Some Critical Information about *SOME* Statistical Tests and Measures of Correlation/Association

This information is adapted from and draws heavily on: Sheskin, David J. 2000. *Handbook of Parametric and Nonparametric Statistical Procedures.* Second Edition. Chapman & Hall/CRC, Boca Raton. 982 pp. I strongly recommend that you consult Sheskin for other statistical tests and measures of correlation. This is simply a **short, incomplete description of some of the limitations and assumptions for commonly used tests and measures of association**. **My objective is to show the relationship between what kind of data you collect, how many samples or populations you include, and how the sample was selected (probabilistic or not) and the kinds of statistical tests you can use to analyze your data. Consult a statistician (preferable) or at least a statistics book before you use any statistical procedure.** Sheskin provides much more detail about these and many other statistical tests and measures. In short, use this information with great care!!! I am not a statistician.

TESTS OF CENTRAL TENDENCY

I. Interval or Ratio Data - Parametric Tests – Means Tests

A. One sample

1. **Single Sample z Test**
   a. What it tests: Whether a sample of subjects or objects comes from a population – does the sample mean equal the population mean?
   b. Limitations: You must know the standard deviation and mean of the population.
   c. Assumptions: The sample represents the population. The sample was randomly selected. The population is normally distributed.

2. **Single-Sample t Test**
   a. What it tests: Whether a sample of subjects or objects comes from a population – does the sample mean equal the population mean?
   b. Limitations: You must know the mean of the population
   c. Assumptions: The sample represents the population. The sample was randomly selected. The population is normally distributed.

B. Two or more independent samples (*independent* samples means the two samples are comprised of *different* subjects or objects)

1. **t Test for *Two* Independent Samples**
   a. What it tests: Do two independent samples represent two different populations with different mean values
   b. Limitations: You can only compare *two* samples, no more
   c. Assumptions: The samples are representative of the populations. The samples were randomly selected. The samples are independent. Both populations are normally distributed. The variances of the two populations are equal.

2. **Single Factor Between-Subjects or One Way Analysis of Variance (ANOVA)**
   a. What it tests: In a group of any number of samples (three, five, ten), do **at least two** of the samples represent populations with different mean values?

b.   Additional procedures: This test does ***not*** tell you which of the means differed – just that there was a difference between some of them. For planned comparisons you may use ***multiple t tests*** to determine which means differ. For unplanned tests you may use ***Fisher's LSD test*** to determine which means differ.

c.   Limitations: Only ***one*** independent variable

d.   Assumptions: Samples are representative of the populations. The samples were selected randomly. The samples are independent. All of the populations are normally distributed. The variances of all of the populations are equal.

3. **Single Factor Between-Subjects Analysis of Covariance (ANCOVA)**

a.   What it tests: It is a form of ANOVA. It allows you to use data about an extraneous (non-experimental) variable that has a linear correlation with the dependent variable to (1) remove variability in the dependent variable and/or (2) adjust the mean scores of the different groups for any pre-existing differences in the dependent variable that were present prior to the administration of the experimental treatments. The most commonly used co-variate (the extraneous or non-experimental variable) is a pretest score for the dependent variable.

b.   Limitations: Only ***one*** extraneous variable. Single factor ANCOVA is sometimes used for a design in which subjects are not randomly assigned to groups (quasi-experimental designs). This use is problematic! This includes in some cases using single factor ANCOVA for inferential designs (ex post facto studies where the group are based on something like sex, income or race). This is even more problematic!

c.   Assumptions: Samples are representative of the populations. All of the populations are normally distributed. The variances of all of the populations are equal.

C. **Two or More Dependent Samples (*Dependent* samples mean that [1] each subject serves as a *member of all* treatment groups and of the control group or that [2] every subject *is paired with* a subject in each of the other test groups, in which case the pairing must be justified. You do this by identifying one or more variables other than the independent variable that you believe are positively correlated with the dependent variable. You "match" subjects based on their similarity in regard to this (or these) variables. Then you randomly assign "matched" members to the control and treatment groups.)**

1. **t Test for Two Dependent Samples**

a.   What it tests: Do two dependent samples represent populations with different mean values

b.   Limitations: Only ***two*** samples (groups, populations)

c.   Samples are representative of the populations. Samples were randomly selected. Both populations are normally distributed. The variances of the two populations are equal.

