WINCROSS STATISTICAL REFERENCE

The Analytical Group Inc.

15300 N. 90th Street • Suite #500 • Scottsdale, AZ 85260

+1.480.483.2700 • www.analyticalgroup.com

TABLE OF CONTENTS

STATISTICAL REFERENCE

General Notation

Unless otherwise specified **n** will denote the number of observations in the data set, and the observations will be denoted by $\mathbf{x_1, x_2, \ldots, x_n}$. If the observations can only take on a specific set of possible values, **k** will denote the number of specific values, the set of specific values will be denoted by $\mathbf{x_1, x_2, \ldots, x_k}$, and the frequencies of occurrence of these specific values in the data set will be denoted by $\mathbf{f_1, f_2, \ldots, f_k}$. In that case,

$$\sum_{i=1}^{k} f_i = n$$

When each of the observations is weighted we will denote the weights by $\mathbf{w_1, w_2, \ldots, w_n}$.

STATISTICS WITH UNWEIGHTED DATA

Mean

The sample mean $\bar{x}$ is calculated as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

An alternative computation, in the case when we have k distinct data values, is

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

Variance

The sample variance s^2 is calculated as

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

A convenient computational equivalent for s^2 is given by the expression

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

An alternative computation, in the case when we have k distinct data values, is

$$s^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1}$$

or, in simpler computational form,

$$s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - n\bar{x}^2}{n-1}$$

Technical Comment:

Variance of a Sample Proportion

When the x_i take on only the possible values 0 and 1, then the numerator of $\bar{x}$ is the count of the 1's, and so $\bar{x} = p$, the proportion of 1's. In that case

$$\sum_{i=1}^{n} x_i^2 = n$$

so that

$$s^2 = \frac{np - np^2}{n-1} = \frac{np(1-p)}{n-1}$$

But the estimate of the variance of the x's in this case should be p(1-p). So we see that by using the formula for s^2 to calculate an estimate of the sample variance in this case produces an overestimate by a factor of n/(n-1). If therefore one uses a computer program that calculates estimated variances using the formula for s^2 when the variables are binary 0,1 variables one must modify the computed variance by multiplying it by (n-1)/n, i.e., the variance should be

$$p(1-p) = [(n-1)/n]s^2$$

In cases in which the variance of a proportion is necessary, such as testing hypothesis about row proportions, WinCross automatically calculates the variance as

$$s^2 = p(1-p).$$

Standard Deviation

The standard deviation of the x's is given by

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

The standard deviation of a proportion p is given by

$$s = \sqrt{p(1-p)}$$

Standard Error
The standard error is defined as the standard deviation divided by the sample size, i.e.,

$$s_{\bar{x}} = s / \sqrt{n}$$

Grouped Median
We are given a table with k rows, with each row associated with a range of possible values of a measurement (e.g., the table has k age groups, with each row representing an age range), and with the ranges listed in ascending value. Let f_i be the count of the number of measurements in row i (in our example, the number in the sample in the age range for row i). Let m denote the row number of the table containing the 50th percentile. Let L_m and U_m denote the lower and upper boundary of the range associated with row m. Let

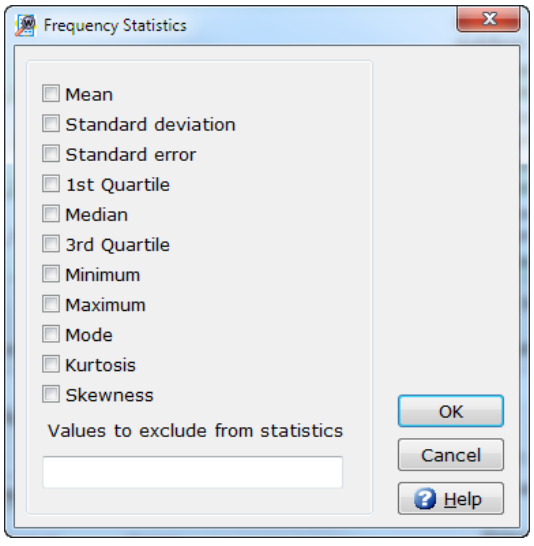
$$F_m = \sum_{i=1}^{m-1} f_i$$

i.e., the cumulative count up to but not including row m.

The grouped median is computed as follows:

$$median = L_m + \frac{(n/2 - F_m)(U_m - L_m)}{f_m}$$

Skewness and Kurtosis
When one selects the **Frequency** option on the **Run** menu and one wishes to augment the frequencies with summary statistics the following window appears, presenting all the statistics that can be calculated for the Frequency.



In particular, note that here in addition to the standard statistics described above WinCross can calculate the Mode (the most frequent value) as well as the values of the Skewness and Kurtosis statistics.

The unbiased estimate of skewness is calculated as:

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$$

The unbiased estimate of kurtosis is calculated as:

$$\left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4\right\}-\frac{3(n-1)^2}{(n-2)(n-3)}$$

(Previous versions of WinCross calculated the skewness and kurtosis statistic more directly by their population counterparts as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4-3$$

Though these are consistent estimates of the population skewness and kurtosis, these estimates have been replaced by the above unbiased estimates to conform to the computations of other commonly used software such as Excel.)

Previous versions of WinCross did not calculate the standard error of each of these statistics. The current version does this calculation as well. The standard error of the skewness estimate is:

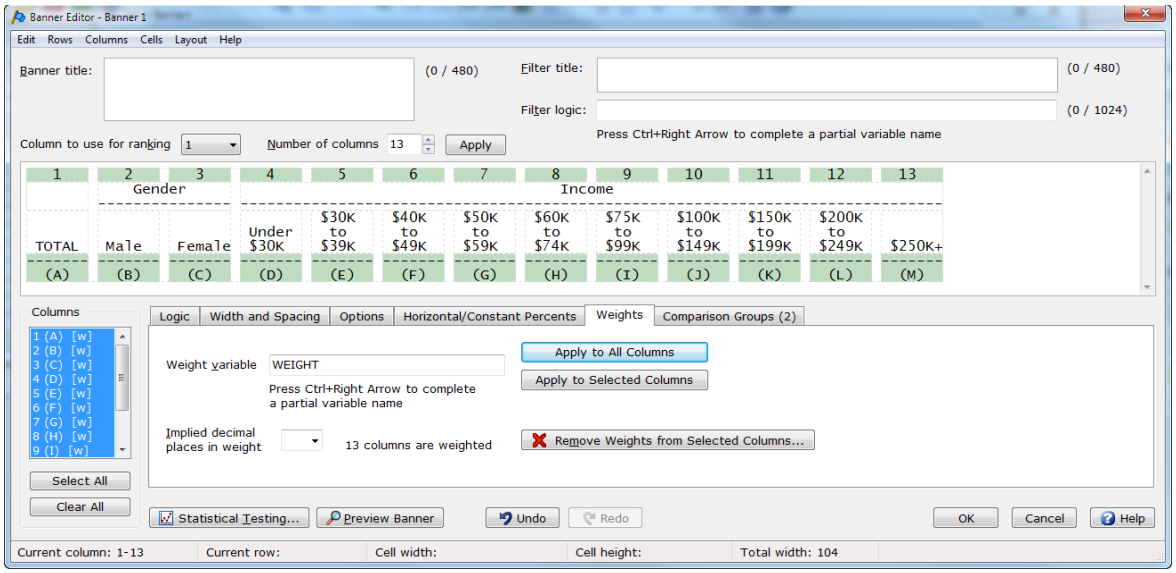
$$se_s=\sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

and the standard error of the kurtosis estimate is:

$$se_k=\sqrt{\frac{4(n^2-1)(se_s)^2}{(n-3)(n+5)}}$$

STATISTICS WITH WEIGHTED DATA

WinCross has the ability to apply separate weights to different variables. It does this using the following **Banner Editor** screen:



In this section we only look at a single weighted variable and describe various statistics calculated by WinCross using that variable's weight. In subsequent sections, we will treat separately, statistical testing where a single weight is applied to all variables and where each variable has a different associated weight.

Weighted Mean

The weighted sample mean is calculated as

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Weighted Variance

The weighted variance is calculated as

$$s_w^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2}{\sum_{i=1}^{n} w_i - 1}$$

When the x_i take on only the possible values 0 and 1, then the numerator of $\bar{x}_{1w}$ is the weighted count of the 1's, and so $\bar{x}_{1w} = p_w$, the weighted proportion of 1's. In this case the weighted variance is given by

$$s_w^2 = p_w(1 - p_w)$$

Weighted Standard Deviation

The weighted standard deviation is calculated as the square root of the weighted variance, namely

$$s_w = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2}{\sum_{i=1}^{n} w_i - 1}}$$

or, when dealing with proportions,

$$s_w = \sqrt{p_w(1 - p_w)}$$

Weighted Sample Size

The weighted sample size is calculated as the sum of the weights of all the observations,

$$\sum_{i=1}^{n} w_i$$

Effective Sample Size

Just as the standard error is defined as the standard deviation divided by the square root of the sample size, some software systems (e.g., SPSS, CfMC) define the weighted standard error as the weighted standard deviation divided by the square root of the weighted sample size. There are strong theoretical arguments to indicate that use of this computation of the weighted standard error is inappropriate. Those arguments are given on our website. Just go to WinCross FAQ's and click on any of the four articles, listed under HELPFUL DOCUMENTS, for in-depth discussion of this topic. These articles are described briefly in Appendix I.

Rather, the appropriate measure of the sample size of weighted data to be used in computing the weighted standard error is a construct which we call the "effective sample size," which is computed as

$$e = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$$

This is sometimes referred to as the "design effect" for weighted sampling.

Weighted Standard Error

WinCross calculates the weighted standard error as the unweighted standard deviation divided by the effective sample size, i.e., as

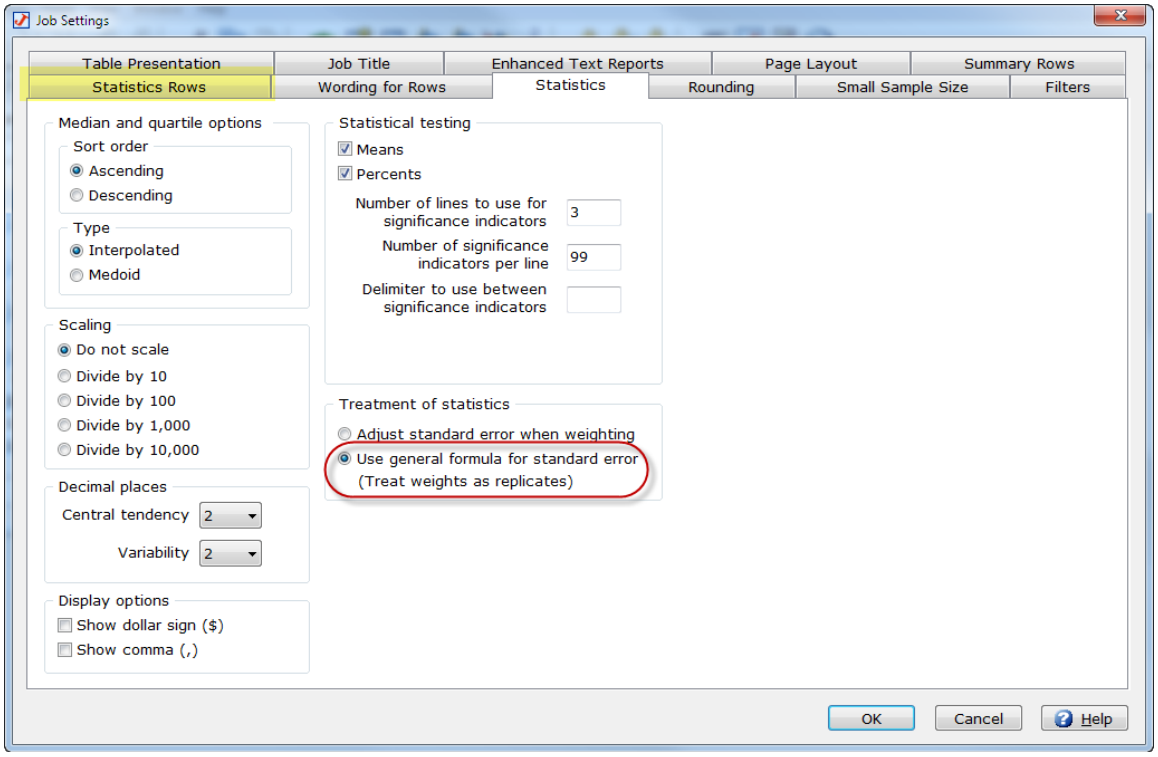
$$s_{\bar{x}_w} = s / \sqrt{e}$$

This estimate is the unbiased minimum variance estimate of the population standard error.

