Journal of
Education and Human Development ISSN 1934-7200
Volume 2, Issue 2, 2008

A Statistical Decision Tree for the Helping Professions

Van Wiesner, Assistant Professor, Department of Educational Leadership and Counseling, Sam Houston State University, wiesner@shsu.edu
Daniel G. Eckstein, Associate Professor, Department of Educational Leadership and Counseling, Sam Houston State University, d.eckstein@yahoo.com
Chi-Sing Li, Assistant Professor, Department of Educational Leadership and Counseling, Sam Houston State University, chisingli@shsu.edu
Raymond Zepp, Professor, Pannasastra University of Cambodia, zepp@rocketmail.com

Abstract

A decision tree is presented for assisting students and practitioners in the helping professions in determining the best-fit statistical method to use in research. By following a series of focused questions, the user will select the most appropriate statistical procedure through this process. Examples of various procedures are drawn from current literature and comments are offered on the use of the statistics. Specific functions within Microsoft EXCEL are also identified. Three supplemental Internet statistical decision making tree resources will also be cited.

Introduction

In the classic world of job categories, occupations are classified into those that deal with data, people, or things. Many in the counseling field find it challenging to navigate the “sea of statistics.” We offer a compass to our fellow sailors.

Warmth, genuineness, and empathy typically are not required for great research. Genuinely speaking, our collective warmth for math is lacking and we seek to increase our empathy of statistics. The primary purpose of the current article is to help bridge the gap often existing between counselors and the data they wish to understand or to acquire. Assisting counseling professionals and students in making better use of statistical procedures in their own research is a major objective of the article.

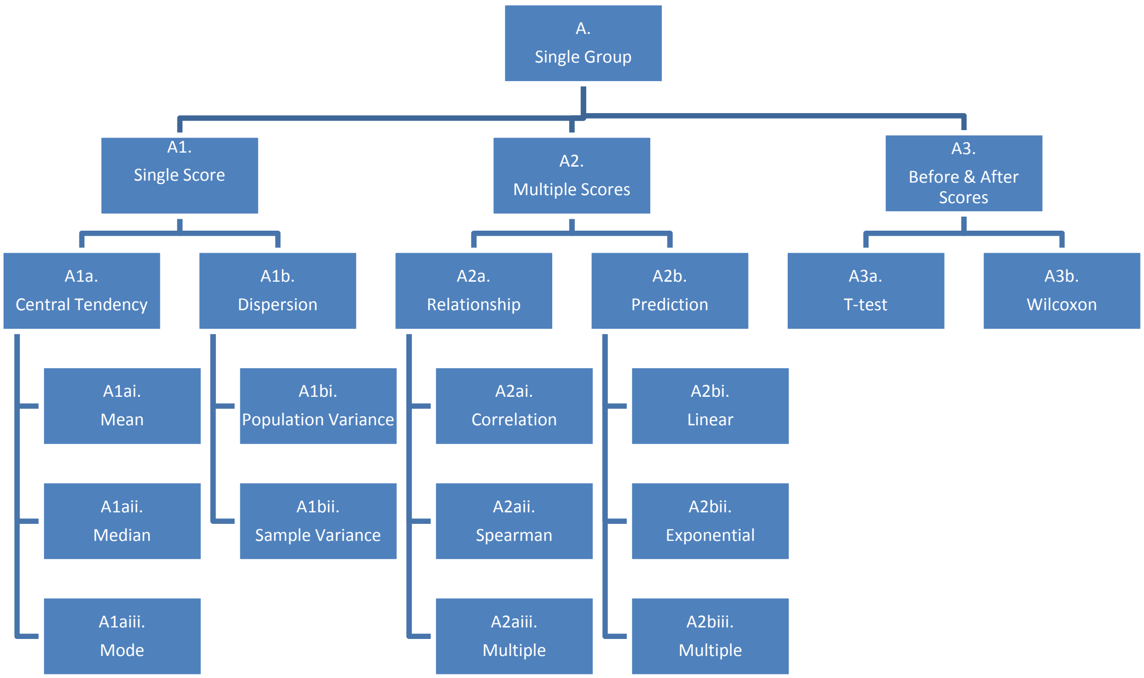
A statistical decision tree with an emphasis on verbal explanation in lieu of mathematical formulae is presented. The main benefit of decision trees is that they provide a series of options based upon answers to fundamental questions. A number of questions that may be considered in approaching research accompany the decision tree presented. Depending on the answer to the questions, the statistical decision tree may assist in the identification of the best-fit statistical procedure(s) for the particular research project. It is meant to be a basic and condensed visual summary for helping the counseling researcher determine the right tool for the right job.

