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]

Topic 2: Probability and Probability Distributions; Sampling Distributions and the Central Limit Theorem

Comment: **Probability** is one of the most fundamental concepts underlying statistical analysis: given a number of events, what is the probability that a specific event will occur or, if there are repetitive events, what is the proportion of a specific kind of event that is likely to occur? For statistics, probability can be approached with a full-blown coverage of probability analysis or with a survey of the basic fundamentals sufficient to provide a sense that the statistical procedures encountered later have an adequate mathematical basis, but not elaborate proofs of everything. I have chosen a route closer to the latter – similar to Triola's approach – which does not require learning elaborate notations or doing excessive calculations. Nevertheless, it is important to pick up the basics. The ultimate conclusions to focus on are the nature of **probability distributions** – including the Standard and Nonstandard **Normal Distribution** – **sampling distributions**, and the **Central Limit Theorem**, which provides the fundamental rationale for relying on samples of a population to make statistical conclusions about the entire population. I have created an additional Excel exercise that illustrates its power.

Readings (in the following order – also note the order in the problem assignments – and do the problems at the end of a section following your reading of that section before going on to reading the next section):

Moore: Part III, Chapters 17-18 and *19-20* – Chance.

Triola: Chapters 3 except section 3-5, 4 and 5, sections 5-1 through 5-4 – Probability and Probability Distributions

Moore gives a very good descriptive introduction to probabilistic thinking, and it is useful particularly for thinking about how probability thinking is used and misused in everyday life. It provides a good real-life introduction to the more technical presentation in Triola. As in Topic 1, I recommend that you skim over his examples, trying to pick up his main point but not getting bogged down in the details unless you find an example particularly interesting to you.

The importance of probabilistic analysis should become clear as you work through Triola's section 3-7, "Counting," which, in my opinion is essential for making sense of probability distributions. The essential concepts here are the basic idea of **probability** as described in the three "Rules" on p. 129, the **Law of Large Numbers**, and the circumstances in which one applies the **addition rule** and **multiplication rule**. I urge you to focus on the actual manipulations and not get bogged down in the statistical notations. I think the multiplication rule is essential for understanding the **counting rules** in section 3-7. In this section, work on the distinction between what situations the two **permutation rules** and the **combination rule** actually refer to and *just accept* the actual formulas; you should know what the expression ***n!*** ("n-factorial") means—ask me if you don't and see p. 29 in Triola for how to express it in Excel. Excel also has formulas for calculating permutations (**PERMUT()**) and combinations (**COMBIN()**), which you can – and should – check using the algebraic/factorial expressions as well.

The "Overview" section of Triola's Chapter 4, Section 4-1, has a very important idea that, again, underlies the basic approach of statistical analysis: "we develop probability distributions that describe what *probably* will happen instead of what actually *did* happen." The ultimate goal of statistics is to compare what happened with what normally *might be expected to happen, based on specific assumptions*.* If the actual event turns out to be improbable according to our expectations, then we conclude our expectations were wrong. In Chapter 4, the important ideas are the differences between **discrete** and **continuous random variables**, the different types of **probability distributions** and their descriptive statistics, **mean**, **variance**, and **standard deviation**. The **expected value** is another way of thinking about the mean of a discrete random variable. Again in all these distributions – which describe different situations – the issue is, what is the **probability of a particular event** (or type of event) occurring?

* This is a "cannon-shot" concept that you will want to keep in mind – and come back to as needed.

Chapter 5 in Triola gets to the fundamental heart of inferential statistics: the **standard normal distribution**, the relationship between ***Z*-scores** and the **area** under the standard normal distribution curve, and finding *Z*-scores and area or **probability** values. It is critically important to be sure you can

correctly do and understand the assigned problems using Table A-2 and Excel's **NORM(S)DIST()** and **NORM(S)INV()** functions and their interpretation, as well as Triola's figures in Sections 5.2-5.4.

Triola: Chapter 5, sections 5-5 through 5-7 – Central Limit Theorem and more on the Normal Distribution and normality.

In my humble opinion, *the **Central Limit Theorem** provides the fundamental rationale for relying on samples of a population to make statistical conclusions about that population.* It is the theoretical basis underlying Moore's assertions about **sampling variability** in his Chapter **3** (pp. 32-35) I have created an Excel exercise that I believe illustrates its power, which you will have the pleasure of working through. The last two sections of Chapter 5, Sections 5-6 and 5-7, deal further with the relationship between the standard normal distribution and determining whether a distribution is normal.

Problem Assignments (page numbers refer to location of the problems):
Note: There are a lot of problems here, but many of them are short answer, often look-up questions and will not require extensive calculations or analysis. For those that do involve calculations, you are welcome to let Excel do the arithmetic. I have selected them to give you a fair amount of practice in working with these essential concepts. Mastery of these will make carrying out the exercises in subsequent chapters much easier, because they will involve similar processes with other statistical concepts. Be conscious of your time on these. You should not struggle longer than 5 minutes or so on any problem if you do not see how to approach it. Some of the exercises, especially those involving Excel, will take time to work through, however, and I would prefer you finish them entirely if can see your way to the end.

