Applying Basic Statistics: A Primer

The purpose of this primer is to introduce you to some basic terms, concepts, and formulas pertaining to understanding and applying statistics to research problems. The primer has seven sections: (a) A basic definition of several key terms; (b) terms and measures pertaining to determining central tendencies, often used as a preliminary examination of a database; (c) terms and measures pertaining to variability, those statistics useful in understanding how a single value in a database compares with others in the database; (d) those statistics useful in understanding the relationship between two or more variables or sets of data; (e) those statistics useful in comparing two or more variables with each other to determine if statistical differences exist between them; (f) a description of the steps you use in choosing appropriate statistical tests; and (g) some final comments and resource suggestions. In addition, two tables are included that provide databases with associated statistical calculations.

Definition of Terms

Database – a collection or distribution of numbers (scores) representing the values assigned to some variable.

Estimation – the use of some descriptive statistical measure, such as a mean, standard deviation, or correlation, to estimate or infer meaning back to a population. This also can be called a statistic.

Parameter – the value of a statistical measure or a database value. For example, if the mean (average) of a group is 17.4, that figure is one parameter. A **Parametric Measure** assumes the data being used in calculations are representative of a larger universe. A **Nonparametric Measure** generally is a distribution-free test whose model does not specify conditions about the representativeness of any sample.

Sample – ideally, a subset of a larger group or numerical database that is equal in all aspects and can therefore represent the larger group.

Statistic – a descriptive measure, such as a mean, standard deviation, or correlation, usually computed from a sample, that is used to estimate or infer meaning back to some population or database.

Statistical Inference – the drawing of generalizations from an available sample group that represents some larger numerical database.

Systematic Research Study (quantitative in nature) – a study so designed that logical reason will not be violated in making a transition from sample findings to generalizations about a universe (larger population or database).

Universe (or population) – a larger group (sometimes a total group) or numerical database of individuals about which various aspects (variables) are known or wish to be known.

Variables: **Dependent variable** – A dependent variable basically depends on or changes as a result of changes in any independent variable.

Independent variable – An independent variable is that factor, aspect, or measure that you, the researcher, changes or manipulates in order to examine what happens (or happened) during your study effort. Generally, you should have only one independent variable to truly understand what took place during the research effort.

For example, if you were measuring the growth of corn under two different types of fertilizer (the independent variable). In essence you are "manipulating" the type of fertilizer. The growth of the plant measured during various times of the growing season would be a dependent variable.

Typical Measures of Central Tendency

Measures of Central Tendency – measures that reflect the clustering of values around certain points in a database or distribution. Commonly used ones are the mean, median, and mode.

Mean (symbolized as $\mathbf{M_X}$) - the sum (Σ or the capital Greek letter sigma) of all distribution or database values (X_1, X_2, X_3, etc.) divided by the total number (N) of these values.

$$\mathbf{M_X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{N} = \mathbf{M_X} = \frac{\Sigma X}{N}$$

Median - that point on a distribution or in a sequentially continuous database above which and below which 50% of the individual values lie.

Mode - that value in a distribution which occurs most frequently. A distribution can be unimodal (only one "most frequent" peak), bi-modal (two peaks), or it can have multiple peaks when comparing values against the frequency of each unique value.

Quartile, Decile, or Percentile - a measure that describes a database at a particular point along a sequentially continuous distribution. For example, the first quartile (Q_1) is that point in a database below which 25% of the values lie. The symbols and calculation for this would be $Q_1 = 1N/4$ (1 for first quarter times N or the total number of data points divided by 4). As another example, the 81st percentile is that point in a database below which 81% of the values lie. The symbol and calculation for this would be $P_{81} = 81N/100$ (81 times N or the total number of data points divided by 100). For example, examine the following distribution of 12 test scores: 34, 38, 39, 40, 44, 49, 56, 57, 61, 68, 69, 70. For those scores, the first quartile would be $Q_1 = 1$ times 12 divided by 4 = 3. Thus, the first three test scores would be in the first quartile of scores.

Skewed Distribution - a non-symmetrical distribution of the values in a database where the number or frequency of identical values are shown against a sequentially continuous arrangement of those values. For example, in a positively skewed distribution, the median and mode will be to the left of the mean value.

