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

Text: Triola, Mario F. (2001). *Elementary Statistics Using Excel*. Boston: Addison-Wesley Longman. [Triola]

Topic 4: More Tests of Hypotheses: Inferences from Two Samples; Multinomial Experiments and Contingency Tables

Introductory Comments: The two tools covered in this Topic are among the most widely used statistical procedures. The first, "Inferences from Two Samples," generally follows quite directly and quite simply from the hypothesis testing of a single sample that was the subject of the previous topic. For a single sample, you compared the mean with a single assumed mean; here you ask one of two questions, (1) "are the means of the populations represented by two different samples the same or different?" or (2) "has the population changed as the result of the treatment?" In the first case, the calculation of the **test statistic** is a bit more complicated than that for a single sample because you need to include the means and variances of both samples. In the second case, because you are often looking at matched pairs, the analysis simplifies to comparing the mean difference with zero. You carry out the analysis using the same **eight steps of hypothesis testing**, and again use either the ***z*-table** or ***t*-table** to determine your ***p*-value** and conclusion. (In the last Topic of this course, yet to come, you'll find out how to make comparisons of means when you have more than two samples.)

The second tool considered in this Topic involves making inferences from data gathered when there are a number of categories, for example colors of M&Ms, races of people or varieties of trees. Again we will consider two different types of questions, (1) "how does the distribution in the categories compare with an assumed distribution?" which applies when you're dealing with a sample from a single population – **a multinomial experiment** – often referred to as a **Chi-Square Test** by biologists, or (2) "are the distributions among the various categories the same or different in the different samples?" which involves analysis of **contingency tables**. These tests involve a new distribution, the **Chi-Square (c^2) Distribution**.

Topic 4a: Inferences from Two Samples

Reading:

Triola, Ch. 8, Sections 1-6 – Various topics on comparing two samples.

Probably the most important things to keep in mind as you work through this chapter are the **assumptions** that apply to each of the tests; these are usually stated early in each section. For comparisons of two means, it is important distinguish between the treatment of **large** and **small independent samples** (covered in *sections 2* and *6*, respectively) and the treatment of **independent samples** compared to **matched pairs** (**paired samples**) (covered in *section 3*). The actual calculation of the test statistic can be tedious, and I would strongly recommend that you NOT do it by hand or on a calculator (there is no educational value in doing so and mistakes are more likely), but use Excel, as described, for example, on p 461, item 1b. Set up the calculation for the test statistic using the appropriate formula and then either use the relevant table and/or the NORMSDIST function to determine the *p*-value. When you read the subsections on the rationale underlying the form of the test statistics, compare the form with that of the corresponding single sample statistic in Ch. 7.

In the preceding chapters, we have skipped over the sections on the analysis of the **variance**, because it added an additional complication, and, I believe, it is rarely used in the analysis of single populations. However, the nature of the variances is important in the analysis of **small independent samples** (covered in *section 6*), and thus you will need to deal with *Section 5*, "Comparing Variation in Two Samples." In this section, you are introduced to yet another test statistic, the ***F*-statistic** and the tables of the values of the respective ***F*-Distribution** (Table A-5). It gets complicated because you need to figure out both the **numerator degrees of freedom** and **denominator degrees of freedom** (do you remember which number in a fraction is the **numerator** and which is the **denominator**?) and then find the appropriate **critical value** of the *F*-statistic in the table. You can also get the *p*-value for the tail in Excel using the FDIST function. For the Coke vs. Pepsi example (pp. 490-492), the straightforward Excel calculation would be done as shown on the accompanying table. (See also the top example in the file **Ch 8&10 AltMeth.xls** sent with these instructions). Since the calculated *F* is smaller than F_{crit}, and the *p*-value is greater than a/2, you do not reject the null hypothesis.

	A	B	C	D
1				
2	n	36	36	
3	mean	0.81682	0.8241	
4	s	0.007507	0.0057	
5				
6	F	1.733927	=B4^2/C4^2	
7	p	0.054079	=FDIST(B6,B2-1,C2-1)	
8				
9	alpha	0.05		
10	F crit	1.961091	=FINV(B9/2,B2-1,C2-1)	