2. **Single Factor Within-Subjects ANOVA**

a.   What it tests: In a group of any number of dependent samples (three, five, ten), do ***at least two*** of the samples represent populations with different mean values?

b.   Additional procedures: This test does ***not*** tell you which of the means differed – just that there was a difference between some of them. For planned comparisons you may use ***multiple t tests*** to determine which means differ. For unplanned tests you may use ***Fisher's LSD test*** to determine which means differ.

c. Limitations: Only *one* independent variable

d. Assumptions: Samples are representative of the populations. Samples were randomly selected. All of the populations are normally distributed. The variances of all of the populations are equal.

D. **Two or More Samples and Two or More Independent Variables or Factors**

1. **Between-Subjects Factorial ANOVA**

a. What it tests: (1) Do at least two of the levels of each factor (A, B, C, etc.) represent populations with different mean values? (2) Is there an interaction between the factors?

b. Additional procedures: Case 1 – there were no significant main effects (no differences between factors) and there is no significant interaction. You can conduct any planned tests, but it is probably fruitless to conduct unplanned tests. Case 2 – there were significant main effects (differences between factors), but there was no significant interaction. In this case you can treat the factors separately – just ignore interaction. Use a **single factor between-subjects ANOVA**. Case 3 – interaction is significant, whether or not there are any significant main effects (differences between factors). Use a **single factor between-subjects ANOVA for all levels of one factor across only one level of the other factor.** (E.g., hold one factor constant while you allow the other to vary.)

c. No limitations of note

d. Assumptions: Samples are representative of the populations. Samples were randomly selected. Samples are independent. All of the populations are normally distributed. The variances of all of the populations are equal.

II. **Ordinal (Rank Order) Data - Nonparametric Tests – Median Tests**

A. **One Sample**

1. **Wilcoxon Signed-Ranks Test**
   a. What it tests: Whether a sample of subjects or objects comes from a population – does the sample median equal the population median?
   b. Limitations: You must know the median of the population.
   c. Assumptions: The sample is representative of the population. The sample was randomly selected. The population distribution is symmetrical.

B. **Two or more independent samples**

1. **Mann-Whitney U Test**
   a. What it tests: Do two independent samples represent two populations with different median values?
   b. Limitations: You can only compare *two* samples, no more. Do *not* use this test for proportions (percentages).
   c. Assumptions: The samples are representative of the populations. The samples were randomly selected. The samples are independent. The original variable that was measured was a *continuous* random variable (this assumption is often violated – no idea if that's OK or not, but Sheskin does not seem to think it is a big deal). The distributions of the populations are identical in shape.

2. **Kruskal-Wallis One-Way Analysis of Variance by Ranks**
   a.   What it tests: In a group of any number of independent samples (three, five, ten), do *at least two* of the samples represent populations with different median values?
   b.   Additional procedures: Like the parametric ANOVA, the Kruskal-Wallace test does not tell you which of the means differed. You must perform pairwise comparisons to determine where the differences lie. See a good statistics book about how to do this. You can use the **Mann-Whitney U test** but there are certain conditions that must be met for this procedure to work. Again consult a statistics book.
   c.   Limitations: Only *one* independent variable
   d.   Assumptions: Samples are randomly selected. Samples are representative of the populations. Samples are independent of one another. The original variable that was measured was a *continuous* random variable (this assumption is often violated – no idea if that's OK or not, but Sheskin does not seem to think it is a big deal). The distributions of the populations are identical in shape.

C. **Two or more dependent samples**

   1. **Wilcoxon Matched Pairs Signed Ranks Test**
      a.   What it tests: Do two dependent samples represent two different populations?
      b.   Limitations: Only *two* samples, no more. You must have *two scores* to compare for this test because it is based on the *difference* between the two. These can be two scores for the same subject (first as a control and then as a treatment) or two scores for matched pairs of subjects (one in the control group and one in the treatment group).
      c.   Assumptions: Samples are randomly selected. Samples are representative of the populations. The distribution of the *difference scores* in the populations is *symmetric* around the median of the population of difference scores.