Measures Pertaining to Variance

Basic Variability - the amount of spread or the scattering of values in a database; typically this is in comparison to some measure of central tendency. Often this is some calculation to determine how much any one value deviates from the mean.

Range - a measure of variability that is simply the difference between the highest and lowest scores in a database or distribution.

Standard Deviation (S) - a commonly used measure of variability. Mathematically, it is the square root (sqrt) of the sum (Σ) of all distribution or database values (X_1, X_2, X_3, etc.) in deviation form ($X - \mathbf{M_X}$) divided by the total number (N) of these values

$$S = \text{sqrt}\ \frac{\Sigma x^2}{N-1} \quad \text{where } x = X - \mathbf{M_X}, \quad N = \text{number of values}$$

To compute the standard deviation (1) compute the mean; (2) subtract the mean from each value in the distribution; (3) square each deviation; (4) sum the squares of the deviations; (5) divide this sum by the number of cases minus one (this is called the degree of freedom and it is always one less than the comparison numbers); and (6) extract the square root.

Variance - the value of the standard deviation calculations before the square root is extracted.

Relationships Between Variables

Coefficient of Correlation - a single value used in a database to represent the relationship between two sets of data pertaining to continuous variable information collected for each individual. Another way of saying this is that it represents the extent to which changes in one variable are accompanied by equal changes in another variable. Correlation scores can range from -1 to (+)1 or from perfect negative correlation to perfect positive correlation. Such a value can be used to (1) indicate the extent to which values of one variable may be predicted from known values of another variable; or (2) make comparisons between variables expressed in different units. There are two main types.

Pearson Product Moment (*r*) (the default correlation coefficient) - this utilizes values as they actually appear in the database. Mathematically, it is represented by the following formula:

$$\boldsymbol{r} = \frac{\Sigma xy}{NS_XS_Y}$$ where "*x*" represents one variable and "*y*" represents another variable

Spearman Rank Order (r_s) - this disregards the actual values and considers only their ranks in the database (1^{st}, 2^{nd}, 3^{rd}, etc.). Mathematically, it is represented by the following formula:

$$\boldsymbol{r}_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$ where D = difference between the ranks (see Table 1)

Table 1 Ranks of Height and Weight for a Group of Students

No.	Height	Rank	Weight	Rank	D	D^2
1	60	1	125	2	-1	1
2	61	2.5	121	1	1.5	2.25
3	61	2.5	136	4	-1.5	2.25
4	62	5	143	8	-3	9
5	62	5	135	3	2	4
6	62	5	140	6	-1	1
7	63	7.5	139	5	-2.5	6.25
8	63	7.5	152	11	-3.5	12.25
9	64	9.5	159	12.5	-3	9
10	64	9.5	143	8	1.5	2.25
11	66	11	144	10	1	1

12	67	12	143	8	4	16
13	68	13	159	12.5	.5	.25
14	69	14	178	14	0	0
15	70	15	179	15	0	0
16	71	16	187	16	0	0
17	72	17	201	17	0	0
						Σ or sum of $D^2 = 66.50$

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6(66.50)}{17(17^2 - 1)} = 1 - \frac{399}{4896} = 1 - .0815 = .919$$

Looking in a table of significance (tables representing different levels of significance typically are found in the rear of statistics books), we will find that the value of .919 is significant at a greater than a 99 % confidence level. Confidence level refers to the odds of stating that there is no relationship, difference, or gain, when, in fact, there is one. Typically, you say that you are 95% confident in your assertion about the statistical relationship (another way you will see it in journal articles is the .05 level) or 99% confident (.01 level). So in the case of Table 1, we are now confident at the 99% level of confidence that there is, in fact, a relationship or a strong correlation between height and weight.