As noted earlier, other software systems compute the weighted standard error as the weighted standard deviation divided by the square root of the weighted sample size, i.e., as

$$s^*_{\bar{x}_w} = s_w / \sqrt{\sum_{i=1}^{n} w_i}$$

WinCross produces the weighted standard error $s^*_{\bar{x}_w}$ given above as a descriptive statistic, but only as an option does it use it in calculating the t statistic for weighted data. To invoke this option, on the **Job Settings|Statistics** tab, select the **Use General formula for standard error (Treat weights as replicates)** option, as noted on the next page:



This statistic, used by SPSS, is a biased estimate of the population standard error. This statistic has been modified by CfMC's Mentor to create from it an unbiased estimate of the population standard error. But, as shown in the articles listed in Appendix I, that estimator is NOT the most efficient (minimum variance) estimator of the population standard error.

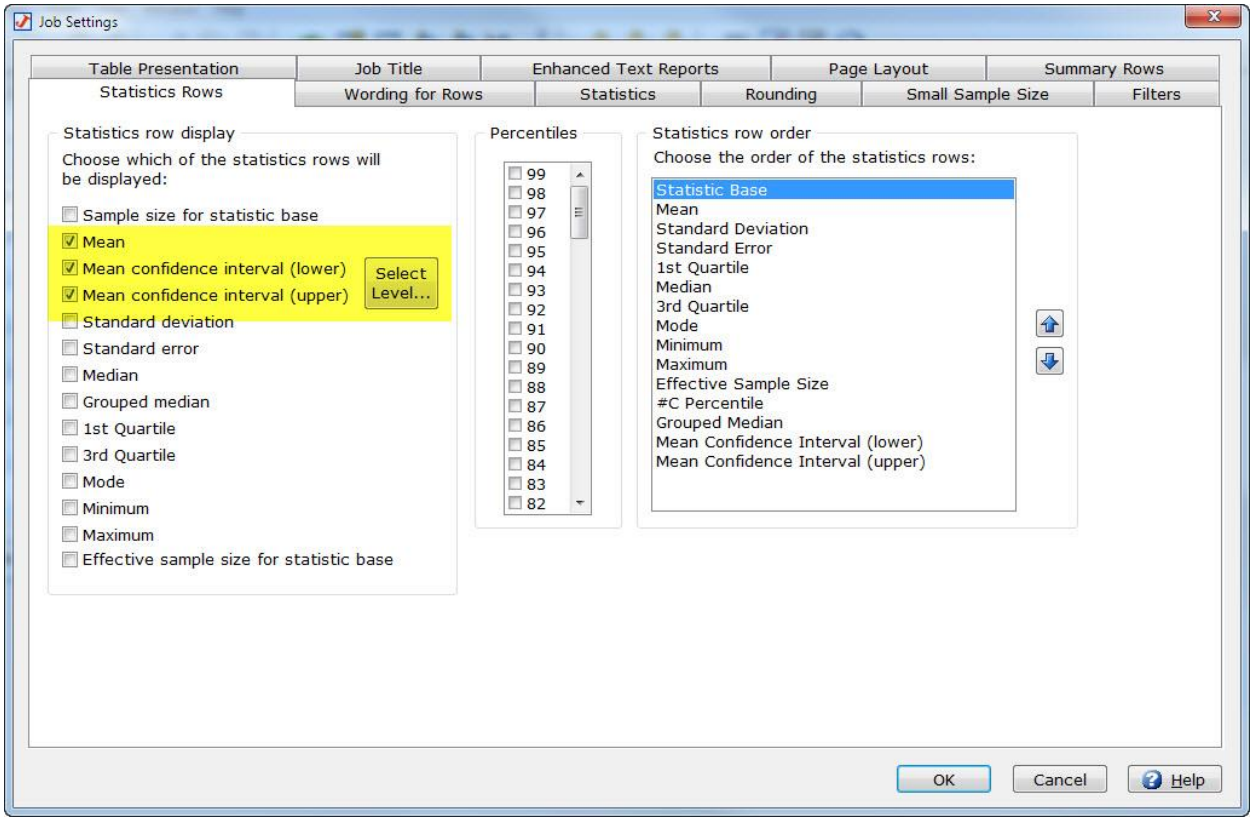
CONFIDENCE INTERVALS FOR MEANS

When a table is completed one may want to exhibit a number of statistics associated with that table. In addition to the various statistics available for selection at a table or row level, WinCross provides a pair of statistics m_1 and m_2, called confidence intervals for the mean, to be displayed in the table. These statistics depend on the choice of a confidence level x, and have the property that in x% of future samples the sample mean will be between m_1 and m_2. The confidence intervals are computed as

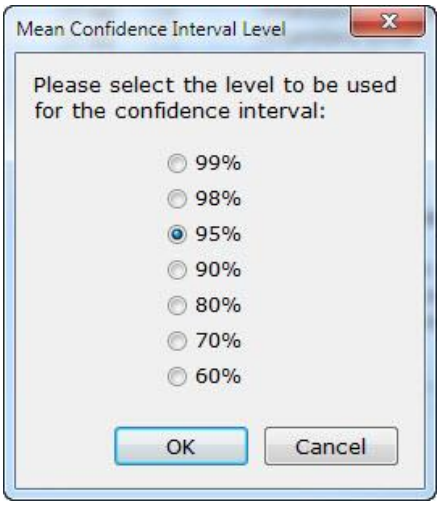
$$m_1 = (\text{sample mean}) - c_x\,(\text{standard error})$$
$$m_2 = (\text{sample mean}) + c_x\,(\text{standard error})$$

where c_x is a factor based on the confidence level x and the sample size (typically, c_x is the two-tailed x-th percentile of the t distribution with appropriate degrees of freedom).

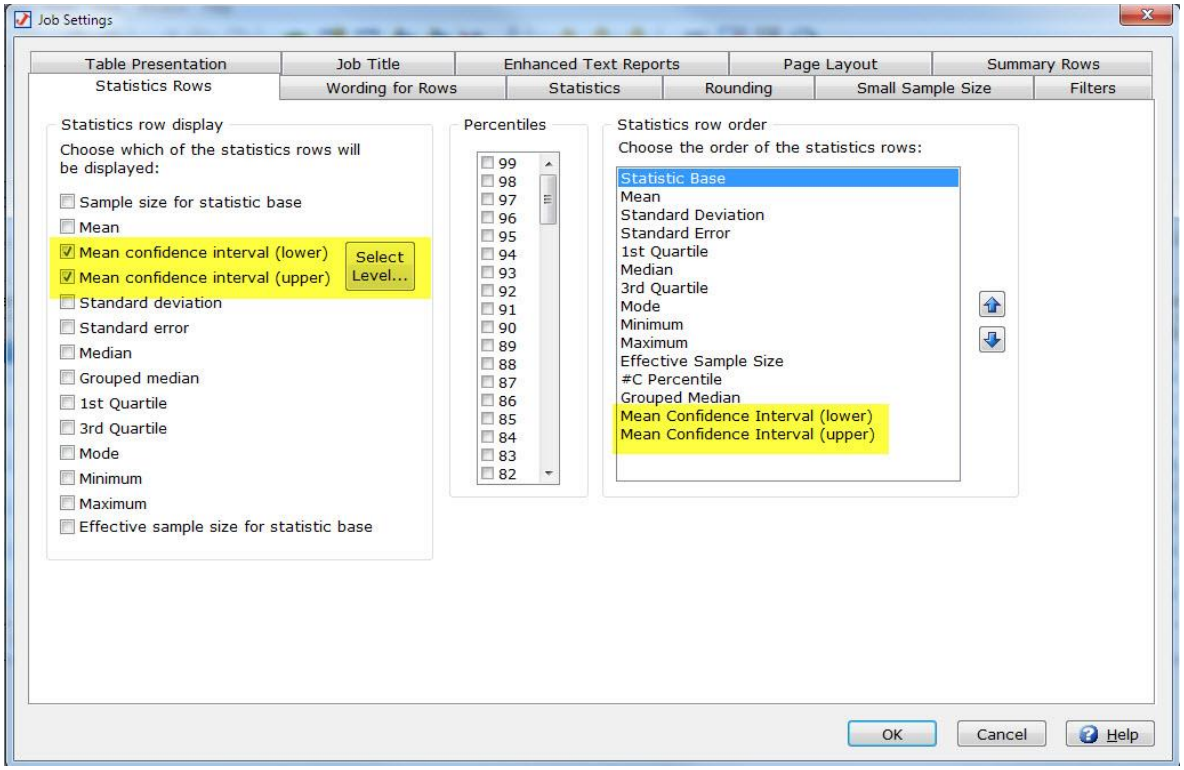
One selects the option of having the confidence interval for the mean calculated for all tables by checking the appropriate boxes in the **Statistics row display** on the **Statistics Rows** tab of the **Job Settings** dialog:



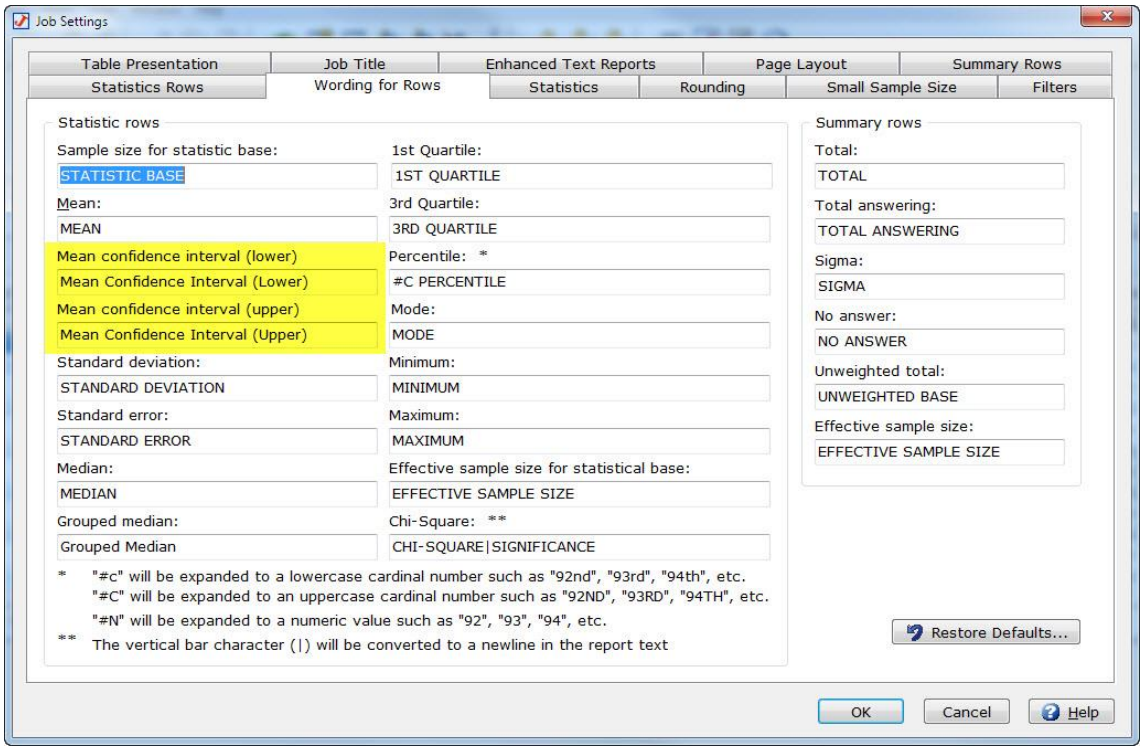
and, one selects the confidence level, first clicking on the **Select Level** button and then choosing the level using the following dialog:



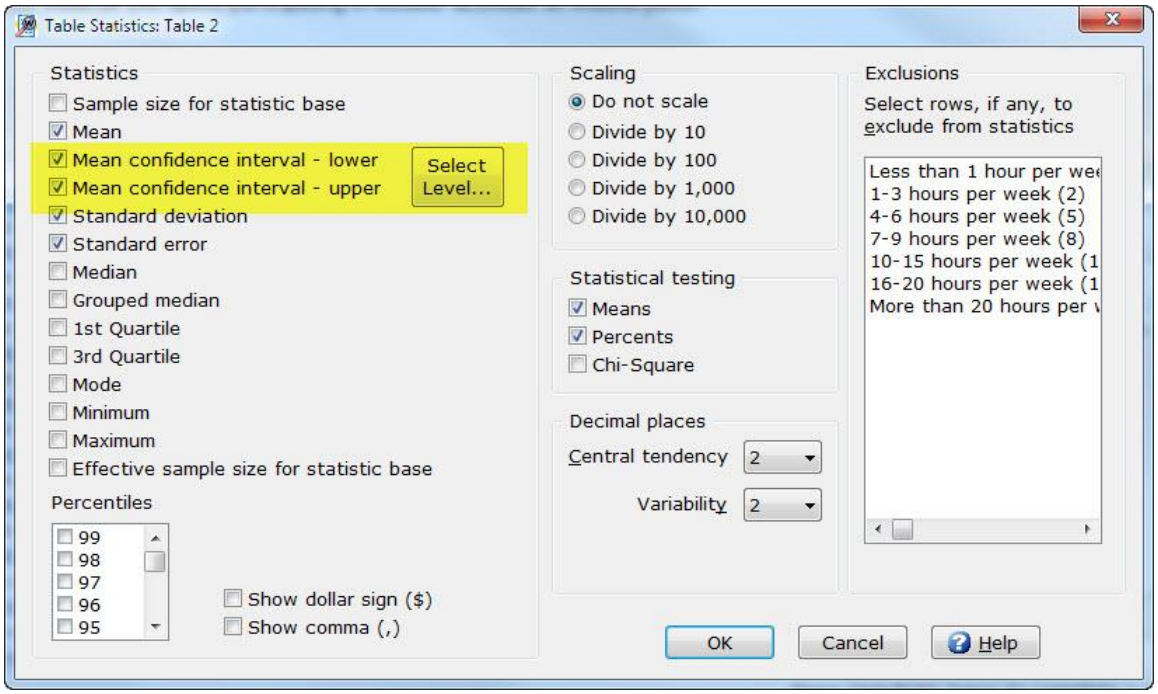
One can also control the order of the roww in the table in which the confidence intervals will appear, by using the arrows in the **Statistics row order** list of the **Statistics row** tab on the **Job Settings** dialog.

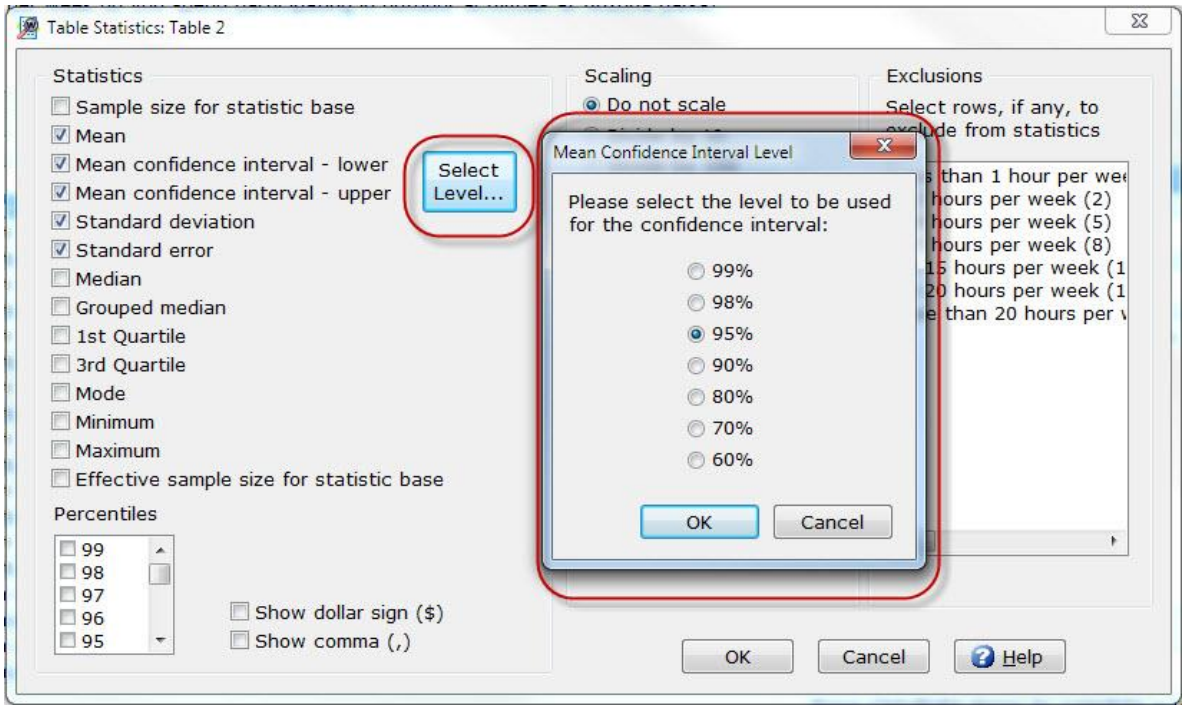


The **Wording for Rows** tab on the **Job Settings** dialog enables one to annotate the rows which contain the confidence intervals with your choice of descriptive text.

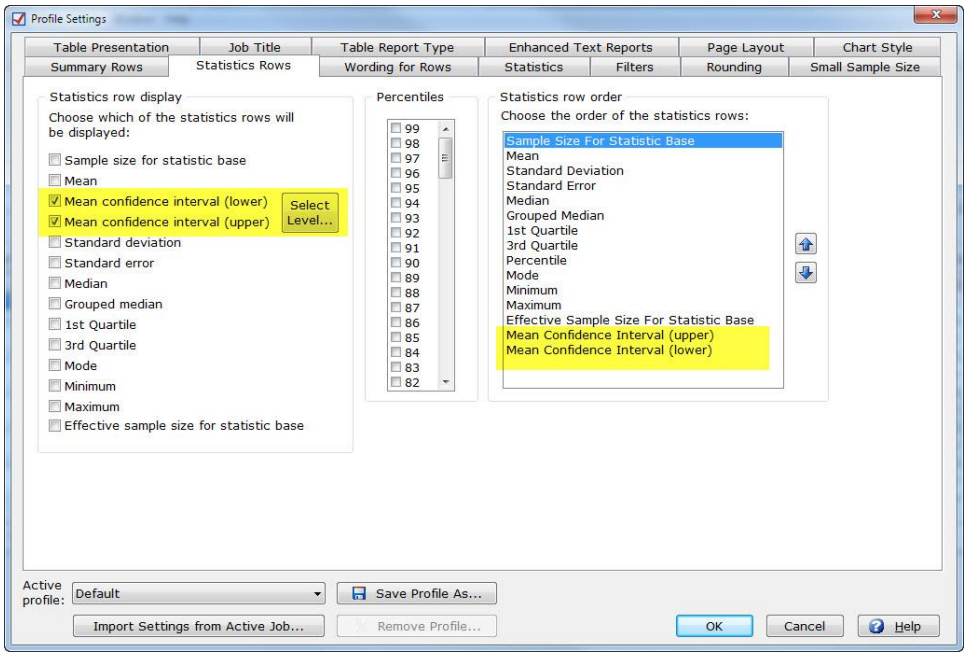


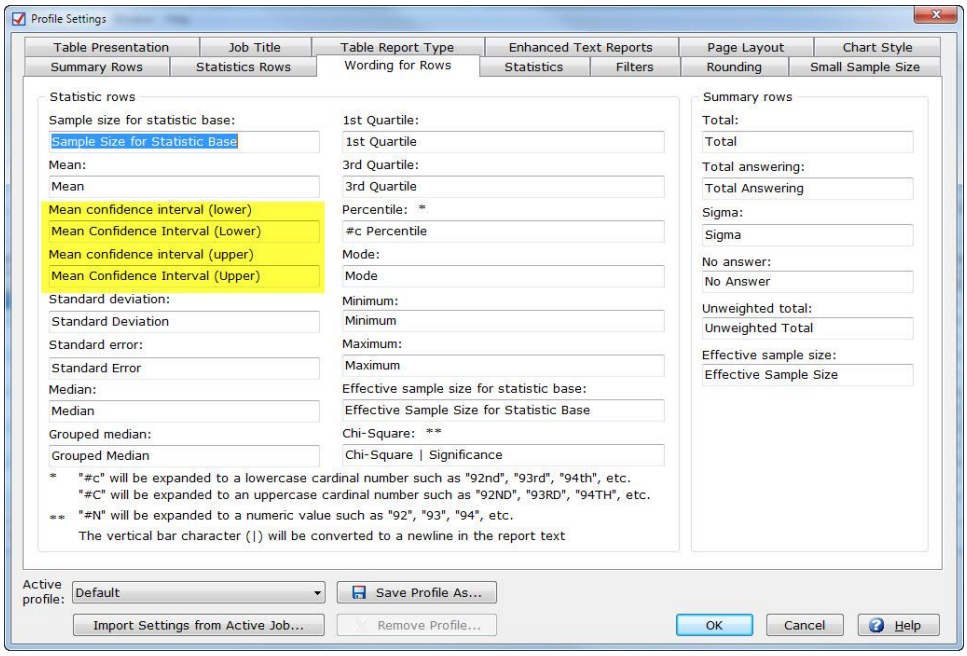
Similarly, one can set up confidence intervals to appear in a specific table by selecting that option from the **Statistics** list in the **Table Statistics** dialog, as illustrated here:



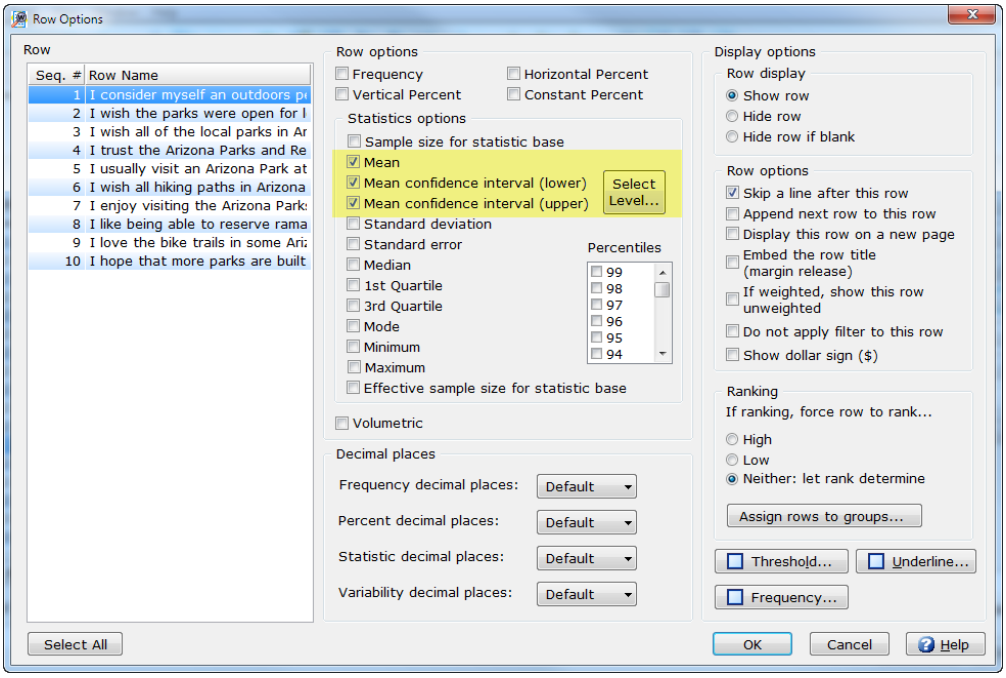


If you know that you will always want to show mean confidence intervals, then you might choose to add these to the profile that you set up when creating a new job. Here is how the **Profile Settings** dialog is to be filled out to enable this feature.



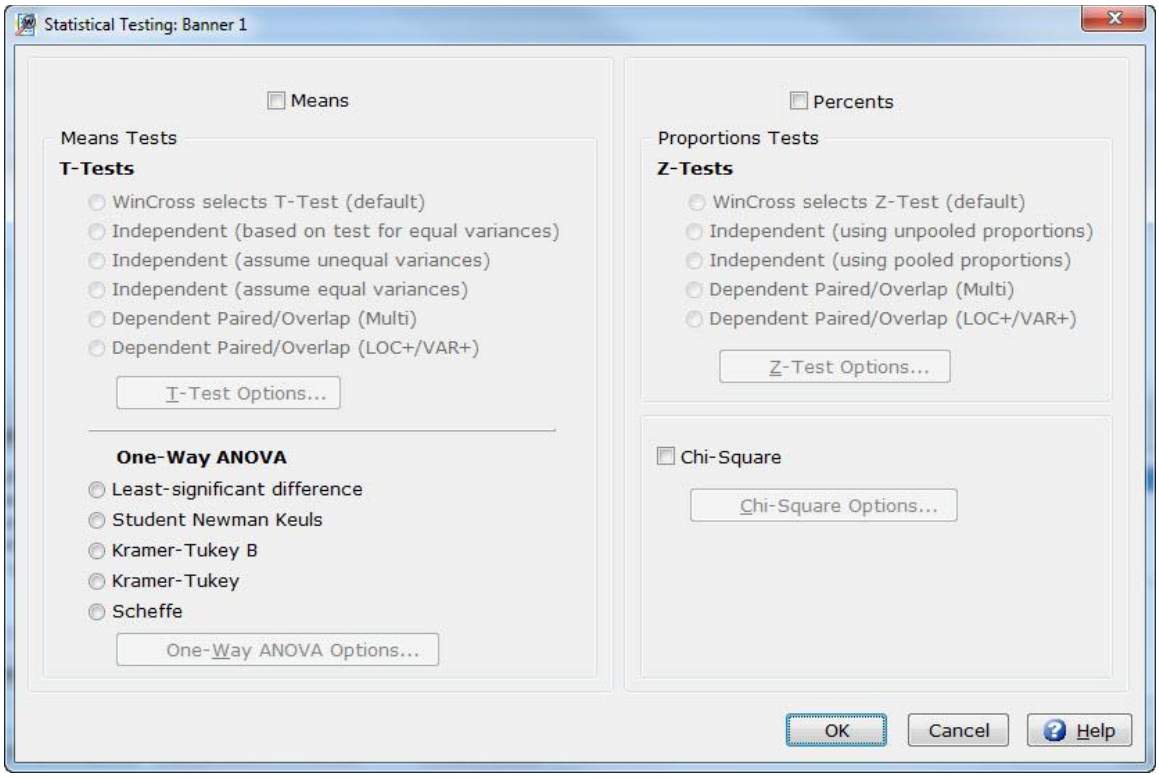


One may also calculate mean confidence intervals at a row level, for example, for each row of a summary of means table. Here is how the **Row Options** dialog is to be filled out to enable this feature.



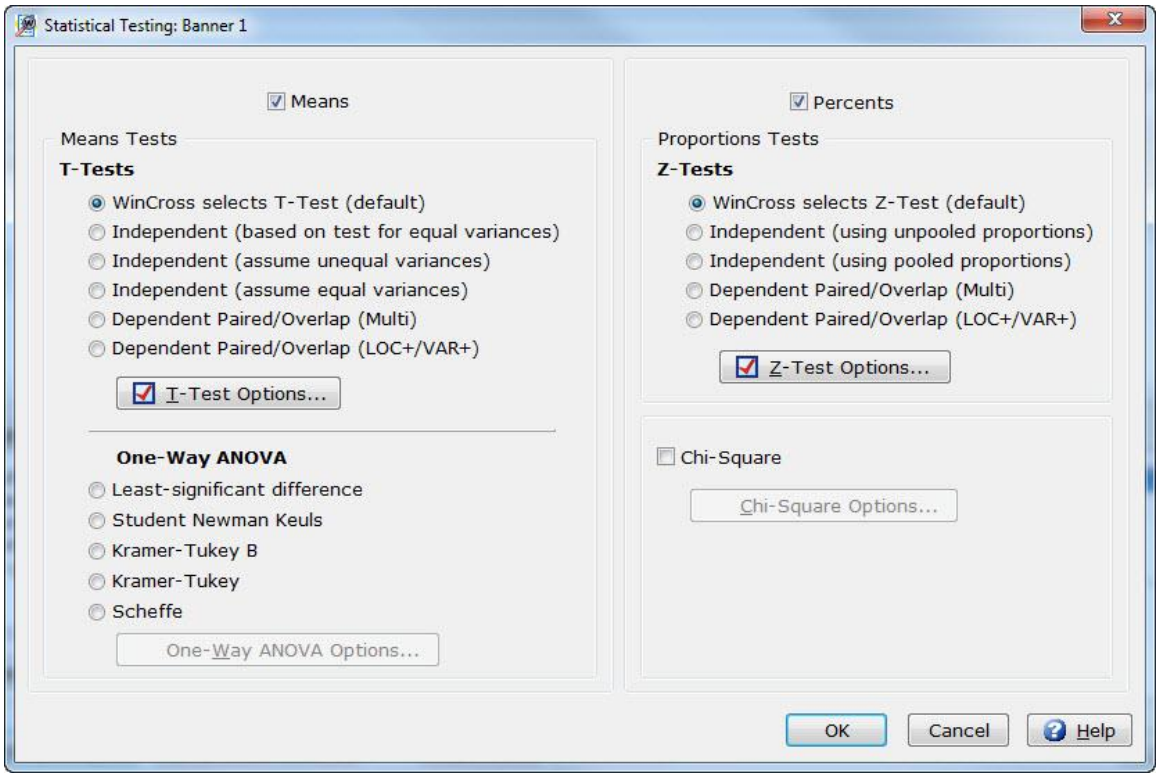
SIGNIFICANCE TESTING

Following is the **Statistical Testing** dialog of WinCross. We will describe the statistical methods underlying each of these items in detail in the sections that follow.



If you want to perform a One-Way ANOVA then you must check the particular form of ANOVA you wish to use. (Detailed description of the various ANOVA methods is given in this manual beginning on page 95.) If you want to perform a Chi-Square test on a table then all you need do is check the Chi-Square box. (Detailed description of the Chi-Square test is given in this manual beginning on page 102.)

Suppose, though, that you want to perform a test on means and/or proportions in the given table. Then, upon clicking the Means and/or Percents box, the **Statistical Testing** dialog looks like this:

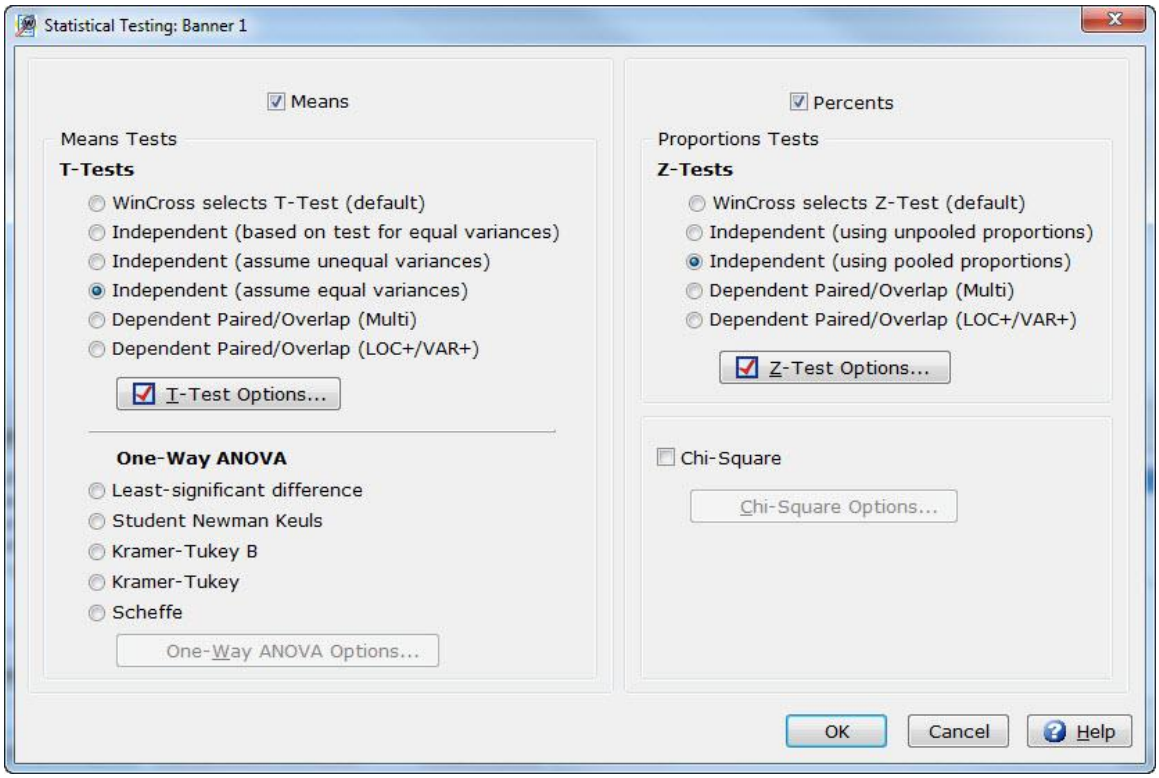


We note here that the first of the test options listed, “WinCross selects T-Test” and “WinCross selects Z-Test” are the “default” options, in that if the user does not check another option, WinCross will determine the appropriate test and perform it. WinCross knows from the structure of the table that a “Dependent Paired” test is to be performed and whether Multi or LOC+/VAR+ is the appropriate test, and so does not need to be informed of this; when this test is called for, the WinCross default test (T or Z) will automatically perform it. (We retain the Multi and LOC+/VAR+ options in case the user wants to select the test.)

