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 3: Introduction to Inferential Statistics – Confidence Intervals; Hypothesis Testing and Tests of Significance

Introductory Comment: The two parts of this Topic introduce you to the essential concepts and tools of **Inferential Statistics**, the science and art of making conclusions about ***populations*** from ***statistical samples*** of those populations. In the first part, the task is to determine the probable range (**confidence interval**) within which the "true" population value of the quantity in question (**parameter**) lies based on analysis of the **statistic** (**point estimate**) obtained from the sample data. In the second part, the process is extended to make comparative statements or conclusions based on **hypotheses** about the population, again on the basis of the sample data from that population. To repeat: these are the essential tools, and it is worthwhile to work through them slowly and thoroughly to be sure you have not only a grasp of the concepts but also the ability to apply them. While the number of assigned problems may seem large, many of them involve simply looking up a number in a table, an exercise that should take less than a minute – or even less once you get the hang of it. Both the Video Tutor files and PowerPoint files should be viewed in connection with the relevant readings to assist your understanding, although those that focus solely on using the TI-83 calculator are less important. One final thought: as in most natural science and mathematics, learning statistics is a cumulative process, *i.e.*, gaining an understanding of later concepts depends on having a good understanding of the previous concepts. If things have made sense so far, you're in good shape. If they're still a little shaky, hopefully they'll become clearer as you apply them to the new material. We don't want to think about the alternative. ☺

Along with this file, I have included an Excel file (**Temps-Table 6-1.xls**) with the temperature data at the beginning of the two Triola chapters for this topic. I recommend that you work the example exercises in the text using this file to help cement the concepts. Although the two tables, Table 6-1 (p. 311) and Table 7-1 (p. 382), are laid out differently, they contain the same data.

Topic 3a: Confidence Intervals

Readings (in the following order):

Moore: Chapters 21, 25 (pp. 485-490), – What is a Confidence Interval?; Sampling distribution of a sample mean; Confidence intervals for a population mean.

Triola: Chapter 6, except section 6-6 – Estimates and Sample Sizes. I recommend doing the assigned problems for each section immediately following your study of that section before going on to the next.

Moore's Chapters 21 (pp. 419-427) and 25 (pp. 485-488) first review some of the material you've seen before regarding the relationship between a sample ***statistic*** and the corresponding ***parameter*** of the population that was sampled and Ch. 25 revisits the **Central Limit Theorem** (esp. pp. 485-487). They then show how to determine ***confidence intervals***, first for population proportions (Ch. 21) and then for population means (Ch. 25, pp. 489-490). You should read these sections for the overview, and should not get bogged down in the details of the calculations. They will be discussed in what might appear to be excruciating detail in Triola.

In Triola's Chapter 6, there are several important topics, (a) **confidence intervals when the sample size is >30**, (b) **confidence intervals when the sample size is £30 and the population is normally distributed**, (c) **determining the sample size needed to get an estimate of the mean within a desired confidence interval**, and (d) **confidence intervals and sample size estimates for population proportions**. The most important concepts here are ***confidence interval*** – and the related ***margin of error*** (***E***) – and ***degree of confidence*** (***level of confidence***, also referred to in Moore and other texts as ***confidence level***; it is related to ***a***, which is referred to later in the text as ***significance level***, by the expression ***1-a***). In the discussion and application of these new concepts, you will see three important concepts that you have seen earlier return: ***z*-scores**, **probabilities of a *z*-value** from the normal probability distribution as given in Table A-2 of your text, and the **standard deviation of sample means** based on the Central Limit Theorem.

When you get to section 6-3, note that the acceptable conditions for dealing with small samples are more stringent, particularly requiring a normally distributed population, but the formulas and approach only differ by calculating *t* values and using a different distribution, the **Students *t* Distribution**, values of which are in Table A-3. As your text describes, it is a probability distribution that is a bit flatter than the normal distribution and depends on the size of the sample, which takes into account the greater statistical uncertainty when you have small samples. ***The real tricky thing here is how the table is formatted***: the values in the body of the table are the values of the distribution, *t*, and the probability values (area under the tail of the curve) and **degrees of freedom** (***df*** = *n*-1, where *n* is the sample size) are in the margins of the table. This arrangement is exactly opposite to that of the normal distribution/*z*-value table, where the *z*-values are obtained from the margin of the table and the areas (probability values) are found in the body of the table. Compare the examples in the PowerPoint files **C05s197.ppt**, slides 20-22, (for *z*) and **C06s397.ppt**, slides 20-22, (for *t*). Note that the last line of numbers in Table A-3 gives the respective *z*-values for the probabilities when

dealing with large samples. In contrast to the difference in table construction, the actual steps of analysis and the equations used are virtually identical. The example in the video for Section 6-3 only shows how to do the analysis using the TI-83 calculator, so I would recommend skipping the example and rely on the PowerPoint show **C06s397.ppt** instead.