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

Note: Unless specifically instructed otherwise, use Excel to do the calculations for all problems identified with the Excel symbol in the margin, when a calculation is required, but do NOT try to generate an imaginary set of numbers as described in your text (Chapter 5 Project), just set up the equation and calculation. Please send the Excel file with these calculations when you submit your work (you can put those from the each section on one worksheet). In some problems, all you need to do is interpret the Excel display given in the text, in which case, do not waste time replicating the calculation, even though the data are available. ***Please state explicitly the two alternative hypotheses and appropriate steps of the hypothesis testing sequence for each problem testing a claim.***

1. *Sect. 8-2* (pp. 462-467): #1-4, 17, 18; *Sect. 8-3* (pp. 472-473): #1, 5, 8; *Sect. 8-4* (pp. 483-487): # 1, 5, 18; *Sect. 8-5* (pp. 493-496): #1, 2, 9, 10; *Sect. 8-6* (pp. 505-510): #1, 2, 5, 6; *Review Exercises* (pp. 511-513): # 5 (You're not told if these are matched pairs data, so test both ways – are the conclusions different?-Do NOT use DDXL, but you may use Excel's Data Analysis Tools.)

Topic 4b: Multinomial Experiments and Contingency Tables

Reading:

Triola, Ch. 10, Sections 1-3 – Multinomial Experiments and Contingency Tables

Probably the most important aspect of this chapter is dealing with data in a different format, **categorical data**, either as counts of a single variable with more than two options (*i.e.*, more than yes or no) – **multinomial experiments** or **one-way frequency tables** – or even more complicated data where the counts are determined by two variables – **contingency tables** or **two-way frequency tables**. As noted earlier, there is a new test statistic and corresponding distribution, **Chi-Square (χ^2)**, introduced in this chapter.

The **null hypothesis** for multinomial experiments takes the form of a set of **expected values** for the results, with the important condition that the total of all expected values is the same as the total of the sample values. For example, one might wonder of the distribution among the various options is constant or whether it meets the conditions of a given theory. The most common example in biology is in genetics, where one might test the actual counts from a mating experiment against the results expected according to Mendelian principles. (It's interesting that the text does not include a genetics problem other than asking if Mendel's results were too perfect – exercise 18, p. 604). You can easily use Excel's **CHITEST()** function for analysis of one-way frequency tables.

Depending on the nature of the data, the question and null hypothesis in the analysis of two-way frequency tables are different – one either does a **test of independence** or a **test of homogeneity**. The major difference is that the *in the experiment* the data are either from a single population (independence) or from several populations (homogeneity). However the calculation of the test statistic is identical in the two cases, and, in fact, is similar to that for the one-way frequency analysis in that the test statistic is χ^2 **= S (Observed value-Expected Value)^2/(Expected Value)**. There is a significant difference, however, between the calculation of expected values in the two methods. Also, while the text (p. 614-615) suggests a arranging the observed and expected values in single columns for two-way tables, I think the calculation is made easier by retaining the table structure and using Excel's *absolute* and *relative* addressing modes as shown below and more completely in the second example in the accompanying file, **Ch 8&10 AltMeth.xls**.

	Observed	Titanic Mortality (p. 589)				
		Men	Women	Boys	Girls	Totals
16	Survived	332	318	29	27	706
17	Died	1360	104	35	18	1517
18	Totals	1692	422	64	45	2223
Row	Column	B	C	D	E	F
	Expected	Values (row total x column total/overall total)				
		Men	Women	Boys	Girls	Totals
23	Survived	=$F16*B$18/$F$18	=$F16*C$18/$F$18	=$F16*D$18/$F$18	=$F16*E$18/$F$18	=SUM(B23:E23)
24	Died	=$F17*B$18/$F$18	=$F17*C$18/$F$18	=$F17*D$18/$F$18	=$F17*E$18/$F$18	=SUM(B24:E24)
25	Totals	=B23+B24	=C23+C24	=D23+D24	=E23+E24	=SUM(B25:E25)
		Note the column pointer to the row total is fixed ($F) and				
		the row pointer to the column total is fixed ($18).				

Problem Assignment (page numbers refer to location of the problems; see also the **NOTE** for Ch. 8 problems):

1. *Sect. 10-2* (pp. 601-605): #1, 9, 10, 18 (hint: see p. 787 and consider if df=5, what is $P(0<\chi^2<1)$?) ; *Sect. 10-3* (pp. 615-620): # 1-2, choose one (1) problem from 3-16 and use the method demonstrated above and in the accompanying file and indicate whether the test is for independence or homogeneity.
2. *Cumulative Review Exercises* (pp. 623-624): # 1-3, 5, 6. Treat this as another self-evaluation exercise. ☺