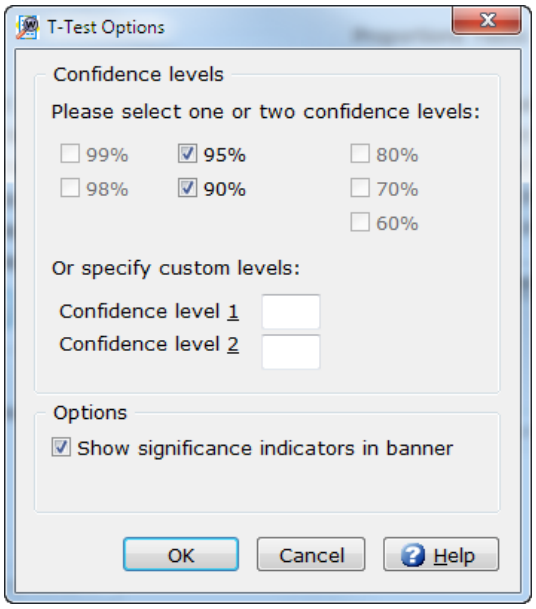
The independent t-test has two variants, one based on the assumption of equal variances and one based on the assumption of unequal variances. In many instances the user does not know whether or not to assume equal variances, and so WinCross has a built-in test which decides which assumption is more based on the sample at hand, and so this option is available to the user. If the user chooses the “default,” i.e., “WinCross selects T-Test” then the default test is the one that assumes unequal variances. The reason for this is that when either of the dependent t-tests is based on independent data then it defaults to the independent t-test based on unequal variances.

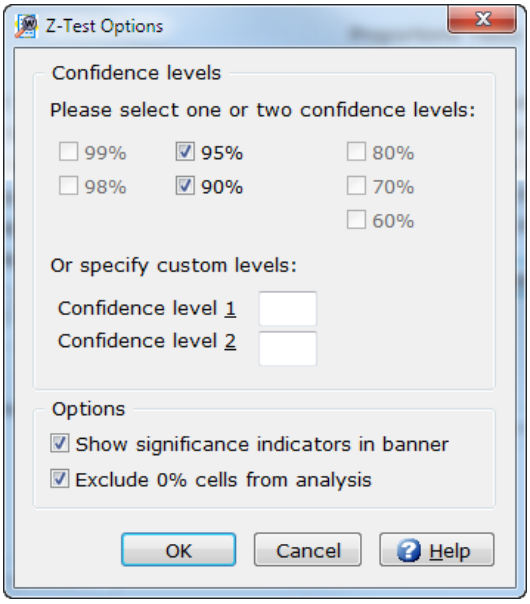
In the case of the z-test, statistical research has shown that the more powerful test is the one that does not pool the proportions of the two samples to estimate the standard error of the difference between the two proportions, and so it is the “default” option. Also, when either of the dependent z-tests is based on independent data then it defaults to the independent z-test based on unpooled proportions.

If one does not use the WinCross default, then one can click on the user-determined test variant, as illustrated below.



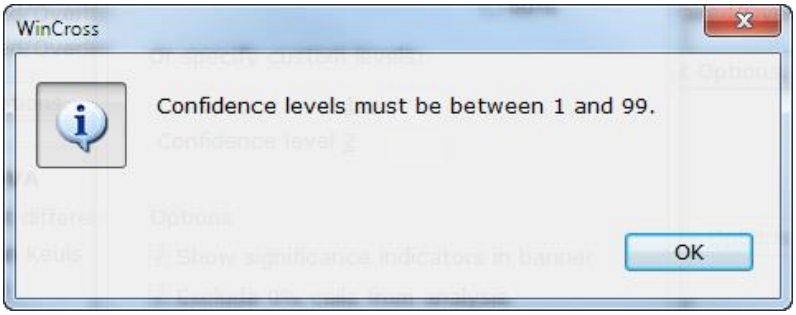
T-TEST OPTONS **Z-TEST OPTIONS**





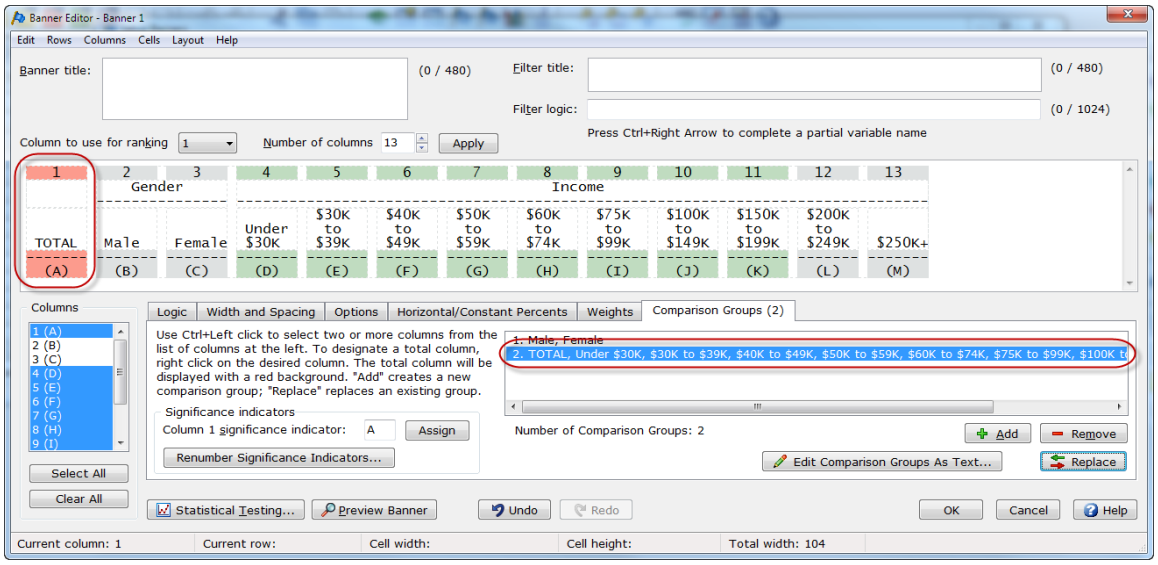
The significance tests are two-tailed tests. In performing either the T-tests or the Z-tests one can select up to two levels of significance, either corresponding to confidence levels of 60%, 70%, 80%, 90%, 95%, 98%, and 99% or any two confidence levels that you specify. If you choose to specify your own confidence level(s), you cannot check one of the preset confidence levels as well. (If you specify your own confidence level(s) WinCross will calculate the corresponding critical value(s) using a Hastings approximation to the t or z percentiles; if you select the preset confidence levels, WinCross will look up the exact critical values in a stored table of t or z percentiles.)

Your specified confidence level(s) must be integers between 1 and 99. Any other specification will lead to the following error window.



Upper or lower case letters under the mean or proportion in a given column indicates the significance between the two columns being compared at either the higher (upper case letters) or lower (lower case letters) level depending on how many confidence levels were selected.

The **Comparison Groups** tab enables one to designate which columns of the table are to be used in the T-tests and/or Z-tests. It also enables one to designate a "Total" column in case you want to perform part-whole comparisons (a description of this test procedure is given below). When m columns are selected, the two sample T or Z tests comparing each of the m(m-1)/2 pairs of designated columns are performed.



There is one caution with respect to using this procedure to separately test each of the m(m-1)/2 pairs of means or proportions. Each time one performs a statistical test there is a probability of making the Type I Error of rejecting the null hypothesis of no difference when in fact there is truly a difference between the means. One normally presets this probability (usually referred to as **α**, the level of significance) at some low level, such as 0.05 or 0.01. If one presets this probability at 0.05, then on average one will make a Type I Error once out of every 20 times one performs a significance test. And if one has m=7 populations and performs m(m-1)/2 = 21 t tests then one will on average reject the hypothesis of no difference when in fact there is no difference between the means being compared. The Oneway anova procedures are designed to circumvent this problem when comparing sets of means.

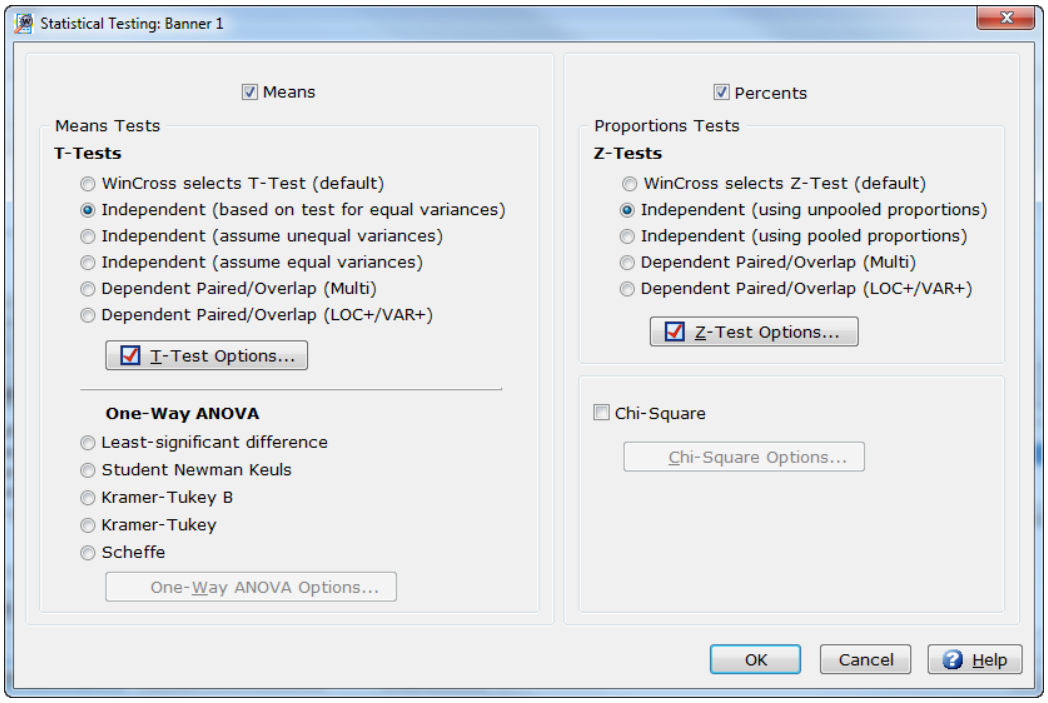
T-TESTS - INDEPENDENT

UNWEIGHTED DATA

General Notation

We consider here the situation in which we have data from two populations, where $\mathbf{n_1}$ is the number of observations in data set 1, $\mathbf{n_2}$ is the number of observations in data set 2, and the data are drawn independently from each of the populations. The means of the two data sets will be designated as $\bar{x}_1$ and $\bar{x}_2$, and the variances of the two data sets will be designated as s_1^2 and s_2^2. The object of this t-test is to test whether the means of the two populations from which the data were drawn are different.