The concept is similar to that of the multi-axial diagnostic decision making trees featured in the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000). Commands are also provided for use with the computer program Excel (Microsoft Corp., 2008), a widely available and commonly used tool for basic statistical research.

The Identification Tree

Just as a tree has a central trunk from which many branches are attached, core questions for determining obvious branches of the research question can be used by researchers. It is often helpful to step back and look at the forest before beginning research. As Keller (2006) noted, statistics really begins with a question rather than with data. The data must be good and needs to come from relevant measures in order for important questions to be answered well. Figures 1and 2 demonstrate the statistical decision tree designed by the authors in the following graphical fashion.

Figure 1. Graphical decision tree for one group.



What follows are a series of questions; the answers to the "if/then" question lead the researcher in identifying different tools ("branches") that are the best fit in examining the core research question(s).

Question I: How many scores or measures does each member of the test group have?

A. For research with only a single group of subjects,

1. If each member of the group has only one score or measure (e.g., a test score)

a. and the investigator wants to describe a central tendency or average of the group

i. In most cases, calculate the *mean* (Excel: *average*) (Microsoft Corp., 2008).

ii. If the group is skewed, with many more subjects on one end than the other (e.g., incomes, ages), calculate the *median* (Excel: *median*) (Microsoft Corp., 2008).

iii. If the investigator wants the most commonly occurring number (e.g., family size), calculate the *mode* (Excel: *mode*) (Microsoft Corp., 2008). The mode is the only measure of central tendency that should be used for most nominal (categorical) data. *Note*: The three measures of central tendency shown here are not the only ones. One may also find the *harmonic mean*, which is used to find average speeds or other rates of change, and the *geometric mean*, used to find the average of price indices or inflation rates. However, the mean (more specifically called the arithmetic mean) is by far the most widely used.

b. If the investigator wants to describe the spread-out-ness or dispersion of the group

i. To describe the group by itself, calculate either the *population variance* [denoted σ^2 (Excel: *varp*)] (Microsoft Corp., 2008) or the *population standard deviation* [denoted σ (Excel: *stdevp*)] (Microsoft Corp., 2008). (This includes those in which the formula divides by n.)

ii. If using the group as a sample to estimate a larger population, calculate the *sample variance* [denoted s^2 (Excel: *var*)] or the *sample standard deviation* [denoted *s* (Excel: *stdev*)] (Microsoft Corp., 2008). The standard deviation is simply the square root of the variance. The standard deviation is preferred for reporting because it has the same unit of measurement as the scores under consideration. The variance, which has squared units, allows easier calculation of the average variance of several group variances.

One also finds other measures of dispersion in use. Very simple measures include the *range* (difference between highest and lowest scores), the *inter-quartile range* (used to avoid extreme scores), and the *mean deviation* (similar to the standard deviation, simpler to use, but not as useful in statistical analysis). The sample standard deviation is almost the same as the population standard deviation, especially for large sample sizes, so for samples above 50 it makes no difference whether *s* or σ is used.

2. Is there more than one score?

If each member of the group has more than one score (e.g., a battery of test scores)

a. And the investigator wants to determine if the sets of scores are related, various forms of correlations can be calculated. Correlations measure relationships, but do not imply cause and effect. One cannot conclude from correlation that one thing *causes* another. For example, there is a high correlation between unresolved anger and people diagnosed with cancer. That does not mean unresolved anger causes cancer or visa verse. The false conclusion that one variable causes the other is perhaps the most common mistake in all statistics. Furthermore, correlations are always numbers between -1 and 1. A negative correlation such as -0.8 indicates an inverse relationship: high scores for one variable relate to low scores on the other, such as lower temperatures at higher elevations.

i. In most cases if *any two* sets of scores are related (e.g., are verbal intelligence test scores related to quantitative scores), then calculate the *correlation coefficient r* (a parametric correlation formally called the Pearson product-moment) (Excel: *correl*) (Microsoft Corp., 2008).

ii. If the scores are rankings or preference ratings, use *Spearman's rank correlation rho* or ρ.

iii. If there are relationships among three or more sets of scores, then use the *multiple correlation coefficient, R*. Correlations measure relationship, but it may not be a *causal* relationship. One cannot conclude from correlation that one thing *causes* another. The false conclusion that one variable causes the other is perhaps the most common mistake in all statistics. Furthermore, correlations are always numbers between -1 and 1. A negative correlation such as -0.8 indicates an inverse relationship: high scores for one variable relate to low scores on the other, such as lower temperatures at higher elevations.