Comparing Variables - Testing Hypotheses

Differences Between Two Means (*t*-test) - a test for significance to ascertain if the means for two groups are statistically different. Mathematically, it is represented by the following formula for groups of equal size:

$$t = \frac{M_X - M_Y}{\text{sqrt}\ \frac{\Sigma x^2}{N_X(N_X - 1)} + \frac{\Sigma y^2}{N_Y(N_Y - 1)}}$$

where X represents one group and Y represents the second group and $\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$ (same formula for Σy^2)

Differences Between Two Sample Frequencies (Chi square) - a test to determine whether the sample frequencies of values in a database are significantly different from those which would result if only chance factors were operating. Used primarily with dichotomous variables. Mathematically, it is represented by the following formula:

$$\chi^2 = \Sigma\left[\frac{(\text{actual frequency} - \text{expected frequency})^2}{\text{expected frequency}}\right]$$

The following (see Table 2) is based on the use of χ^2 (Chi-square) to determine if there are statistical differences:

H_o (the symbol for a null hypothesis or a belief that there is no difference between two samples) = There are no differences in the amount of time spent skiing among people who live north of the Ithaca in comparison with those who live south of Ithaca in the State of New York.

100 people were surveyed; 68 lived south of Ithaca, New York. The average hours spent skiing in the past year for the total group was 24. Thus, if you hypothesize that there will be no difference between the two groups, then you would "expect" that half of the 68 people or 34 people would be below the mean and half would be above the mean (remember how mean or average is determined) and the same logic for those living north of Ithaca.

Table 2 Hours of skiing compared with the geographic location

No. Hours of Skiing				
Where Lives	**<mean**	**Expected**	**>mean**	**Expected**
Lives S. of Ithaca	36	34	32	34
Lives N. of Ithaca	10	16	22	16

$$\chi^2 = \Sigma\left[\frac{(\text{actual frequency} - \text{expected frequency})^2}{\text{expected frequency}}\right]$$

$$\chi^2 = \Sigma\left[\frac{(36-34)^2}{34} + \frac{(32-34)^2}{34} + \frac{(10-16)^2}{16} + \frac{(22-16)^2}{16}\right]$$

$$\chi^2 = 4/34 + 4/34 + 36/16 + 36/16 = 4.74$$

Degrees of Freedom = number of rows - 1 times the number of columns - 1
In a 2 x 2 contingency table this equals 1

Looking in a table of Chi Square, a value of 4.74 with one degree of freedom is at the 95% level of confidence. In essence, the null hypothesis is not confirmed and it appears those who live north of Ithaca are more likely to ski than those who live south of the Ithaca.

You also will see authors talking about a Type I error (when the null hypothesis of no difference between two variables is rejected when it is in fact true – i.e., we conclude that two drugs are different in outcome after clinical tests, when they actually are not) and Type II error (when the null hypothesis is not rejected when it is in fact true – i.e., we conclude that there is no difference, but in fact there really is a difference).

Steps in Choosing Appropriate Statistical Tests

1. Determine the number of independent variables.

2. Determine the number of dependent variables.

3. Determine whether the variables are nominal, ordinal, or interval.

    a. Nominal scales are those where distinctive numbers or identifying labels have been assigned to groups or classes of objects (for example, male -vs- female).
    b. Ordinal scales are those where numbers assigned to objects are rank-ordered according to some characteristic (for example, ranked grades for achievement such as A, B, C, etc.).

c. Interval scales are those where numbers assigned to objects are rank-ordered but there are equal differences between the existing numbers) (for example, ages, valid IQ tests, etc.).