WinCross gives the user the option to determine whether to assume that the variances of the two populations are equal or unequal, and then applies the appropriate test. This is done by selecting either the **Independent (assume equal variances)** or **Independent (assume unequal variances)** option on the **Statistical Testing** dialog. WinCross also gives the user the option to let the program determine, using a preliminary test for equality of variances, which of these two options is appropriate for the data. This is done by selecting the **Independent (based on test for equal variances)** option on the **Statistical Testing** dialog:



Assume equal variances

If one assumes that the two populations have a common variance σ^2, then the best estimate of the common variance is the pooled variance

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The pooled standard error is given by

$$s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

so that the t statistic is

$$t=\frac{\bar{x}_1-\bar{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

This statistic has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Assume unequal variances

If one cannot assume that the two populations have a common variance, then the t statistic is

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

When performing a two sample t test without assuming equality of variances the computation of the number of degrees of freedom is not so straightforward.
The degrees of freedom is given by

$$df_S=\frac{(n_1-1)(n_2-1)}{(n_1-1)(1-c)^2+(n_2-1)c^2}$$

where

$$c=\frac{s_1^2/n_1}{s_1^2/n_1+s_2^2/n_2}$$

Technical Comment:

A Note on Degrees of Freedom

The preferred approach is the Welch approximation[1], developed specifically for the two sample t test. The degrees of freedom of the Welch approximation is given by

$$df_W=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1+1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2+1}}-2$$

[1] B. L. Welch 1938 The Significance of the Difference Between Two Means when the Population Variances are Unequal Biometrika, Vol. 29, No. 3/4 (Feb), pp. 350-362.

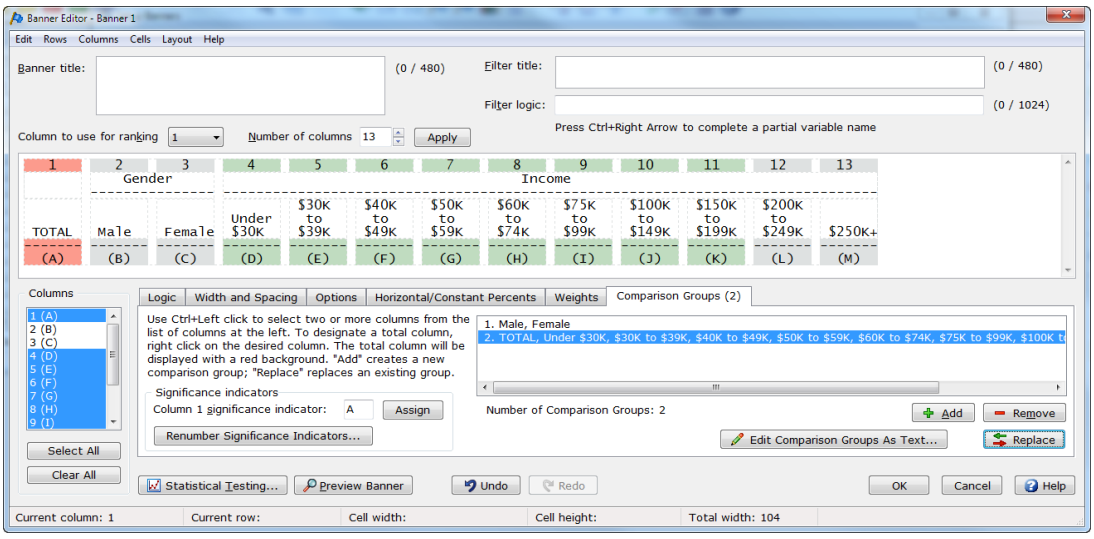
However, SPSS uses a different approximation, the Satterthwaite approximation[2], which is a specialization to the two sample t test of a more general approximation useful in analysis of variance situations. The degrees of freedom of the Satterthwaite approximation is as given on the previous page. Given the widespread use of SPSS, WinCross has adopted the Satterthwaite approximation as the basis for its computation of the degrees of freedom for the two sample t test when equality of variance is not assumed.

Allow WinCross to determine whether variances are equal or not

WinCross performs the F-test for equality of variances to determine whether the population variances are equal or not. The F-test compares the ratio s_1^2 / s_2^2 to the 2.5% point and 97.5% point of the F distribution with n_1-1 and n_2-1 degrees of freedom. If the ratio is within these bounds, WinCross concludes that the variances are equal; if the ratio is either lower than the 2.5% point or higher than the 97.5% point then WinCross concludes that the variances are unequal. WinCross then performs the t test consistent with this determination about the variances.

Part-Whole Comparisons

One sometimes wants to compare the mean $\bar{x}_1$ of a subsample (e.g., a sample from division of a company) with the mean $\bar{x}$ of the full sample (e.g., a sample from the entire company). These means are not independent, and so a special statistical procedure is necessary to implement this comparison. In particular, one has to designate which column of the table contains the totals. WinCross is told that one of the columns being used in a statistical test is a **Total** column by right-clicking on that column in the **Banner Editor**, as in this example:



Let m be the sample size of the subsample and n be the sample size of the full sample. Let s^2 be the sample variance from the full sample.

[2] F. E. Satterthwaite 1946 An Approximate Distribution of Estimates of Variance Components Biometrics Bulletin, Vol. 2, No. 6 (Dec.), pp. 110-114

Assuming equality of variance across the entire population, the proper t statistic for testing whether the subpopulation mean differs from the population mean is

$$t = \frac{\bar{x}_1 - \bar{x}}{s\sqrt{\frac{1}{m} - \frac{1}{n}}}$$

Since the sample variance is based on the complete sample, n-1 is the degrees of freedom for this test.

(If one erroneously used the independent t test one would calculate

$$t = \frac{\bar{x}_1 - \bar{x}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

The denominator of this t statistic is larger than that of the correct t statistic, so that one will be calculating a smaller-than-appropriate test statistic and erroneously saying that the two means are not significantly different when in fact they are.)

If one does not assume equality of variances then WinCross separately calculates the sample variance s_m^2 of the subsample and s_{n-m}^2 of the rest of the n-m observations not included in the subsample. The independent t-test in this case is given by

$$t = \frac{\bar{x}_1 - \bar{x}}{\sqrt{(\frac{1}{n} - \frac{1}{m})^2 m s_m^2 + \frac{n-m}{n^2} s_{n-m}^2}}$$

Using the Satterthwaite approach, the degrees of freedom is given by

$$df_S = \frac{(m-1)(n-m-1)}{(m-1)(1-c)^2 + (n-m-1)c^2}$$

where

$$c = \frac{s_m^2 / m}{s_m^2 / m + s_{n-m}^2 / (n-m)}$$

SINGLY and MULTIPLY WEIGHTED DATA

General Notation

We consider here the situation in which we have data from two populations, where $\mathbf{n_1}$ is the number of observations in data set 1, $\mathbf{n_2}$ is the number of observations in data set 2, and the data are drawn independently from each of the populations. The weighted means of the two data sets will be designated as $\bar{x}_{1w}$ and $\bar{x}_{2w}$. These means may be calculated using a single weight for the observations from the two populations or separate weights

applied to the data from each of the populations. The unweighted variances of the two data sets will be designated as s_1^2 and s_2^2.

Assume equal variances

When the samples are weighted, the best estimate for the pooled standard error is based on the unweighted pooled variance given above, and is given by

$$s\sqrt{\frac{1}{e_1}+\frac{1}{e_2}}$$

where e_1 and e_2 are the effective sample sizes of the two samples and

$$s=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

The t statistic is then

$$t=\frac{\bar{x}_{1w}-\bar{x}_{2w}}{s\sqrt{\frac{1}{e_1}+\frac{1}{e_2}}}$$

This statistic has a t-distribution with $n_1 + n_2 – 2$ degrees of freedom.

Assume unequal variances

If one cannot assume that the two populations have a common variance, then the t statistic is once again based on a standard error calculated from the unweighted sample variances

$$t=\frac{\bar{x}_{1w}-\bar{x}_{2w}}{\sqrt{\frac{s_1^2}{e_1}+\frac{s_2^2}{e_2}}}$$

where e_1 and e_2 are the effective sample sizes of the two samples.

The degrees of freedom, based on the Satterthwaite approximation, is given by

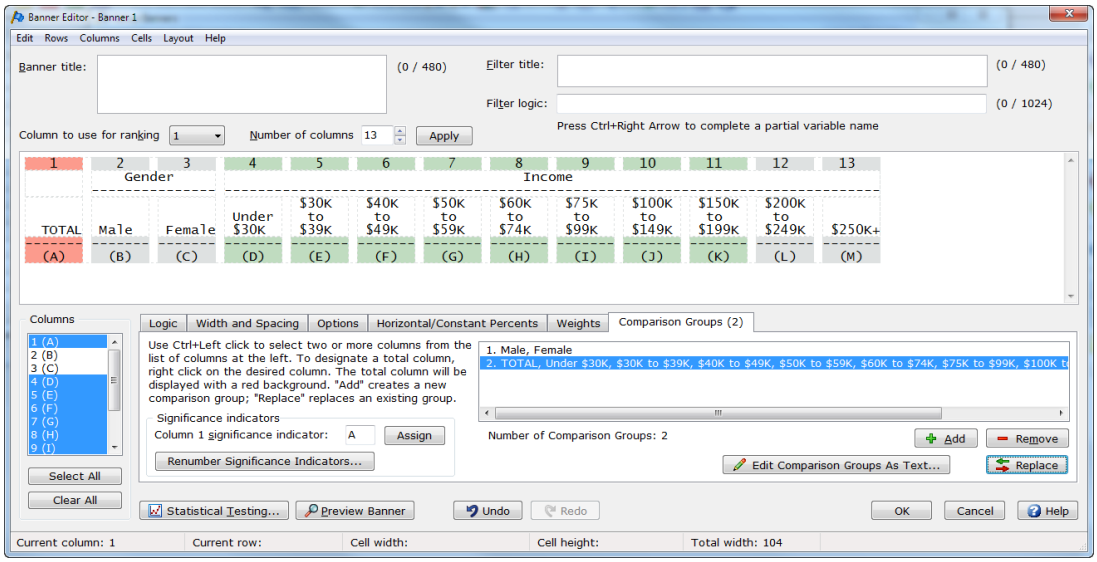
$$df_S=\frac{(n_1-1)(n_2-1)}{(n_1-1)(1-c)^2+(n_2-1)c^2}$$

where

$$c=\frac{s_1^2/n_1}{s_1^2/n_1+s_2^2/n_2}$$

Part-Whole Comparisons

One sometimes wants to compare the weighted mean $\bar{x}_{1w}$ of a subsample (e.g., a sample from division of a company) with the weighted mean $\bar{x}_w$ of the full sample (e.g., a sample from the entire company). These means are not independent, and so a special statistical procedure is necessary to implement this comparison. ***WinCross only applies the part-whole comparison test when a single weight is applied to all of the observations***. In particular, one has to designate which column of the table contains the totals. WinCross is told that one of the columns being used in a statistical test is a **Total** column by right-clicking on that column in the **Banner Editor**, as in this example:



Assuming equality of variance across the entire population, the proper t statistic for testing whether the subpopulation mean differs from the population mean is

$$t = \frac{\bar{x}_{1w} - \bar{x}_w}{s\sqrt{\frac{1}{e_1} - \frac{1}{e}}}$$

where $\bar{x}_{1w}$ is the weighted mean of the subsample, $\bar{x}_w$ is the weighted mean of the whole sample, s is the unweighted standard deviation of the whole sample, e_1 is the effective sample size of the subsample, and e is the effective sample size of the whole sample. This statistic has a t-distribution with n – 1 degrees of freedom.

Assume unequal variances, if one cannot assume that the two populations have a common variance, then, the t statistic is once again based on the standard errors calculated from the unweighted sample variances

$$t = \frac{\bar{x}_{1w} - \bar{x}_w}{\sqrt{(\frac{1}{e_1} - \frac{1}{e})e_1 s_m^2 + \frac{e - e_1}{e^2} s_{n-m}^2}}$$

where s_m^2 is the sample variance of the subsample, s_{n-m}^2 is the sample variance of the rest of the n-m observations not included in the subsample, e_1 is the effective sample size of the subsample and e is the effective sample sizes of the full sample.