b. If the investigator wants to predict a value from other known variables (e.g., predicting a subject's performance based on intelligence test scores), then a regression equation can be utilized.

i. Most cases involve predicting a value termed the dependent variable from one known score termed the independent or predictor variable. In this case use *linear regression* (Excel: *linest*) if the same case as above is true (Microsoft Corp., 2008).

ii. Exponential regression is used if the independent variable shows exponential growth, such as population or income growth.(Excel: *growth*) (Microsoft Corp., 2008).
iii. Multiple regression is used for predicting a value from more than one known predictor variable. For example, you want to predict how well prospective students applying to a graduate program will do in a graduate program. One way to measure the level of success of students in a graduate program would be to examine their grade on a comprehensive examination that takes place toward the end of the masters program. Another method might be to look at their graduate grade point average upon program completion. Yet another way to measure success might be to take the average score of a student assessment instrument based on professors who taught the various students.

Once you determine how you will measure success, you must also consider what data is available, logical, and reasonably fair for predicting success. You might choose to consider such variables as: the student's undergraduate grade point average (UGPA), a (quantitative) score assigned to their letters of references (REF), a score assigned to a screening interview (INT), their verbal graduate record exam score (VGRE), and/or their quantitative graduate record exam score (QGRE).

Suppose you chose to measure success as the student's comprehensive exam score (COMP). You could look at the data available for students who have already finished the program and allow linear multiple regression to determine the best way to weigh various factors in predicting COMP scores. For example the formula:
COMP = A*UGPA + B*REF + C*INT + D*VGRE + E*QGRE + constant

This would produce a predicted COMP score for each student using five pieces of data. The values of the regression weights (A, B, C, D, & E) and of the constant would be determined by a computer program such as SPSS and these weights would be values that produced predicted values of COMP that came closest to the actual COMP scores. The correlation between these predicted scores and the actual scores is R. The higher R is the better the formula predicts actual scores for the available data.

One could use the formula to predict the score of potential students on their comprehensive examination should be admitted into the program. The problem of shrinkage will plague the accuracy of this equation to predict accurately for future sets of students. One could equate two more equations based on the other two measures of success so that three predicted values are obtained. Those deciding on admission could they use the results of three multiple regression equations to help (not substitute) them in the selection process.

For multiple correlation, researchers often report R^2 instead of R. If R is reported, it is always positive, since a negative multiple correlation has no meaning. Regression is related to correlation. In making a good prediction, a strong relationship must exist. Thus, one needs a high correlation for making a good prediction. Another way of stating this is that if the correlation is low, the prediction will not be of much use. In multiple regression, only one variable depends on several independent variables. An important assumption is that the independent variables must be truly independent. A high correlation between supposedly independent variables can lead to serious problems with the results.

If pre- and post-test scores are being researched

1. In comparing the *before* and *after* scores of a sample

    a. If the data is distributed according to a normal distribution, calculate all the before-after differences and perform a paired t-test using paired observations.

    b. If data is not normally distributed, rank the scores and perform a Wilcoxon t-test. A common mistake is to treat the *before* scores as one group and the *after* scores as a second group and then compare groups using an independent-test. This incorrect method treats the two sets of scores as independent and loses the valuable fact that two scores belong to the same person.

B. When the research compares two or more groups

1. If the data are frequencies, that is, if they describe how many times something occurs, such as how many people in each group answer *yes* and how many answer *no,* use the *chi-square (χ^2) test*. The chi-square test is one of the easiest and most common methods for comparing groups. No assumptions about normal distribution are made since the data are only frequencies. In computer statistical packages like SPSS, chi-square usually comes after cross-tabulations or cross-tabs, from a table of the frequencies from which the chi-square test can be performed.

2. If the data are scores are within a normal distribution, that is, if most scores cluster around the average, with only a few very high or very low scores (e.g., SAT results have most scores around 500, with a few over 700 or below 300).

    a. If the investigator is asking about differences between two groups, then an independent t-test is the best-fit. (Excel: ttest) (Microsoft Corp., 2008). When one has two groups, Analysis of Variance (ANOVA) and t-test will yield the same results.

    b. If the investigator is asking about differences among three or more groups, then ANOVA should be used. ANOVA is preferred over multiple t-test because the risk of making a Type I error is compounded by using many t-tests. A Type I error is when a true null hypothesis is rejected. A Type II error is when one fails to reject a false null hypothesis. Because this is not easily done with Excel, a better suggestion is to use SPSS for analysis of variance.

        i. If differences are found, and the investigator wants to analyze which groups are significantly different (from a statistical standpoint) from others, use a *post hoc* (also called *a posteriori*) test such as Tukey's HSD or Scheffe ("Scheffe's method," 2008).

        ii. If no statistically significant differences are found, then the investigator reports that no differences were found at the given alpha level. Researchers should not assert that there are no differences.