   2. **Binomial Sign Test for Two Dependent Samples**
      a.   What it tests: Do two dependent samples represent two different populations?
      b.   Limitations: Only *two* samples. You need two scores. This test is based on whether the subject's (or matched pairs of subjects) score increases or decreases – by the *sign* (positive or negative). You can use this test with the assumption of symmetric distribution for the Wilcoxon Matched Paris Test is violated.
      c.   Assumptions: Samples are randomly selected. Samples are representative of the populations.

   3. **Friedman Two-Way Analysis of Variance by Ranks**
      a.   What it tests: In a group of any number of dependent samples (three, five, ten), do *at least two* of the samples represent populations with different median values?
      b.   Additional procedures: Like the parametric ANOVA, the Kruskal-Wallace test does not tell you which of the means differed. You must perform pairwise comparisons to determine where the differences lie. See a good statistics book to learn how to do this. You can use the **Wilcoxon matched pairs signed ranks test** or the **binomial sign test for two dependent samples**. See a statistics book to learn how to do this.
      c.   Assumptions: Samples are randomly selected. Samples are representative of the populations. The original variable that was measured was a *continuous* random variable (this assumption is often violated – no idea if that's OK or not, but Sheskin does not seem to think it is a big deal).

III. Nominal (Categorical) Data – Nonparametric Tests
   A.   NONE – There is no central tendency to test

TESTS OF DISPERSION

I.   Interval or Ratio Data - Parametric Tests – Variance

   A. **Single sample**

      1. **Single Sample Chi-Square Test for Population Variance**
         a.   What it tests: Does a sample come from a population in which the variance equals a known value?
         b.   Limitations: You must know the variance of the population.
         c.   Assumptions: The sample was selected randomly. The sample is representative of the population. The population is normally distributed.

   B. **Two or more independent samples**

      1. **Hartley's F(max) Test for Homogeneity of Variance**
         a.   What it tests: Are the variances of two or more populations equal?
         b.   Assumptions: The samples were selected randomly. The samples are representative of the populations. The populations are normally distributed. Sample sizes should be equal or approximately equal.

   C. **Two or more dependent samples**

      1. **The t test for Homogeneity of Variance for Two Dependent Samples**
         a.   What it tests: Are the variances of two populations equal?
         b.   The samples were selected randomly. The samples are representative of the populations. The populations are normally distributed.

II. Ordinal or Rank Ordered Data - Nonparametric Tests - Variability

   A. **Single samples**

      1. **NONE**

   B. **Two or more independent samples**

      1. **The Siegel-Tukey Test for Equal Variability**
         a.   What it tests: Do two independent samples represent two populations with different variances?
         b.   Limitations: You must know or be willing to make some assumptions about the medians of the two populations (see assumption 3 below).
         c.   Assumptions: The samples were randomly selected. They are representative of the populations and they are independent. The samples represent populations with equal medians. If you know the medians of the populations and they are not equal, you can perform some adjustments and still use this test. If you do not know the medians and you are unwilling to assume they are equal (probably normally the case), do not use this test.

2. **Moses Test for Equal Variability**
   a.   What it tests: Do two independent samples represent two populations with different variances?
   b.   Limitations: The data for the dependent variable must have been interval or ratio data originally that were later transformed to ordinal data and the dependent variable must have been a **continuous** variable (not discrete).
   c.   Assumptions: The samples were randomly selected. The samples are independent and representative of the populations. The original data for the dependent variable were interval or ratio data (they were transformed to ordinal data later). The original data for the dependent variable were **continuous** (could assume any value). The distribution of two or more populations must have the same general shape (although it need not be normal).

C. **Two or more dependent samples**
   1.   None that I know

III. **Nominal (Categorical) Data**

A. **None – no adequate measures of variability**

Tests of Distribution

I.   **Interval or Ratio Data** – **Parametric Tests**

A. **One Sample**

   1. **Single Sample Test for Evaluating Population Skewness**
      a.   What it tests: Does the sample come from a population distribution that is symmetrical (not skewed)?
      b.   Limitations: None
      c.   Assumptions: The sample is representative of the population. The sample was randomly selected.

   2. **Single Sample Test for Evaluating Population Kurtosis**
      a.   What it tests: Does the sample come from a population distribution that is mesokurtic (not peaked)?
      b.   Limitations: None
      c.   Assumptions: The sample is representative of the population. The sample was randomly selected.