Using the Satterthwaite approach, the degrees of freedom is given by

$$df_S = \frac{(m-1)(n-m-1)}{(m-1)(1-c)^2 + (n-m-1)c^2}$$

where

$$c = \frac{s_m^2 / m}{s_m^2 / m + s_{n-m}^2 / (n-m)}$$

T-TESTS - DEPENDENT PAIRED/OVERLAP (LOC+/VAR+)

Terminology

WinCross uses the terms LOC+, VAR+ and MULTI as shorthand for describing the contexts in which one applies statistical tests to a pair of columns in a table wherein the observations across columns are correlated. For the WinCross descriptions of the use of these terms, see the *WinCross Online Help*. We describe the statistical basis for each of these contexts in the **General Notation** sections of this manual.

General Notation

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_2 independent observations $x_{21},...,x_{2n_2}$ from population 2. Suppose further that the first n_0 observations from the two populations are paired (for example, population 1 is a "treatment," population 2 is a "control," and the first n_0 observations are taken from the same respondent; for another example, population 1 is ratings of Coke, population 2 is ratings of Pepsi, and the first n_0 pairs of ratings are taken from the same respondent).

The two sample means are

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad \bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

The two sample variances are given by

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \quad s_2^2 = \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

The object of this t-test is to test whether the means of the two populations from which the data were drawn are different.

UNWEIGHTED DATA

t-Test for Means with Partial Pairing

Because there are n_0 pairs of observations $(x_{11}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{1n_0}, x_{2n_{2_1}})$ that are correlated, we must calculate the covariance between the sample means as part of the standard error computation. WinCross calculates the sample covariance between the two sets of paired observations as

$$c = \frac{\sum_{i=1}^{n_0}(x_{1i} - \bar{x}_{10})(x_{2i} - \bar{x}_{20})}{n_0 - 1}$$

$$= \frac{\sum_{i=1}^{n_0} x_{1i}x_{2i} - n_0\bar{x}_{10}\bar{x}_{20}}{n_0 - 1}$$

where $\bar{x}_{10}$ is the mean of the first n_0 observations on population 1 and $\bar{x}_{20}$ is the mean of the first n_0 observations on population 2. This uses only the means of the n_0 paired observations in the computation, and produces an unbiased estimated of the population covariance. However, it does not use the full set of data to estimate the means of the two populations.

The variance of the difference between the two sample means is given by

$$V\bar{x}_1 + V\bar{x}_2 - 2Cov(\bar{x}_1, \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - \frac{2}{n_1 n_2} Cov(\sum_{i=1}^{n_1} x_{1i}, \sum_{i=1}^{n_2} x_{2i})$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - \frac{2n_0 Cov(x_1, x_2)}{n_1 n_2}$$

The variance of the difference between the two sample means is estimated by

$$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - \frac{2n_0 c}{n_1 n_2}$$

The t-statistic to test the difference between the two means is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - \frac{2n_0 c}{n_1 n_2}}}$$

The degrees of freedom computation is in two parts. The first part is an application of the Satterthwaite approximation to the sample sizes of the unique observations from the two populations, and is given by

$$\frac{(n_1 - n_0 - 1)(n_2 - n_0 - 1)}{(n_1 - n_0 - 1)(1-c)^2 + (n_2 - n_0 - 1)c^2}$$

where

$$c = \frac{s_1^2/(n_1 - n_0)}{s_1^2/(n_1 - n_0) + s_2^2/(n_2 - n_0)}$$

(This only applies if there are two or more observations in each of the sets of observations from the two populations. **If n_i is 1 or less then s_i^2 cannot be calculated (i=1 or 2), and the test is not performed**.) The second part is just n_0-1, the degrees of freedom for the overlap set of observations. The degrees of freedom are given by the sum of these component parts, namely

$$df = \frac{(n_1 - n_0 - 1)(n_2 - n_0 - 1)}{(n_1 - n_0 - 1)(1-c)^2 + (n_2 - n_0 - 1)c^2} + (n_0 - 1)$$

Thus if there is perfect pairing then $n_1 = n_2 = n_0$, and the first term is not to be calculated. And **if n_0=0** the degrees of freedom are those of the Satterthwaite formula in the two independent sample comparison, and **the test reduces to the independent t test with <u>unequal</u> variances.**

As noted in the document A NOTE ON SPURIOUS SIGNIFICANCE on our web site, there is the possibility of spuriously finding "significant" differences due only because of the degree of overlap of the two samples. WinCross has adopted the safeguard of declaring all such differences not significant if a factor, described in that document, based on the fraction of unique observations from population 1 and from population 2 is less than 5%.

<u>Technical Comment</u>:

On Calculating Covariances

There are a number of other ways of calculating the sample covariance between the two sets of paired observations. One such is the following:
The population covariance between two variables u and v is defined as

$$Cov(x_1, x_2) = E(x_1 - Ex_1)(x_2 - Ex_2),$$

where E denotes the expected value operation. This can equivalently be expressed as

$$Cov(x_1, x_2) = Ex_1x_2 - Ex_1Ex_2$$

Thus, to estimate $Cov(x_1, x_2)$ one might use the best estimates of Ex_1x_2, Ex_1, and Ex_2 in the computation. The best estimate of Ex_1x_2 is the mean of the products of the $x_1$and x_2 across the n_0 observations where we have data on both of these variables. The best estimate of Ex_1 is the mean of <u>all</u> the x_1; the best estimate of Ex_2 is the mean of <u>all</u> the x_2. Putting all this together we obtain as an estimate of the sample covariance between the two sets of paired observations

$$c^* = \frac{\sum_{i=1}^{n_0} x_{1i} x_{2i}}{n_0} - \bar{x}_1 \bar{x}_2$$

$$= \frac{\sum_{i=1}^{n_0} x_{1i} x_{2i} - n_0 \bar{x}_1 \bar{x}_2}{n_0}$$

Unfortunately, this is not an unbiased estimate of the population covariance and the unbiasing factor is quite complex.

If we were to use the form $Cov(x_1, x_2)=E(x_1-Ex_1)(x_2-Ex_2)$ as the template for building our estimate, we would be led to the following computation of the sample covariance between the two sets of paired observations:

$$\tilde{c} = \frac{\sum_{i=1}^{n_0} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{n_0 - 1}$$

$$= \frac{\sum_{i=1}^{n_0} x_{1i} x_{2i} - n_0 (\bar{x}_{20} \bar{x}_1 + \bar{x}_{10} \bar{x}_2 - \bar{x}_1 \bar{x}_2)}{n_0 - 1}$$

This estimate requires the extra computation of these means, and is also not unbiased, and therefore is not recommended.

Technical Comment:

A Note on Perfect Pairing

In the case where $n_1 = n_2 = n_0 = n$, say, i.e., when all the observations are paired, all these computations simplify considerably. Indeed, there is no need to calculate the covariance, for, letting $d_i = x_{1i} - x_{2i}$, we see that

$$\bar{x}_1 - \bar{x}_2 = \bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

Given this, the standard deviation of the differences between the paired observations is given by

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

so that the t-statistic to test the difference between the two means is given by

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

and it has a t distribution with n-1 degrees of freedom.

Part-Whole Comparisons

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_j independent observations $x_{j1},...,x_{jn_j}$ from population j, j=2,…,m. We want to compare the mean of population 1 with the mean across all m populations.

The two means are

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad \bar{x}_T = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{ji}}{\sum_{j=1}^{m} n_j}$$

For each of the m-1 pairs of observations (x_{1i}, x_{ji}) i=1,..., n, j=2, …, m there are n_{0j} that are paired (for example, population 1 is ratings of Coke, population 2 is ratings of Pepsi, population 3 is ratings of Seven-Up, and there are n_{02} sets of ratings from the same respondent for Coke and Pepsi and n_{03} sets of ratings from the same respondent for Coke and Seven-Up). The two sample variances are given by

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \quad s_T^2 = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_T)^2}{\sum_{j=1}^{m} n_j - 1}$$

For each of the m-1 pairs of observations (x_{1i}, x_{ji}) i=1,..., n, j=2, …, m are correlated, we must calculate the covariance between the sample means as part of the standard error computation. WinCross calculates the sample covariance between the two sets of paired observations as

$$c_j = \frac{\sum_{i=1}^{n_{0j}}(x_{1i} - \bar{x}_{10j})(x_{ji} - \bar{x}_{j01})}{n_{10j} - 1}$$

where $\bar{x}_{10j}$ is the mean of item 1 and $\bar{x}_{j01}$ is the mean of the first n_0 observations on population j. This uses only the means of item j from the n_{10j} observations from respondents who answered both item 1 and item j

The variance of the difference between the two sample means is given by

$$V\bar{x}_1 + V\bar{x}_T - 2Cov(\bar{x}_1, \bar{x}_T)$$
$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_T^2}{n_T} - \frac{2}{n_1 n_T} \sum_{j=1}^{m} Cov(\sum_{i=1}^{n_1} x_{1i}, \sum_{i=1}^{n_j} x_{ji})$$
$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_T^2}{n_T} - \frac{2\sum_{j=1}^{m} n_{10j} Cov(x_1, x_j)}{n_1 n_T}$$

The variance of the difference between the two sample means is estimated by

$$\frac{s_1^2}{n_1} + \frac{s_T^2}{n_T} - \frac{2c}{n_1 n_T}$$

where

$$c = \sum_{j=1}^{m} n_{10j} c_j$$

The t-statistic to test the difference between the two means is given by

$$t = \frac{\bar{x}_1 - \bar{x}_T}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_T^2}{n_T} - \frac{2c}{n_1 n_T}}}$$

The degrees of freedom computation is made complicated by the fact that n_T is not reflective of the sample sizes used in calculating the covariances. The total set of items which are paired with column 1 is given by

$$n_c = \sum_{j=2}^{m} n_{10j}$$

We apply the Satterthwaite approximation to n_1 and n_c to obtain the degrees of freedom of this test, using

$$df_S = \frac{(n_1 - 1)(n_c - 1)}{(n_1 - 1)(1 - g)^2 + (n_c - 1)g^2}$$

where

$$g = \frac{s_1^2 / n_1}{s_1^2 / n_1 + s_T^2 / n_c}$$

As noted in the document A NOTE ON SPURIOUS SIGNIFICANCE on our web site, there is the possibility of spuriously finding "significant" differences due only because of the degree of overlap of the part to the whole, WinCross has adopted the safeguard of declaring all such differences not significant if the fraction of the part to the whole is less than 5% or greater than 95%.

SINGLY WEIGHTED DATA

General Notation

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_2 independent observations $x_{21},...,x_{2n_2}$ from population 2. Suppose further that the first n_0 observations from the two populations are paired (e.g., population 1 is a "treatment", population 2 is a "control," and the first n_0 observations are taken from the same respondent). Finally, suppose that each of the respondents has an associated weight, with $w_{11},...,w_{1n_1}$ the weights for the respondents from population 1, $w_{21},...,w_{2n_2}$ the weights for the respondents from population 2, and where the weights applied to each of the observations on the first n_0 respondents are identical for both observations, i.e., $w_{11} = w_{21} = w_1,...,w_{1n_0} = w_{2n_0} = w_{n_0}$.