3. If the data are scores with a non-normal distribution, such as large numbers of scores at the high end and/or the low end, first test whether data are normally distributed. Perform a *goodness of fit* test. Divide the scores into interval groups and predict how many scores should fall into each group according to the normal distribution. Test the frequencies of these expected scores against the observed scores using a chi-square test.

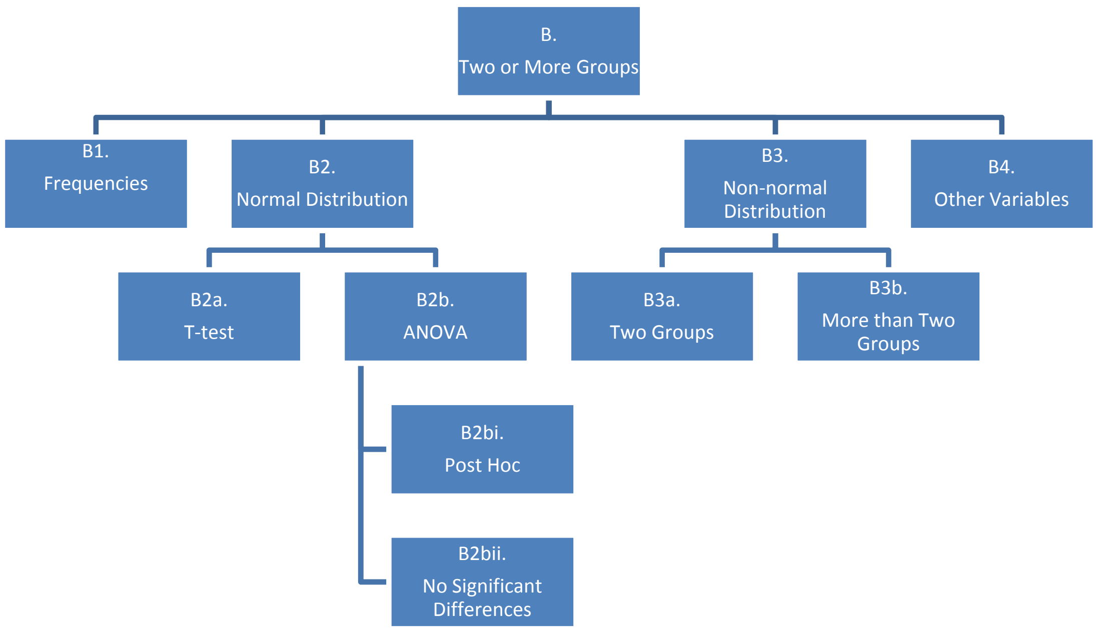
    a. For two groups, rank the scores and use the *Mann-Whitney U test*.

    b. For three or more groups, rank the scores and use the *Kruskal-Wallis test*.
    With three or more groups, the ANOVA test reveals differences among the groups, but it does not identify which specific groups are different from others. A *post hoc* test identifies where the differences lie. A common mistake is to report the findings as, "There are no statistically significant differences between groups." Assuming a 0.05 alpha level has been chosen, it is better to state, "No statistically significant differences were found at the

$p<0.05$ level.”

4. If the data depend on other variables (e.g., test scores depend on scores on an intelligence test), then *analysis of covariance* methods (ANCOVA) should be used. ANCOVA is a rather advanced method of multivariate statistics. One simple use of ANCOVA is to compare the effects of a treatment on two groups that differ, say, in age or intelligence. In measuring the pure treatment effect (i.e., without the effects of error), one can factor out the effects of age or intelligence for arriving at a pure or adjusted score, which can then be analyzed using ANOVA.

Figure 2. Graphical decision tree for two or more groups.



