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]

Topic 5: Two Important Analysis Tools: Analysis of Variance (ANOVA) and Correlation & Regression

Introductory Comments: The two tools described here are again two of the most important tools used in statistical analysis, and for the purposes of this course, they complete our study of basic statistical analysis techniques.* You will study them in the reverse order that they are discussed in the text, primarily because **Analysis of Variance** (**ANOVA**) is a logical extension to multiple samples of the hypothesis testing of sample data encountered in one and two sample *z*- and *t*-tests. In contrast **Correlation** and **Regression** deal with the analysis of a different type of data: pairs of sample data, such as size of a restaurant bill and size of the tip that's left or time and distance for a mode of travel or various track events. In both cases, the actual calculations can get fairly complicated, so your primary focus should be on the basic principles behind the analysis, understanding the underlying assumptions and limits of the analysis, and interpretation of the results, leaving it to the computer to do the calculations. The analyses can often be done with the basic functions and/or "data analysis" add-in tools in Excel, without resort to the more complete picture (including the H_0 *vs.* H_a hypothesis test conclusion) produced by DDXL: in most cases, unless specified, I would prefer that you do not use DDXL except as a check on your own conclusion.

Topic 5a: Analysis of Variance (ANOVA)

Reading:

Triola, Ch. 11, Sections 1-3 [Note: the accompanying file, **Chs 11 & 9.xls**, contains the data for the examples in this chapter; you should use it to duplicate the text's computer displays and check your understanding. Please send me a copy of the file with your replication of the text's displays.]

One-Way ANOVA		
FACTOR		
Category A	Category B	Category C
Unit A-1	Unit B-1	Unit C-1
Unit A-2	Unit B-2	Unit C-2
.	.	.
.	.	.

		Two-Way ANOVA		
		FACTOR 1		
		Category A	Category B	Category C
FACTOR 2	Group 1	Unit A1-1	Unit B1-1	Unit C1-1
		Unit A1-2	Unit B1-2	Unit C1-2
		.	.	.
		.	.	.
	Group 2	Unit A2-1	Unit B2-1	Unit C2-1
		Unit A2-2	Unit B2-2	Unit C2-2
		.	.	.
		.	.	.
	Group 3	Unit A3-1	Unit B3-1	Unit C3-1
		Unit A3-2	Unit B3-2	Unit C3-2
		.	.	.
		.	.	.

The one and two sample *z*- and *t*-tests we studied in chapters 7 and 8 appeared to focus on the comparison of means, either against a fixed hypothesized value or against each other. However, if you review the basics of the analysis carefully, the tests actually can be viewed as using the size of the variation or standard deviation in the sample data to determine if there is (a) sufficient overlap of the tail (determined by the desired **level of confidence**) of an hypothesized normal distribution attributed to the sample with the hypothesized population mean in the case of a single sample or (b) overlap of the two tails in the case of comparison of samples from two separate populations. So, in a sense, those tests can be characterized as analysis based on the variance in the sample data. The analysis tool formally called **analysis of variance** (or **ANOVA** for short), extends those techniques to more than two samples but uses the variances directly in its analysis. The **test statistic** of concern here is the ***F* distribution**, which you met in Triola's *Section 8-5*.

As noted above, the calculations get fairly complicated, and the chapter also describes two types of ANOVA (see table): **one-way ANOVA** – in which different populations are divided into categories based on a single **factor** or **treatment** (size of car, type of injury, M&M color, or gender) – and **two-way ANOVA** (often referred to as **block** design) – in which there are two

* There is a whole realm of statistics that we will not cover, **Nonparametric Statistics**, which comprises tools that do not require assumptions about the nature or shape of the data. They are the only tools that can be used with nominal and ordinal (ranked) data, such as one often encounters in opinion questionnaires: "rate your impression of the actress's performance on a scale of 1-10." Some of the basic nonparametric tests are described in Triola's Chapter 13.

different factors or treatments (divided into "categories" of one factor and "groups" of the other) and a set distinguishing each separate sample – gender ("category") *and* M&M color ("group") or type of injury ("category") *and* size of car ("group")). If there are c "categories" for one factor and g "groups" for the other, then the total number of different treatments or **blocks** is $n = c \cdot g$. The analysis can be based on observation or a designed experiment, *but it is critical that you examine and understand the assumptions for each type of analysis and be sure the data fit the assumed conditions for the statistical test – this is the responsibility of the person who designs the survey and survey methods or experiment, as well as the statistician.*