   3. **D'Agostino-Pearson Test of Normality**
      a.   What it tests: Does the sample come from a population that is normally distributed?
      b.   Limitations: None
      c.   The sample is representative of the population. The sample was randomly selected.

B. **Two or More Independent Samples**
   1.   Use the **Single Sample Test for Evaluating Population Skewness,** the **Single Sample Test for Evaluating Population Kurtosis** and the **D'Agostino-Pearson Test of Normality** for each sample.

C.   **Two or More Dependent Samples**
    1.   Use the **Single Sample Test for Evaluating Population Skewness,** the **Single Sample Test for Evaluating Population Kurtosis** and the **D'Agostino-Pearsone Test of Normality** for each sample.

II.   **Ordinal (Rank Order) Data – Nonparametric Tests**

A.   **One Sample**

    1.   **Kolmogorov-Smirnov Goodness-of-Fit Test for a Single Sample**
        a.   What it tests: Does the distribution of scores in a sample conform to a **specific theoretical or empirical (known)** population distribution?
        b.   Limitations: You must know the distribution of the population. This can be a theoretical distribution (such as the normal distribution) or an empirical (real) distribution. The dependent variable must be **continuous** (not discrete). This tests takes continuous the continuous variable and converts the data into a cumulative frequency (hence it becomes nonparametric data) – but you must start with a continuous variable.
        c.   Assumptions: The samples were randomly selected. The samples are independent and representative of the populations. The original data for the dependent variable were **continuous** (could assume any value).

    2.   **Lillefor's Test for Normality**
        a.   What it tests: Does the distribution of scores in a sample conform to a population distribution for which either the mean or the standard deviation (or both) must be estimated (an unknown distribution)?
        b.   Limitations: The dependent variable must be **continuous** (not discrete). This tests takes continuous the continuous variable and converts the data into a cumulative frequency (hence it becomes nonparametric data) – but you must start with a continuous variable.
        c.   Assumptions: The samples were randomly selected. The samples are independent and representative of the populations. The original data for the dependent variable were **continuous** (could assume any value).

B.   **Two or More Independent Samples**
    1.   Use the **Kolmogorov-Smirnov Goodness-of-Fit Test** or the **Lillefor's Test for Normality** for each sample if the data for the dependent variable are **continuous.**
    2.   Use the **Chi-Square Goodness-of-Fit** or the **Binomial Sign Test** for each sample if the data for the dependent data are not **continuous.**

C.   **Two or More Dependent Samples**
    1.   Use the **Kolmogorov-Smirnov Goodness-of-Fit Test** or the **Lillefor's Test for Normality** for each sample if the data for the dependent variable are **continuous.**
    2.   Use the **Chi-Square Goodness-of-Fit** or the **Binomial Sign Test** for each sample if the data for the dependent data are not **continuous.**

III. Nominal (Categorical) Data

A. **Single Sample**

1. Chi-Square Goodness of Fit Test
a. What it tests: Are the observed frequencies different from the expected frequencies?
b. Limitations: You must know the expected frequency for each category of responses. This can either be based on a theoretical (probability-based) distribution or based on some pre-existing empirical information about the variable you are measuring.
c. Assumptions: The sample was randomly selected and represents the population. The categories are mutually exclusive. Each observation is represented only once in the data set. **The expected frequency of each cell is five or greater.**

2. **Binomial Sign Test for a Single Sample**
a. What it tests: Are the observed frequencies for **two** categories different from the expected frequencies?
b. Limitations: You must know the expected frequency for each category of responses. This can either be based on a theoretical (probability-based) distribution or based on some pre-existing empirical information about the variable you are measuring. This test is just like the Chi-Square Goodness-of-Fit test, but is used for small sample sizes (cell frequency does not have to be five or greater).
c. Assumptions: The sample was randomly selected and represents the population. The categories are mutually exclusive. Each observation is represented only once in the data set.

