Errata for *Doing Statistics with SPSS*

Howard Seltman, Fall 2006

Note: Please do not get discouraged by these errata. Most of the book is well written and correct.

Chapter 2, Descriptive Statistics

- page 11: The formula for range, which subtracts the minimum data value from the maximum, then adds one is non-standard. Most statisticians do not add one.

- page 18: The statement "...with a skewed distribution, the median will provide a more appropriate indication of the typical score." is too strong. Depending on the use of the "typical score", the median *may* be more appropriate that the mean.

  The statement that "Outliers are scores in the distribution that are more than 1.5 box lengths from the 25th or 75th percentile..." is misleading. This applies only to the conventional picture of a boxplot. A more general discussion of what it means to be an outlier is much more complex.

  SPSS defines "extreme values" as those points that are more than 3 (not 4) IQRs from the 25th or 75th percentiles.

- page 19: The sentence "The choice of which measure of central tendency to employ is not restricted..." should read "The information used to make the choice of which measure of central tendency to employ is not restricted...".

- page 20: The term "categorical" is an important one which includes both nominal and ordinal scales and variables. The term "quantitative" is the opposite of categorical; it includes both continuous (also known as "interval") variables and discrete variables. Discrete variables are quantitative because the values for two cases can be meaningfully added or subtracted, but they differ from continuous variables in that they take on only whole number values, and fractions are meaningless as data values. Certain statistical procedures are specifically designed for discrete variables.

- page 21: You can ignore the statement "...sometimes the value 1 is added" for the range formula.

The statement “Unfortunately, if the distribution is symmetrical, there will be as many scores greater than the mean as there are scores smaller than the mean” is somewhat misleading. We know this: 1) if a distribution is (perfectly) symmetrical, then its mean and median are equal, and 2) any distribution has as many values above as below the median, by definition. More importantly, **you should know the following fact**: for *any* set of values, regardless of the shape of the distribution, the sum of the deviations from the mean is zero.

- page 22: The definition of variance with “n” rather than “n-1” in the denominator does not apply to *samples* of data from a larger population, which is almost always what we have. (see next comment)

- page 24: The explanation of why we use “n-1” in the denominator of the variance and standard deviation for a sample is **flat out wrong!** It has nothing to do with the possibility that “an error could have been made” or that we want a “conservative estimate”. Instead, we know that many different samples of the same size as our sample could have been drawn from the population, and each would have a different (rarely the same) variance. If we use “n-1” in the denominators, the average of all of these estimates of the variance will equal the true population variance. If we use “n”, the average of the estimates will be smaller than the true population variance. So, in order to achieve “unbiasedness” of our estimates, we use “n-1”. (The reason for this fact is rather technical.)

  The middle formula on this page is not useful to us because we will use a computer to calculate variances.

Chapter 3, The Normal Distribution

- page 29: The phrase “will produce what is known as a normal distribution” should be changed to “tends to produce what is known as a normal distribution”. This statement is based on the “central limit theorem” which has some technical conditions that must be met for it to be true, and these are not always met.

  The formula on this page is **garbage**. The left hand side should be “the density of Y”, not Y, and the right hand side should have “1” not “n”.

- page 30: The book gives the impression that the normal distribution is special because mean +/- 1 (or 2) standard deviations encompasses a fixed percentage of the data. This is actually true of *all* distributions. Of course the percentages inside those limits differ from distribution to distribution.

Chapter 4, Intro to Experimental Research

- page 43: The statement “this analysis measures the precise probability of getting the observed differences in the dependent variable under the various levels of the independent variable by chance” is correct only if “getting the observed differences” is replaced by “getting differences as large as or larger than the observed differences”. I hope that you can see that the chance of getting exactly the observed differences is always very small. This also applies to the paragraph spanning pages 43 and 44.

- page 45: The first two sentences under “Type I and II Errors” is imprecise. Pay attention to how we talk about the chance of making a mistake in class. Particularly, the phrase “when the null hypothesis is true” is missing in several places throughout the chapter when talking about chances of making a mistake.

- page 45, figure 4.1: This diagram is quite misleading. In particular, it reinforces the *totally incorrect* idea that setting the significance level (alpha) to 5% gives 95% power. The important correct statement is at the top of page 46: “...any attempt to reduce the probability of committing a Type I error by reducing the significance level will result in an increase in the possibility of committing a Type II error.”