In one-way ANOVA (*Triola, Sect. 11-2*), the null hypothesis, H_0 is that $\mu_1 = \mu_2 = \mu_3 = \ldots$, and the alternative hypothesis, H_a, is that the means are *not* all equal or that *at least two of the means differ* (or *at least one* of the means *differs* from *at least one* of the others. **Note that your text does not explicitly state H_a, but states conclusions as "reject the null hypothesis" or "fail to reject the null hypothesis."** The basic analysis described in the text does not enable you to analyze which of the means are not equal to the other(s), but there are some advanced techniques (**multiple comparison procedures**) that can provide such measures, or at least indications toward them.

In two-way ANOVA (*Triola, Sect. 11-3*), the analysis gets quite a bit more complicated, and allows one to determine (a) if there is **interaction** between the two factors (H_0: there is no interaction), (b) if there is an effect (difference) due to one factor independent of the other factor, and (c) if there is an effect due to the other factor, independent of the first. Note that if it determined that there is interaction (test (a) – rejection of H_0), analysis of each factor independently (tests (b) and (c)) cannot be carried out. This is an important point: if there is interaction, analysis stops (Figure 11-3, p. 653).

Excel's Analysis Tools are best for doing ANOVA calculations. They do the calculations and present the values (F, p-value, and F-critical) you need to make the selection between H_0 and H_a. Particularly for two-way ANOVA, even more than indicated by your text, DDXL presents problems in ANOVA analysis.

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

1. *Sect. 11-2* (pp 644-648): #1, 2, 3 (do the simple calculation as shown on pp. 638-9 using Excel to do the arithmetic, then check using Excel's Data Analysis tool as shown on pp. 643-4), 8, 16 (tough but straightforward); *Sect. 11-3* (pp. 656-657): #1-3, 7, 8 (note that these do not require calculations, but they do require you to state explicitly the 8 steps in the test of hypothesis).
2. *Additional non-calculation question*: why is it inappropriate to do a two-way ANOVA analysis on the crash data in Table 11-1 (p. 632)?

Topic 5b: Correlation and Simple Regression

Reading:

Triola, Ch. 9, Sections 1-4 only: Correlation and Regression [Note: the data for the tipping calculations, as well as an expanded version of Table A-6 (Critical Values of the Pearson Correlation Coefficient) are also in the accompanying file, **Chs 11 & 9.xls**. You should carry out the example calculations using the Excel functions and tools described in the text to increase your understanding of the results.]

Correlation and regression deal with sets of pairs of numbers which are almost always **continuous numerical interval or ratio data** (see Section 1-2 for a reminder of this terminology). The main question in correlation analysis is whether the data pairs represent variables that are related – *i.e.*, the behavior of one of the variables corresponding to one of the members of each data pair is somehow related to the behavior of the other. The best description of such relationships is provided by the various scatterplots in Figure 9-1 (p. 524). The significant statistic is the **correlation coefficient** (or **Pearson product moment correlation coefficient**). The actual calculation gets pretty tedious, so the best thing to do is use Excel's **CORREL()** function or the equivalent **PEARSON()** function (which your text does not mention). (The more elaborate Correlation Data Analysis Tool provides minimal additional information and is not worth the effort.) Important concepts in Section 9-2 are the interpretation of the correlation coefficient (actually through the value of its square, r^2), the formal hypothesis test of whether r is significantly different from zero, – *i.e.* there is significant correlation – and the caution that *correlation does NOT mean causality*. Note also the notation that the correlation coefficient, $\boldsymbol{r}$, for the sample data is an estimate of the correlation coefficient, r, for the population that is sampled.

The correlation coefficient is essentially a descriptive statistic; to determine its statistical significance, you need to do the hypothesis test, comparing the r-value (or corresponding t-value) to a statistically determined critical value, usually based on the null hypothesis, H_0: $r = 0$, *i.e.*, no correlation. In doing hypothesis tests for correlation, you should know that the simple test using r_{crit} values (Table A-6, p. 794) is relatively unique to this text, *i.e.*, most texts do not do the conversion from t_{crit}, but expect you to calculate t using the formula on p. 530. I have produced an expanded version of Table A-6 in **Chs 11 & 9.xls**, along with a formula to

calculate the *p*-value from *r* and *n*. This is shown in columns **G-I** on the **Correlation table** worksheet in the Excel file, including a calculation of the *p*-value for the tipping example. This analysis is based on the approach described in problem 9-2:23 (p. 540); basically Table A-6 is a converted *t*-table. Note also that problem 9-2:24 shows how to construct confidence intervals for *r*, a quite complex procedure.