3. **Chi-Square Test of Independence**
a. What it tests: Is there a relationship between two variables measured for the same sample; e.g., does response for one of the variables predict response for the other variable?
b. Limitations: Does not work for very small samples (see assumptions).
c. Assumptions: The sample was randomly selected and represents the population. The categories are mutually exclusive. Each observation is represented only once in the data set. **The expected frequency of each cell is five or greater.**

B. **Two or More Independent Samples**

1. **Chi-Square Test for Homogeneity**
a. What it tests: Whether or not two or more samples are homogeneous with respect to the proportion of observations in each category of response
b. Limitations: Does not work for very small samples (see assumptions)
c. Assumptions: The sample was randomly selected and represents the population. The categories are mutually exclusive. Each observation is represented only once in the data set. **The expected frequency of each cell is five or greater.**

2. **Fisher Exact Test**
a. What it tests: Whether or not two or more samples are homogeneous with respect to the proportion of observations in each category of response
b. Limitations: None – commonly used to replace the **Chi-Square Test for**

Homogeneity for small samples

c.   Assumptions: The sample was randomly selected and represents the population. The categories are mutually exclusive. Each observation is represented only once in the data set.

C. **Two or More Dependent Samples**

1. **McNemar Test**
a.   What it tests: Do two dependent samples represent two different populations?
b.   Limitations: Only **two** samples. Not good for small samples.
c.   Assumptions: The samples were randomly selected and are representative of the populations. Each observation is **independent** of all other observations. The scores (dependent variable data) are dichotomous. Categories are mutually exclusive.

2. **Cochran Q Test**
a.   What it tests: Among several (three, five, whatever) different samples, do at least two of them represent different populations?
b.   Limitations: This test does not tell you **which** samples differed. You must perform additional tests to determine where the differences lie. You can use the **McNemar Test** to make pairwise comparisons.
c.   Limitations: Not good for small samples.
d.   The samples were randomly selected and are representative of the populations. Each observation is **independent** of all other observations. The scores (dependent variable data) are dichotomous. Categories are mutually exclusive.

MEASURES OF ASSOCIATION

I.   **Interval or Ratio Data – Parametric Tests**

A. **Bivariate Measures**

1. **Pearson Product-Moment Correlation Coefficient**
a.   What it tests: Is there a significant linear relationship between two variables (X or predictor and Y or criterion or predicted) in a given population?
b.   Other calculations needed: The "size" of the Pearson correlation coefficient (r) in an of itself may or may not indicate a **statistically significant** relationship between predictor variables and the criterion variable. At a minimum, you should use a **Table of Critical Values for Pearson r** and report this value when you use this statistic. The values vary for **one-tailed and two-tailed** hypotheses. Large r values can be meaningless. Alternatively, small values can be meaningful! You may also need to conduct one or more other tests for evaluating the value of the coefficients. Failure to take this step is common and makes many presentations of measures of association fairly useless. The "r" value alone is not enough!
c.   Limitations: This is a bivariate measure – only **two** variables
d.   Assumptions: The sample was randomly selected and represents the population. The two variables have a **bivariate normal distribution** – each of the two variables and the linear combination of the two variables are normally distributed. The relationship between the predictor (X) and criterion (Y or predicted) variables is of equal strength across the whole range of both variables (**homoscedasticity**). There is no **autocorrelation** between the two variables.

B. **Multivariate Measures**

1. **Multiple Correlation Coefficient**
    a.   What it tests: Is there a significant linear relationship between two or more predictor (X) variables and a criterion (Y or predicted) variable in a given population?
    b.   Other calculations needed: The "size" of the multiple correlation coefficient (R) in an of itself may or may not indicate a **statistically significant** relationship between predictor variables and the criterion variable. At a minimum, you should compute the R^2 statistic – the **coefficient of multiple determination.** Then compute the F statistic for R^2. Use a **Table of the F Distribution** to determine significance. Large R values can be meaningless. Alternatively, small values can be meaningful! You may also need to conduct one or more other tests for evaluating the value of the coefficient. Failure to take this step is common and makes many presentations of measures of association fairly useless. The "R" or "R^2" value alone is not enough!
    c.   Limitations: Although you can use a large number of predictor variables, the additional predictive power gained from adding more variables to the model decreases greatly after a few "good" predictors have been identified.
    d.   Assumptions: The sample was randomly selected and represents the population. The variables have a **bivariate normal distribution** – each of the variables and the linear combination of the variables are normally distributed. The relationship between the predictor (X) and criterion (Y or predicted) variables is of equal strength across the whole range of both variables (**homoscedasticity**). There is no **multicollinearity** between the predictor variables – they are not strongly correlated **to each other**.