- page 49: It is not true that ”parametric tests” *always* “assume ... a normally distributed population”. Rather, they always assume a particular distribution for a population, and this distribution is often assumed to be normal. Similarly, equal variance is not a requirement for a parametric test, and an interval (quantitative) scale is not required. Instead, for a test to be “parametric” the variance relationship and the scale just need to be specified.

- Table 4.1: It is *not* true that you *must* use non-parametric tests for ordinal and nominal data, nor that you *cannot* use them for quantitative data.

- page 50: In the last paragraph, remember that “error” in this context means variation, not mistakes.

- page 54: In the summary, again change ”obtaining the observed difference” to “obtaining differences as large as or larger than the observed differences”. Again, the statement that “... there is a 5% probability that the wrong decision has been made” is imprecise–see class notes.

Chapter 5, Sampling Error

- page 58: In the first sentence of paragraph 2, we do *not* assume that the sample mean is the same as the population mean.

- page 62: The formula for t at the bottom of the page is wrong. There should be “n” not “n-1” in the denominator. (The “n-1” is already in the s^2.)

- page 63: The variance and t calculations at the bottom of the page give the right t value for the wrong reasons. The variance of the four differences is 2, not 1.5, and the standard error (denominator of t) should be calculated as $\sqrt{s^2/n} = \sqrt{2/4} = 0.707$.

- page 64: (Interpreting the Value of t) Replace “observed value of t” with “observed or more extreme value of t”.

- page 65, last paragraph: After “same experiment were conducted in the future”, add “and the null hypothesis is really true”, and change “observed difference” to “observed or more extreme difference”. Continuing on the next page, “only a 5% probability that a mistake is being made” is true only when “the null hypothesis is really true”.

- page 67: It is *completely wrong* to interpret p-values as “the probability that the null hypothesis is correct.” Also, clearly the chance that the exact observed mean difference would be obtained in future sampling is not equal to the p-value.

- page 70: You don’t need to know the formulas, but the correct formula for the independent t-test is

$$t_{(n_1-1)+(n_2-1)} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Chapter 6, Between Groups 1-Way ANOVA

- page 81: I would call #4 “Treatment”, not “Treatment Variability” because I reserve the latter for describing the added variability caused by not applying a treatment uniformly to all subjects within a given treatment group.
- page 82: Although the mnemonic F=$\frac{\text{error+treatment}}{\text{error}}$ is a good one, the statement that “Total variance = between-groups variance + within-groups variance” is wrong. As you can see from close examination of the ANOVA table for any one-way analysis of variance, both the SS and df values are additive (total=between+within), but the MS values (which are the variance estimates) are not additive. The statement “Total variance = between-groups variance + within-groups variance” is true only if we replace “between-groups variance” with “treatment variance” (obtained by subtracting out the error variance from $MS_{between}$).
- page 83: In the between-groups SS formula, the book uses $\bar{x}$ where I used $\bar{X}_i$ for the group means, and it uses $\bar{X}$ where I used $\bar{\bar{X}}$ for the grand mean.
- page 91: The cutoff for F_{max} in the last paragraph is incorrect. It should be 3, not 9.

Chapter 7, Analytic Comparisons

- page 102-103: This section on A Priori Tests is excellent! It includes the correct technical details for planned comparisons that 1) they are chosen before running the experiment, 2) there are no more than k-1 of them, and 3) they are independent of each other.
- page 104, after equation 7.1: The word “product” is not intended to imply multiplication here–it just means “result”.

Chapter 8, Within Group 1-Way ANOVA

- page 117-118: I suggest ignoring the stuff with square brackets. This relates to easier manual calculation, but we will be using the computer. SS_{subjects} is simply the sum of squared deviations of individual subject means from each other.

- page 120: The statement about within subjects ANOVA being “more conservative” could easily be misinterpreted. It applies to comparing the incorrect to the correct analysis for a within subjects design. A within subjects design is more powerful whenever the subject-to-subject variation is not small. And if you set up an experiment with a within-subjects design, you MUST analyze it with a within-subjects analysis.

- p 121: Be aware that Mauchly’s test of sphericity is very sensitive to violations of the normality assumption. So a small p-value for the sphericity test could be due to a bit of non-normality rather than a variance/covariance problem. And in that case, the F test is robust, and the correction would result in loss of power. Also, with small sample size, assumption tests in general don’t have enough power and therefore almost always give non-significant p-values even if assumptions are violated.

- p. 127: I don’t think there are many situations where the “Helmert” contrasts would be of interest.