Correlation provides primarily a numerical analysis of the relationship between the two variables in the data sets. Simple linear regression carries the analysis one step further, and generates the formula for the best fit graph line – the **regression line** – and the corresponding linear **regression equation** of the classis form, $\boldsymbol{y = mx + b}$, where *m* is the **slope** of the line and *b* is the ***y*-intercept** (the value of *y* when *x* is zero). Again note the corresponding notation for the sample and population parameters for the slope and intercept in the box on the bottom of p. 541. One of the critical things to be concerned about in regression analysis is related to the causality question: which variable (nominally the **independent variable** – or the one specified by the experimenter) determines the value of the other variable (the **dependent variable** – the one measured by the experimenter in response to the independent variable). The regression line is determined by a "**least-squares**" analysis, that is, by determining the equation that provides the smallest sum of the squares of the distance from each point to the line given by the equation; the distances are the **residuals**, as described in subsection beginning on p. 548 and illustrated in Figure 9-8 (p. 549). A valuable extension of regression analysis is the ability to make predictions using the regression equation, predict the value of the dependent variable for a specific value of the independent variable, a value that may or may not be one of those in the original data. It is generally OK to do this for values within the range of the original data (often described as **interpolation**), but it gets more and more unreasonable and unreliable when the calculation goes outside the range of original data (**extrapolation**).

Basic least squares regression analysis provides the best fit line, but as in the basic correlation coefficient calculation, it is primarily descriptive in nature. The inferential aspect again comes in with the hypothesis testing, usually in the form of both (a) H_0: the slope, *m*, is 0 and (b) H_0: the *y*-intercept, *b*, is 0, although there are instances where one might be interested in knowing whether the slope and/or intercept are statistically different from specified values other than zero. In the tipping example in the text, for example, the slope of the line is statistically different from zero ($p = 0.0418$), but the intercept is not ($p = 0.9339$) (see the box on the top of p. 550). In addition to the hypothesis testing of the of the slope and intercept, one can not only determine their confidence limits, but also the confidence limits of the intermediate predicted (calculated) values. This analysis is described in Section 9-4, and can get pretty elaborate with relatively tedious calculations. Fortunately, Excel provides some of the basic numbers needed for the calculation as shown on pp. 557ff. Again, in making predictions, the critical issue is determining the most probable range within which the actual (population) value occurs.

Note that we are skipping the last two sections of Chapter 9: 9-5 Multiple Regression and 9-6 Model Building. Multiple regression analyzes data that have multiple independent variables; model building examines situations where the relationship between the variables is not necessarily linear, *e.g.*, polynomial – as in $y = ax^2 + bx + c$ (**note the typo in the polynomial expression on p. 574**: the first term should be $\boldsymbol{ax^2}$, <u>not</u> a^2).

<u>Problem Assignment</u> (page numbers refer to location of the problems):

Sect. 9-2 (pp 536-541): #1-3 (work through the steps of the hypothesis test explicitly, use Table A-6 to determine r_{crit}, and state the conclusion; then check your conclusion by using the formula on the **Correlation table** worksheet in **Chs 11 & 9.xls** – just enter the data in the first empty row in columns **G** and **H**, and then drag the formula down in column **I**), 6 (you'll need to enter the data and let Excel do the calculations for #3 and 6), 17-20; *Sect. 9-3* (pp. 551-554) 3, 6 (<u>do not reenter the data from the corresponding problems in Sect 9-2</u>); *Sect. 9-4* (pp. 562-565): #1-7 (it looks like a lot, but there are minimal calculations required; for 5-7, you mainly need to interpret the DDXL results in the graphic in the middle of page 562), 8 (to save the elaborate calculation, assume the 95% prediction interval is ±68.435; your only task is to write the requested statement).

RAC 02/08/04 (Rev 2U); 02/19/05 (Rev U5)