Remember too that both Moore and Triola have answers (but not the solution process) to odd-numbered problems at the back of the book. Please work through the problem first before checking their answer. *But there is one thing I would caution you about*: there is no absolute guarantee that their answers are always correct – there might be a typo or error, so if after a reasonable struggle you can't get the book's answer, just move on, but show your work and conclusion. ***It is critically important to my review of your work that you include all your calculations and procedure. Try to write clearly as you work through a problem and do NOT do your calculations on a separate sheet of paper. If you use Excel to do your arithmetic, please send the worksheet file along as well*** – I encourage you to use multiple worksheets in a single Excel file with each section's problems together on a separate sheet, except that the Central Limit Theorem exercise should have its own, separate file.

Before you start these problems, I recommend you copy the files The Uniform Distribution.pdf, Excel-Central Limit Theorem.pdf, and 16Sheets.xls from the folder labeled "**Richards Files**" on the extra CD-ROM disk I sent you with the Video Tutor disks to an appropriate folder on your hard drive. You will need them for problem groups (6) and (8) below. You will also need the Adobe Acrobat Reader, available at http://www.adobe.com/products/acrobat/readermain.html. The Windows version of the reader is also available on the disk as ar505enu.exe.

1. Moore, Ch. 17 (pp 357-361), #17.6, 17.7, 17.20; Ch. 18 (pp. 369-375), #18.3, 18.5, 18.11, 18.19.
2. **Additional Excel Exercise #1**: Carry out the Excel Demonstration of the Law of Large Numbers as described on Triola's pp. 131-132. When you have your chart formatted to your satisfaction, click on any cell outside the chart and press the function key F9. What happens? Do this multiple times and write a brief statement of what you observe and what you conclude from your observations in terms of the Law of Large Numbers. Include this as a worksheet in the Excel file with this Topic's problems.
3. Triola, Ch. 3, *Sect. 3-2* (pp. 138-141), #1, 2, 14, 23; *Sect. 3-3* (p. 148), #1-4, 7, 8; *Sect 3-4* (pp. 156-158), #1, 2, 5, 16; *Sect. 3-6* (pp. 168-169), #3, 4, 7*; *Sect. 3-7* (pp. 179-181), # 1-4, 9, 23.
   * Hint (see also Triola pp.167-8) Use Excel's RANDBETWEEN() function to create 3 columns of 100 numbers each, with each number randomly between 1 and 6. In the next column of cells, insert the total (SUM()) of the three cells containing the random digits in the respective row. Then using the COUNT() function, find the total number of responses (the result should be obvious), and using the COUNTIF() function, find the number of cells containing a total of 10. Then calculate the ratio to determine the answer. Press F9 as described above to see the variation. The text's answer of "approximately 0.125" is a theoretical calculation based on the number of ways 10 can be obtained (27) compared to all the possible combinations for the 3 dies (${}_6C_1{}^3$=216): 27/216 = 0.125
4. Moore, Ch. 20 (p. 402), #20.11; Part III Review (p. 408), #III.3. Triola, Ch. 4, *Sect. 4-2* (pp. 207-210), #1, 2, 15, 23; *Sect. 4-3* (pp. 217-219), #2, 3, 5, 8, 11, 14, 17, 20, 21, 24, 28, 29; *Sect 4-4* (pp. 223-226), #1, 4, 14, 17; *Sect. 4-5* (p. 231), #10.

5. Triola, Ch. 5, *Sect 5-2* (pp. 254-256), #2, 7, 10, 11, 25, 26, 41, 43a&c.
6. **Uniform Distribution exercise** – Print out the one page document, The Uniform Distribution.pdf (see above). Work through the Uniform Distribution worksheet and do the requested calculations. Save this to submit with the other assigned work for this topic. You will need to do this *before* you do the Central Limit Theorem exercise below, since you need to refer to your calculations here in that exercise.
7. Triola, Ch. 5, *Sect. 5-3* (pp. 261-264), #1, 8, 22, 25, 26; *Sect 5-4* (pp. 267-270), #1, 4, 15, 16.
8. **Additional Excel Exercise #2**: ***Central Limit Theorem***. Do this exercise ***after*** finishing your reading of Triola's section 5.5.
   A. Open the **Excel-Central Limit Theorem.pdf** file (see above). Print out this file on paper. You will find it much easier to follow the instructions when they're printed, rather than trying to go back and forth between computer screens.
   B. Now work through and carry out the Central Limit Theorem Demonstration instructions. In doing so, you should read the left-hand columns of each instruction step completely, before carrying out the manipulations in that step. When you have completed each step, read the explanation in the right-hand paragraph to help you understand what happened. It should take you about 2-3 hours to complete the exercise, depending on your familiarity and dexterity with Excel. You will use the **16Sheets.xls** file in the exercise. I apologize for the detailed steps in the computer instructions, but they're important for those not fully familiar with Excel and its tricks. Many of them will be useful in other applications.
   C. It is critical that, in addition to doing the steps outlined in the instructions, you write the summary and analysis described in the final "**Finishing up**" paragraph.
   D. The final file you have at the end of the exercise is pretty large. I would not recommend trying to e-mail it; just copy it to a disk and submit it with the paper work for this Topic's assignments.

   Good luck. Be sure and consult with me if you run into problems. I would also appreciate it if you would identify any typos or other errors in the instructions.
9. Triola, Ch. 5, *Sect. 5-5* (pp. 278-281), #3, 4 – add a part **c** to 4 by doing the calculation in **b** for 37 women and explain the difference in the two answers – 11, 17 – in 17, explain the logic for the book's answer of 0.0069, rather than 0.486; *Sect. 5-6* (pp. 290-292), #1, 4, 11, 12, 23; *Sect. 5-7* (pp. 297-298), #1-4, 7, 11.