Representative Studies

The following section briefly describes studies used in some of the more common statistical procedures in counseling research. Readers are encouraged to consider these, as well as many others, in detail as exemplars of good statistics. For example, Detillion, Craft, Glasper, Prendergrast, and DeVries (2004) described how the most complex phenomena can sometimes be approached via more basic inferential statistical analyses such as t-tests and the ANOVA, resulting in meaningful graphics for interpretation. Shackelford and Besser (2007) demonstrated the use of multiple regression to predict attitudes toward homosexuality based on the variables of educational level, age, conservatism, fundamentalism, and geographic immobility. Lutz, Greischar, Rawlings, Ricard, and Davidson (2004) exhibit the cutting edge of behavioral science with t-tests and ANOVA. Hinrichsen, Morrison, Waller, and Schmidt (2007) used chi-square tests in a straight-forward manner to determine that bulimic individuals reported core feelings of both shame/guilt and of anxiety/worry more frequently than either feelings of boredom or depression prior to self-induced vomiting. Conversely, Johnsen and Lutgendorf (2001) offered a more complicated study that demonstrates the use of t-tests, ANOVA, and regression analyses in the study of a fascinating topic.

Bangert and Baumberger (2005) presented specific statitistical techniques used in the *Journal of Counseling and Development* from 1990-2001. In 25% of the articles they surveyed, a total of 745 statistical methods were coded. The most frequently identified statistical procedures for all articles reviewed were those classified as basic (62%), followed by intermediate (22%), and advanced (16%). Basic statistics were found in articles at a significantly greater frequency when compared to intermediate or advanced statistical procedures. The most frequently tabulated statistics across all research reviewed were descriptive procedures (31%), Pearson correlation (12%), multiple regression (8%), one-way analysis of variance (ANOVA; 7%), and one-way multivariate analysis of variance/ covariance (MANOVA/MANCOVA; 7%).

Internet Research Statistical Trees

Here are three supplemental internet websites that cover information on statistical analysis decision trees:

1. http://www.math.nsc.ru/AP/datamine/eng/decisiontree.htm
   This site is from the Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences. It has very comprehensive explanations to create a stat decision tree and provides pdf document tutorials for the researcher wanting to learn more about this process.
2. http://www.microsiris.com/
   This site lists private companies that have provided freeware (no cost programs!) that help the researchers decide what statistical measure to use.
3. http://www.socialresearchmethods.net/Selstat/ssstart.htm
   This website has a ready-to-use html-based statistical decision tree.

Conclusion

The present article has been written to assist both counseling practitioners/mentors as well as students in determining possible “best fit” research measures. Answers to various questions posed in the article help funnel or focus the researcher toward the most desired research tool. Microsoft EXCEL applications are identified as are three related internet statistical decision tree resources.

References

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders (4th ed. TR)*. Washington D.C.: American Psychiatric Association.

Bangert, A. W., & Baumberger, J. P. (2005). Research and statistical techniques used in the Journal of Counseling & Development: 1990-2001. *Journal of Counseling & Development, 83,* 480-487.

Detillion, C. E., Craft, T. K. S., Glasper, E. R., Prendergrast, B. J., & DeVries, A. C. (2004). Social facilitation of wound healing. *Psychoneuroendocrinology, 29*(8), 1004-1011.

Hinrichsen, H., Morrison, T., Waller, G., & Schmidt, U. (2007). Triggers of self-induced vomiting in bulimic disorders: The roles of core beliefs and imagery. *Journal of Cognitive Psychotherapy*, *21*(3), 261-272.

Johnsen, E. L., & Lutgendorf, S. K. (2001). Contributions of imagery ability to stress and Relaxation. *Annals of Behavioral Medicine, 23*(4), 273-281.

Keller, D. K. (2006) *The Tao of Statistics: A Path to Understanding (With No Math).* Thousand Oaks, CA: Sage.

Lutz, A., Greischar, L. L., Rawlings, N. B., Ricard, M., & Davidson, R. J. (2004). Long-term Mediators self-induce high-amplitude gamma synchrony during mental practice. *Proceedings of the National Academy of Sciences, 101*(46), 16369-16373.

Microsoft Corp. (2008). *Math Functions.* Retrieved May 12, 2008, from http://office.microsoft. com/en-us/assistance/CH062528291033.aspx

Selecting statistics. (2008). Retrieved May 12, 2008, from http://www.socialresearchmethods. net/Selstat/ssstart.htm

Scheffe's method. (2008). Retrieved May 12, 2008, from http://www.itl.nist.gov/div898/ handbook/prc/section4/prc472.htm

Shackelford, T. K., & Besser, A. (2007). Predicting attitudes toward homosexuality: Insights from personality psychology. *Individual Differences Research, 5*(2), 106-114.

Statistical decision tree. (2008). Retrieved May 12, 2008, from http://www.microsiris.com/