Problem Assignment (page numbers refer to location of the problems):

Note 1: I have only assigned problems from Triola except for a few from Moore at the very end of the assignment. Again there are lots of problems here, but, as before, many of them are short answer, often just table look-up questions and will not require extensive calculations or analysis. I have selected them in order 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.

Remember too that Triola has answers (but not necessarily the solution process) to odd-numbered problems at the back of the book. Remember that there is no absolute guarantee that the answers are always correct – there might be a type 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.***

Note 2: Excel uses an approximation method for the distribution functions and your text tables are rounded off, so any answers you get using Excel may differ in the 3rd or 4th or higher significant figure from those you get with your hand calculator or those given in the back of the book. However, you can be confident you got the same answer if the first three digits agree, esp. when you apply the round-off rules.

1. Triola, Ch. 6: *Sect. 6-2* (pp. 325-328), #3-6 (use Table A-2), 11, 16, 21 (use **COLA.xls**†), 27 (if you can do #27, you've got a good grasp of the concepts); *Sect. 6-3* (pp. 336-339), #1, 4, 7, 10 (use Tables A-2 or A-3 as appropriate for these 4), 12 ***(no calculation required)***, 17 ***(use Excel)***, 18 (use Table, check with Excel); *Sect 6-4* (pp. 343-345), #1, 3, 4 (do these on paper and confirm using Excel as described below*), 11 (use **BOSTRAIN.xls**†); *Sect. 6-5* (pp. 353-358), #3-6, 9-12, select 1 odd and 1 even problem from 21-32 (your choice of topic; no need to use the random number table). [†**NOTE: *COLA.xls* and *BOSTRAIN.xls*** *are in the* ***DATASETS/EXCEL*** *directories on the CD-ROM disk in the back of your text as well as the Course CD-ROM with the instruction files*.]

   * When you use Excel to check your calculations for problems 1, 3, & 4 in Sect 6-4, follow the method described on pp. 342-343, with three modifications: (1) use the NORMSINV function rather than the NORMINV function – NORMSINV gives value of the standard (that's what the *S* means) normal distribution rather than that of a normal distribution where you have to specify the mean and standard deviation of the distribution (note the text has to put the 0 and 1 for the mean and standard deviation into NORMINV to get the standard normal distribution); (2) don't enter the numbers into column A and the functions into column B. but put a label for each variable in column A, the actual value in column B and the function in column C, so you know what each number refers to; and (3) use Excel's integer function (INT), which returns the integer part of a decimal number, to do the rounding up by adding 1 to the integer returned In actually entering this into the worksheet, be careful that the cell locations referred to in the function statement reflect the cell where the respective variable is. The table below shows how the worksheet would be set up if you start in cell A6. I recommend you set this up as shown below, which uses the example of the text. I got 294.88549 instead of the 294.69444 given in the example box in the text; it still rounds up to 295.

Row\Col	A	B	C
6	s (std dev)	20	n
7	E (marg of Error)	3	**=(NORMSINV(B8+0.5*(1-B8))*B6/B7)^2**
8	Conf Level	0.99	**=INT(C7)+1**

   One advantage of this is configuration, which you should utilize, is if you copy the full calculation set-up (cell range A6:C8 in this case) to another location (starting in A12, for example), you can just fill in a different set of numbers for a different problem. Thus, once you have this set up, you can do all three problems almost instantly.
2. When you have finished the problems for Ch. 6, write a brief paragraph explaining the relationships among the following quantities: **standard error** (as defined in Ch 5), **point estimate**, **confidence interval**, **margin of error**, and **degree of confidence**.

Topic 3b: Hypothesis Testing and Tests of Significance

Readings (in the following order):

Moore: Chapters 22, 25, 23 (in that order) – Inference: What Is a Test of Significance?; Inference about a Population Mean; Use and Abuse of Statistical Inference

Triola: Chapter 7, except section 7-6 – Hypothesis Testing. Again, I recommend doing the assigned problems for each section immediately following your study of that section before going on to the next.

Moore's Chapters 22 and 25 introduce the concept of hypothesis testing and its application to hypotheses about means and proportions. Again, you should read these sections for the overview and not get bogged down in the details of the calculations, as the topics will be covered in more detail in Triola. Moore's Chapter 23 presents a good set of ideas and

cautions about applying statistical considerations. You might want to review it again *after* finishing the assigned Triola problems.