2. **Partial Correlation Coefficient**
    a.   What it tests: What is the strength of the relationship between **one** predictor variable of several and the criterion variable? Put another way, you hold the values for all other predictor variables constant and then measure the strength of the one variable that interests you. It is sort of the reverse of multiple correlation.
    b.   Other calculations needed: The "size" of the partial correlation coefficient (r) in an of itself may or may not indicate a **statistically significant** relationship between the predictor variable and the criterion variable. At a minimum, you should compute the value for t and then use a **Table of Student's t Distribution** to determine significance. The values vary for **one-tailed and two-tailed** hypotheses. Large r values can be meaningless. Alternatively, small values can be meaningful! You may also need to conduct one or more other tests for evaluating the value of the coefficients. Failure to take this step is common and makes many presentations of measures of association fairly useless. The "r" value alone is not enough!
    c.   Assumptions: The sample was randomly selected and represents the population. The variables have a **bivariate normal distribution** – each of the variables and the linear combination of the variables are normally distributed. The relationship between the predictor (X) and criterion (Y or predicted) variables is of equal strength across the whole range of both variables (**homoscedasticity**).

II. **Ordinal or Rank Order Data – Nonparametric Measures**

A. **Bivariate Measures**

1. **Spearman's Rank-Order Correlation Coefficient**
    a.   What it tests: In a sample from a population is there a correlation (relationship)

between subjects' scores on two different variables? Put another way, does a test subject's score for Variable 1 (X) predict his/her score for Variable 2 (Y)?

b.   Other calculations needed: The "size" of the Spearman's rank-order correlation coefficient (r_s) or Spearman's Rho in an of itself may or may not indicate a **statistically significant** relationship between the two variables. You use a **Table of Critical Values for Spearman's Rho** to determine significance. There are equations you can use, too, one of which gives a t value and one of which gives a z value. The values vary for **one-tailed and two-tailed** hypotheses. Large r_s values can be meaningless. Alternatively, small values can be meaningful! You may also need to conduct one or more other tests for evaluating the value of the coefficients. Failure to take this step is common and makes many presentations of measures of association fairly useless. The "r_s" value alone is not enough!

c.   Limitations: Only **two** variables

d.   Assumptions: The sample was randomly selected and represents the population. The relationship between the predictor (X) and criterion (Y or predicted) variables is of equal strength across the whole range of both variables (**homoscedasticity**).

B. **Multivariate Measures**
    1. **Kendall's Coefficient of Concordance**
        a.   What it tests: In a sample from a population is there a correlation (relationship) between subjects' scores on three or more different variables? Put another way, does a test subject's score for Variables 1, 2, 3 ... (X1, X2, X3...   ) predict his/her score for Variable 2 (Y)?

        b.   Other calculations needed: The "size" of the Kendall's coefficient of concordance (W) in an of itself may or may not indicate a **statistically significant** relationship between the two variables. You use a **Table of Critical Values for Kendall's Coefficient of Concordance** to determine significance. You can also computer the significance using the Chi-square statistic and a **Table of the Chi-Square Distrition**. The values vary for **one-tailed and two-tailed** hypotheses. Large W values can be meaningless. Alternatively, small values can be meaningful! You may also need to conduct one or more other tests for evaluating the value of the coefficients. Failure to take this step is common and makes many presentations of measures of association fairly useless. The "W" value alone is not enough!

        c.   Assumptions: The sample was randomly selected and represents the population. The relationship between the predictor (X) and criterion (Y or predicted) variables is of equal strength across the whole range of both variables (**homoscedasticity**).

III. Nominal or Categorical Data – Nonparametric
    There are several measures of association for nominal or categorical data. They are all related to the **Chi-Square Test for Homogeneity**. They include the **Contingency Coefficient**, the **Phi Coefficient**, **Cramer's Phi Coefficient, Yule's Q** and the **Odds Ratio.** All of these measures measure the degree to which frequencies in one cell of a contingency table are associated with frequencies in other cells or categories – that is, the degree of association between the two variables. These measures provide you with precise information about the magnitude of the treatment effect. The **Contingency Coefficient** can be applied to more than two variables. The **Phi Coefficient, Cramer's Phi Coefficient** and **Yule's Q** can only be used for two variables. The **Odds Ratio** can be used for more than two variables, but usually is not because it becomes difficult to interpret the results. See a good statistics book if you need to use these measures.