (Scatter) Chart type option and the lower left hand Chart sub-type ("Scatter with data points connected by lines). Click "Next" and note the sample chart already has the correct form. Again click next and repeat the various Chart option steps: Title the axes as desired, select X-axis gridlines rather than Y-axis gridlines, unselect "Show legend," and select Y Value Data Labels. Click "Finish." Move and resize the chart so that it fills cells G1:J17, which should make it the same size as the first chart.

Compare the two charts. How would you describe the difference? Which do you think is a better presentation of the data? Why? (Hint: think about the difference between ordinal and interval scales.)

As a last experiment, <u>right</u> click on the first chart, select "Chart Options…" and the "Axes" tab. Select "Time Scale" rather than "Automatic" and click OK. Notice the similarity of the two charts now, although the units on the x (time) axis of the first chart are in dates (m/d/y) rather than years. (You can adjust the time scale values by right-clicking on the time axis and "Format axis.") This shows that Excel's default Line Chart option can be formatted with the x-axis as an interval scale *as long as you are aware of the difference and why you would need it rather than just an ordinal scale.*

Appendix version 2, 2/28/05