Triola's Chapter 7 introduces the essential tool of inferential statistical analysis, the use of calculations of **statistical significance** based on sample data, to compare two alternative, mutually exclusive **hypotheses** about the population or populations from which the sample data were obtained. It is critical that you make sense of this chapter and the methods used for **hypothesis testing**. In particular, you should almost get to the point where you can recite the 8 steps outlined in Figures 7-5 (p. 399) and 7-9 (p. 405) in your sleep. I don't think you need exactly to memorize them, but they need to be an essential component of each statistical analysis that you do, and each step should be explicitly stated as you go through the analysis. Once you have this down, then you will find that the types of analysis you encounter in the future differ primarily in how you calculate the **test statistic** (step 6 in the hypothesis testing procedures).

In some ways, the first tests introduced are similar to finding the confidence interval for the population mean. Specifically, the tests can be thought of as finding the probability that a hypothesized population mean (as stated in the **null hypothesis**, H_0) falls within the confidence interval calculated from the sample data, or outside that range (as stated in the **alternative hypothesis**, H_1 – referred to in some texts as H_a). The situation can get a bit complicated depending on whether you're testing for equality *vs*. inequality or *vs*. a value that is greater than or less than the hypothesized value. You should not get confused by the two different approaches described in sections 7-3 and 7-5 as the "traditional" method and "*P*-Value" method. They should result in the same conclusion, and differ only in how the conclusion about the null hypothesis is made from the test statistic.

In this chapter the analyses involve tests around a single population mean or proportion. Later chapters introduce tests comparing two or more sample means or proportions. And again, in this chapter, you will use a ***z* test** or ***t* test**, depending on the size of the sample obtained. And, as in the preceding chapter, the conditions necessary for acceptable analysis of small samples are more rigid, especially requiring a relatively normally distributed population.

Problem Assignment (page numbers refer to location of the problems, and the **Notes** for the Topic 3a problem assignment above apply here as well):

1. Triola, Ch. 7, *Sect 7-2* (pp. 394-397): #3, 6, 10, 16, 17, 22, 27, 30, 35, 39; *Sect. 7-3* (pp. 411-416): #1-3, your choice of two (2) problems from 9-20, at least one of which should be even numbered, and use both the ***Traditional*** and ***P-value Method*** for each problem, without repeating the steps common to both (*i.e.*, you need only do the different parts of step 6 each way for problem); *Sect. 7-4* (pp. 422-427): #1, 4, 5, 8 (again, you need only do the different parts of step 6 each way for the pairs 1&5 and 4&8), 23, 30; *Sect. 7-5* (pp. 431-435): your choice of three (3) problems from #1-16, at least one of which should be even-numbered, again doing the different parts of step 6 each way for each problem, and after you have done the calculations and reached the conclusion, also use DDXL to check your result as shown on p. 431, also do problem #21.
2. Explain the difference between the two methods of hypothesis testing, indicate which you think is preferable, if either, and explain your choice. Then answer Moore, Ch 23, (p.463), #23.12, using the following as an example:
   Test $H_0 : p = 0.5$ versus $H_a : p \neq 0.5$ in each of the following cases, finding both the *P*-values and 95% confidence intervals using Excel if you wish
   (a) 1000 tosses give 505 heads ($\hat{p} = 0.505$).
   (b) 10,000 tosses give 5050 heads ($\hat{p} = 0.505$).
   (c) 100,000 tosses give 50,500 heads ($\hat{p} = 0.505$).

   Does that change your conclusion?
3. Answer the following questions from Moore's 4th Ed.:
   8.57 When asked to explain the meaning of "statistically significant at the a = 0.05 level," a student says, "This means there is only probability 0.05 that the null hypothesis is true." Is this an essentially correct explanation of statistical significance? Explain your answer.
   8.58 Another student, when asked why statistical significance appears so often in research reports, says, "Because saying that results are significant tells us that they cannot easily be explained by chance variation alone." Do you think that this statement is essentially correct? Explain your answer.

Self-evaluation exercise *You should set aside two hours in a continuous period for this – I would hope it would take less time than that*: When you're all done with the above and have it ready to send to me, relax for a day or so and then do the four (4) problems in *Cumulative Review Exercises* for Triola, Ch. 7. (pp. 447-448) as a midterm evaluation of your understanding of this material. Please do NOT look at the answers at the end of the book UNTIL after you have completed all four problems. Record and report the amount of time it took for each problem. Then compare your answers with the book's answers, and write a short paragraph of self-evaluation for your work in statistics so far. Be sure and save it for use as the end of the term.