Chapter 9, Factorial ANOVA

- p. 132: Note that there are many ways to divide sources of variability for different purposes. The three sources given here are not in conflict with the four we learned in class.

- p. 134: The phrase “both the main effects and the interaction influence fitness” is misleading. In reality, if there is an interaction, then changing levels of both factor A and factor B have some effects on the outcome, and for at least one level difference for A, the size of the effect on the outcome differs depending on the level of factor B (and vice versa).

- p. 135: The criterion of kurtosis (or skewness) divided by its standard error exceeding 1.96 is approximately equivalent to checking if the kurtosis exceeds twice the standard error.

- p. 138: The line “Total” in table 9.7 is not included in most program’s output. In fact, the quantity labeled “corrected total” is usually labeled “Total”.

Chapter 11, Linear Regression

- p. 163: It is a common mistake to claim that "linear" regression is only used to find a "linear" relationship between two variables. Actually, the "linear" in linear regression is a technical term indicating that the parameters of the mean model (coefficients) are in their simple forms (not to powers, logs, etc.). There is no such restriction on the dependent variable or the independent variables, so it is quite easy to model a non-linear relationship with linear regression.

- p. 164, last paragraph: The "regression" line is not "equidistant" from all of the points. Actually, it is the chosen as the line that minimizes the sum of squares of the vertical distances from the points to the lines.

- p. 166: The notation of adding the subscript "y" to "a" and "b" is non-standard. Usually we just use a and b (or b and m, respectively, or b_0 and b_1).

- p. 166, bottom: We will use the more standard notation $\mathrm{E}(Y|X)$, read "the expected value of Y for a given X value", rather then Y'.

- p. 169, top: The phrase "where the extrapolated line cuts through the Y-axis" hints at the problem with the preceding sentence. It is *dangerous* to extrapolate! Unless there are data points with X values on both sides of zero, it is better to say that the intercept has no substantive meaning for a particular problem. If the X values cross zero, then it is correct to say that the intercept is the expected (or predicted) value of Y when X is zero.

- p. 169, middle: "This means that if someone's self confidence score is 3, their predicted performance score would be 4.15." This is better thought of like this: Our model states that if we happen to observe a large number of subjects all with self confidence scored 3, we can expect that the distribution of their performance scores will be N(4.15,σ^2), i.e., normally distributed with mean 4.15 and variance equal to the common variance (the same for all X values).

- p. 170, top: Note that it is only a coincidence that the slope and correlation coefficient have the same value in this example.

- p. 171, bottom: Multiplying the square of r by 100 to obtain a percentage is optional. It is equivalent to look at the coefficient of determination (R^2 or r^2) as falling between 0 and 1 or between 0 and 100%.

- p. 172: The adjusted r square (adjusted R^2) is better thought of as a very useful adjustment needed when there is more than one explanatory variable. In that case it adjusts for the fact that adding any variable (useful for explaining the outcome or not) will increase R^2 at least a small amount, while the "adjusted R^2" will not suffer from this misleading property.

- p. 172, bottom: Ignore the alternate formula for calculating r: we will let the computer do the calculations.

Chapter 12, Multiple Regression

- p. 179, para. 4: The usefulness of the log transformation does not mean that it is always the best transformation. Other transformations such as reciprocal, square and square root are commonly used.

- p. 181: Again, the book's notations is non-standard. We will use b_0 instead of a_y (for the intercept or constant).

- p. 187: The last word, "This", should be "When the null hypothesis is true, this".

- pp. 188-191: Here is a comment about the example. The EDA scatter plot shows a clear, strong relationship between Trees and Attractiveness, but Trees was dropped from the model. It would have been useful to look at a scatter plot of pairs of independent (explanatory) variables. Presumably, Trees was highly correlated with at least one of the other explanatory variables. We should be careful not to "believe" the final model, other than as an aid for predicting Attractiveness for a new similar landscape. When two explanatory variables are correlated, and stepwise regression is used, a repeat experiment may easily result in a different set of explanatory variables. This problem is NOT seen with designed experiments when the explanatory variables (at least those under the experimenter's control) are not correlated.

- p. 192: The “shared variance” concept is not as pertinent as “percent of variance explained”. Think of $R^2 = 0.49$ as telling us that 49% of the total variability in the outcome is explained by regressing on the combination of explanatory variables.

- p. 192, “Standard Error”: This paragraph is talking about “residual standard error” (which is also the square root of MSE, or mean squared error”). Without the word “residual” this is ambiguous because there are other standard errors calculated in regression.