4. Determine if basic parametric assumptions are met.

5. Refer to the appropriate statistical choice table and select one or more tests.

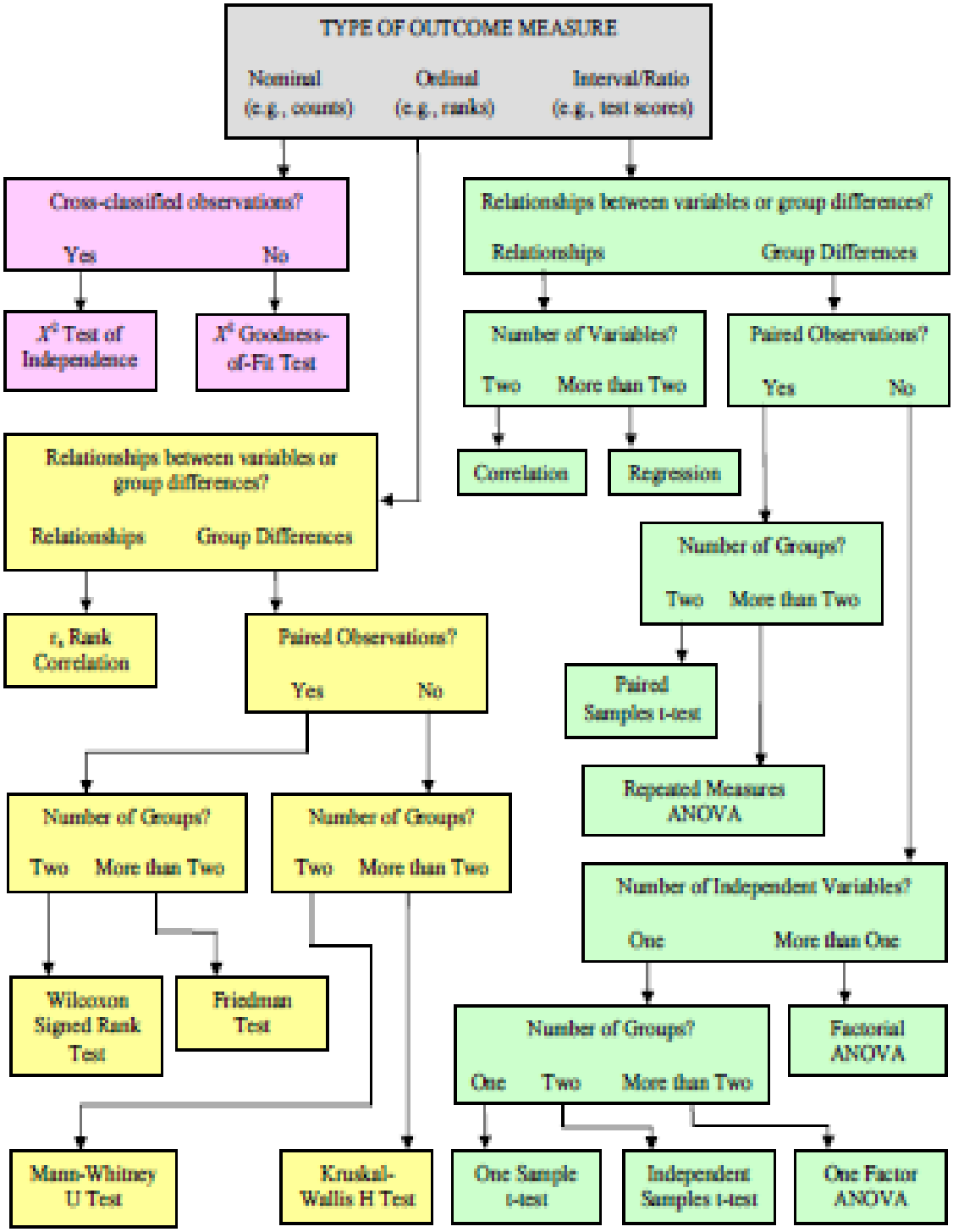
Example: A study of the effects on achievement of programmed materials for a group of randomly chosen Adult Basic Education students. A randomly chosen comparison group was taught in a traditional lecture mode without the use of programmed materials. The IQ for each student is known. Achievement has been measured on a standardized achievement test (it is an interval scale). Following the steps:

1. Two independent variables - IQ and programmed materials
2. One dependent variable – achievement
3. IQ - interval
   Programmed materials -vs- traditional teaching - nominal
   Achievement - interval

NOTE: Because interval and nominal variables are mixed together, in choosing appropriate statistics it is easiest to work with if you make the independent variables both at the same level. Thus, you convert the IQ to a nominal variable by categorizing the students as high IQ -vs- low IQ. A general rule of thumb: You can convert down in order but not up (you can't turn a nominal variable into an ordinal or interval variable).

4. Parametric assumptions have been met
5. Referring to the following flowchart for number and type of variables and you would choose ANOVA (Analysis of Variance).

Flowchart for Selecting the Appropriate Hypothesis Test



Some Final Comments

It is impossible to "teach" a statistics course in a short period of time. The purpose of this primer is to give you a "working" acquaintance with statistical manipulations so that when you read future journal articles or even plan your own future research, you will have an idea about what is said or what you can plan.

The Web is filled with many excellent resources to help you learn more about statistics. Check out this one:

http://www.statsoft.com/textbook/stathome.html

_____
Roger Hiemstra