The two sample means are

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad \bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

The two weighted sample means are

$$\bar{x}_{1w} = \frac{\sum_{i=1}^{n_1} w_{i1} x_{1i}}{\sum_{i=1}^{n_1} w_{1i}}, \bar{x}_{2w} = \frac{\sum_{i=1}^{n_2} w_{2i} x_{2i}}{\sum_{i=1}^{n_2} w_{2i}}$$

The two unweighted sample variances are given by

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \quad s_2^2 = \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

t-Test for Means with Partial Pairing

The unweighted sample covariance between the two sets of paired observations is given by

$$c = \frac{\sum_{i=1}^{n_0} (x_{1i} - \bar{x}_{10})(x_{2i} - \bar{x}_{20})}{n_0 - 1}$$

$$= \frac{\sum_{i=1}^{n_0} x_{1i} x_{2i} - n_0 \bar{x}_{10} \bar{x}_{20}}{n_0 - 1}$$

where $\bar{x}_{10}$ is the mean of the first n_0 observations on population 1 and $\bar{x}_{20}$ is the mean of the first n_0 observations on population 2.

The best estimate of the variance of the difference between the two sample weighted means is given by

$$\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0 c}{e_1 e_2},$$

where e_1 and e_2 are the effective sample sizes for the samples from populations 1 and 2, namely

$$e_1 = \frac{(\sum_{i=1}^{n_1} w_{1i})^2}{\sum_{i=1}^{n_1} w_{1i}^2}, \quad e_2 = \frac{(\sum_{i=1}^{n_2} w_{2i})^2}{\sum_{i=1}^{n_2} w_{2i}^2}$$

and e_0 is the effective sample size for the observations common to populations 1 and 2, namely

$$e_0 = \frac{(\sum_{i=1}^{n_0} w_i)^2}{\sum_{i=1}^{n_0} w_i^2}$$

The t-statistic to test the difference between the two means is given by

$$t = \frac{\bar{x}_{1w} - \bar{x}_{2w}}{\sqrt{\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0 c}{e_1 e_2}}}$$

The degrees of freedom computation is in two parts. The first part is an application of the Satterthwaite approximation to the sample sizes of the unique observations from the two populations, and is given by

$$\frac{(n_1 - n_0 - 1)(n_2 - n_0 - 1)}{(n_1 - n_0 - 1)(1-c)^2 + (n_2 - n_0 - 1)c^2}$$

where

$$c = \frac{s_1^2/(n_1 - n_0)}{s_1^2/(n_1 - n_0) + s_2^2/(n_2 - n_0)}$$

(This only applies if there are two or more observations in each of the sets of unique observations from the two populations. **If n_i is 1 or less then s_i^2 cannot be calculated (i=1 or 2), and the test is not performed.**) The second part is just n_0-1, the degrees of freedom for the overlap set of observations. The degrees of freedom are given by the sum of these component parts, namely

$$df = \frac{(n_1 - n_0 - 1)(n_2 - n_0 - 1)}{(n_1 - n_0 - 1)(1-c)^2 + (n_2 - n_0 - 1)c^2} + (n_0 - 1)$$

Thus if there is perfect pairing then $n_1 = n_2 = n_0$, and the first term is not to be calculated. And **if n_0=0** the degrees of freedom are those of the Satterthwaite formula in the two

independent sample comparison and **the test reduces to the independent t test with <u>unequal</u> variances.**

<u>Technical Comment</u>:

A Note on Perfect Pairing

In the case where $n_1 = n_2 = n_0 = n$, say, i.e., when all the observations are paired, all these computations simplify considerably. Indeed, there is no need to calculate the covariance, for, letting $d_i = x_{1i} - x_{2i}$, we see that

$$\bar{x}_{1w} - \bar{x}_{2w} = \bar{d}_w = \frac{\sum_{i=1}^{n} w_i d_i}{\sum_{i=1}^{n} w_i}$$

Given this, the unweighted standard deviation of the differences between the paired observations is given by

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

so that the t-statistic to test the difference between the two means is given by

$$t = \frac{\bar{d}_w}{s_d / \sqrt{e}},$$

where the effective sample size e is given by

$$e = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$$

This statistic has a t distribution with n-1 degrees of freedom.

Part-Whole Comparisons

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_j independent observations $x_{j1},...,x_{jnj_2}$ from population j, j=2,…,m. Suppose further that the first n_0 observations from each of the populations are paired (e.g., the first n_0 observations are taken from the same respondent). Finally, suppose that each of the respondents has an associated weight, with $w_1,...,w_n$ and where the weights applied to each of the observations on the first n_0 respondents are identical for all the observations, i.e., $w_{11} = w_{21} = w_1,...,w_{1n_0} = w_{2n_0} = w_{n_0}$.

The two sample means are

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad \bar{x}_T = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{ji}}{\sum_{j=1}^{m} n_j}$$

The two weighted sample means are

$$\bar{x}_{1w} = \frac{\sum_{i=1}^{n_1} w_{i1} x_{1i}}{\sum_{i=1}^{n_1} w_{1i}}, \bar{x}_{Tw} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} w_i x_{ji}}{\sum_{j=1}^{m} \sum_{i=1}^{n_j} w_i}$$

The two unweighted sample variances are given by

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \quad s_T^2 = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_T)^2}{\sum_{j=1}^{m} n_j - 1}$$

As above, we use the unweighted sample covariance in calculating the variance of the difference between the two means. This is given by

$$c_j = \frac{\sum_{i=1}^{n_{0j}} (x_{1i} - \bar{x}_{10j})(x_{ji} - \bar{x}_{j01})}{n_{10j} - 1}$$

where $\bar{x}_{10j}$ is the mean of item 1 and $\bar{x}_{j01}$ is the mean of the first n_0 observations on population j. This uses only the means of item j from the n_{10j} observations from respondents who answered both item 1 and item j

The variance of the difference between the two sample means is given by

$$\begin{aligned} & V\bar{x}_{1w} + V\bar{x}_{Tw} - 2Cov(\bar{x}_{1w}, \bar{x}_{Tw}) \\ &= \frac{\sigma_1^2}{e_1} + \frac{\sigma_T^2}{e_T} - \frac{2}{(\sum_{j=1}^{n} w_{1j})(\sum_{j=1}^{m} \sum_{i=1}^{n_j} w_i)} \sum_{j=1}^{m} Cov(\sum_{i=1}^{n_1} w_{1i} x_{1i}, \sum_{i=1}^{n_j} w_{ji} x_{ji}) \\ &= \frac{\sigma_1^2}{e_1} + \frac{\sigma_T^2}{e_T} - \frac{2\sum_{j=1}^{m} Cov(x_1, x_j) \sum_{i=1}^{n_j} w_{10ji}^2}{(\sum_{j=1}^{n} w_{1j})(\sum_{j=1}^{m} \sum_{i=1}^{n_j} w_i)} \end{aligned}$$

where w_{10ji}^2 is the square of the weight for the i-th respondent who answered both questions 1 and j.

The best estimate of the variance of the difference between the two sample weighted means is given by

$$\frac{s_1^2}{e_1} + \frac{s_T^2}{e_T} - \frac{2c}{e_1 e_T},$$

where

$$c = \sum_{j=1}^{m} c_j \sum_{i=1}^{n_j} w_{10\,ji}^2$$

and where e_1 and e_T are the effective sample sizes for the samples from populations 1 and the set of m populations, namely

$$e_1 = \frac{(\sum_{i=1}^{n_1} w_{1i})^2}{\sum_{i=1}^{n_1} w_{1i}^2}, \quad e_T = \frac{(\sum_{j=1}^{m} \sum_{i=1}^{n_2} w_i)^2}{\sum_{j=1}^{m} \sum_{i=1}^{n_2} w_i^2}$$

The t-statistic to test the difference between the two means is given by

$$t = \frac{\bar{x}_{1w} - \bar{x}_{Tw}}{\sqrt{\frac{s_1^2}{e_1} + \frac{s_T^2}{e_T} - \frac{2c}{e_1 e_T}}}$$

The degrees of freedom computation is made complicated by the fact that n_T is not reflective of the sample sizes used in calculating the covariances. The total set of items which are paired with column 1 is given by

$$n_c = \sum_{j=2}^{m} n_{10\,j}$$

We apply the Satterthwaite approximation to n_1 and n_c to obtain the degrees of freedom of this test, using

$$df_S = \frac{(n_1 - 1)(n_c - 1)}{(n_1 - 1)(1 - g)^2 + (n_c - 1)g^2}$$

where

$$g = \frac{s_1^2 / n_1}{s_1^2 / n_1 + s_T^2 / n_c}$$

MULTIPLY WEIGHTED DATA

General Notation

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_2 independent observations $x_{21},...,x_{2n_2}$ from population 2. Suppose further that the first n_0 observations from the two populations are paired (e.g., population 1 is a "treatment", population 2 is a "control," and the first n_0 observations are taken from the same respondent). Finally, suppose that each of the respondents has an associated weight, with $w_{11},...,w_{1n_1}$ the weights for the respondents from population 1, $w_{21},...,w_{2n_2}$ the weights for the respondents from population 2, and where the weights applied to each of the observations on the first n_0 respondents are not necessarily identical, i.e., $w_{11} \neq w_{21},...,w_{1n_0} \neq w_{2n_0}$.

The two sample means are

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad \bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

The two weighted sample means are

$$\bar{x}_{1w} = \frac{\sum_{i=1}^{n_1} w_{1i} x_{1i}}{\sum_{i=1}^{n_1} w_{1i}}, \bar{x}_{2w} = \frac{\sum_{i=1}^{n_2} w_{2i} x_{2i}}{\sum_{i=1}^{n_2} w_{2i}}$$

The two unweighted sample variances are given by

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \quad s_2^2 = \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

t-Test for Means with Partial Pairing

The unweighted sample covariance between the two sets of paired observations is given by

$$c = \frac{\sum_{i=1}^{n_0} (x_{1i} - \bar{x}_{10})(x_{2i} - \bar{x}_{20})}{n_0 - 1}$$

$$= \frac{\sum_{i=1}^{n_0} x_{1i} x_{2i} - n_0 \bar{x}_{10} \bar{x}_{20}}{n_0 - 1}$$

where $\bar{x}_{10}$ is the mean of the first n_0 observations on population 1 and $\bar{x}_{20}$ is the mean of the first n_0 observations on population 2.

In analogy with the way we estimate the variance of the difference between the two sample weighted means when the weights applied to each of the observations on the first n_0 respondents is identical, our estimate in this case is given by

$$\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2c\sum_{i=1}^{n_0} w_{1i} w_{2i}}{\sum_{i=1}^{n_1} w_{1i} \sum_{i=1}^{n_2} w_{2i}},$$

where e_1 and e_2 are the effective sample sizes for the samples from populations 1 and 2, namely

$$e_1 = \frac{(\sum_{i=1}^{n_1} w_{1i})^2}{\sum_{i=1}^{n_1} w_{1i}^2}, \quad e_2 = \frac{(\sum_{i=1}^{n_2} w_{2i})^2}{\sum_{i=1}^{n_2} w_{2i}^2}$$

and

$$e_0 = \frac{\sum_{i=1}^{n_1} w_{1i} \sum_{i=1}^{n_2} w_{2i}}{\sum_{i=1}^{n_0} w_{1i} w_{2i}}$$

The t-statistic to test the difference between the two means is given by

$$t = \frac{\bar{x}_{1w} - \bar{x}_{2w}}{\sqrt{\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0 c}{e_1 e_2}}}$$

T-TESTS - DEPENDENT PAIRED/OVERLAP (MULTI)

General notation

Suppose we wanted to compare the mean of a respondent's attribute (e.g., age) on for those responding to item 1 (e.g., drank Coke) with the mean of that attribute for those responding to item 2 (e.g., drank Pepsi). Here we deal with a single measurement and compare averages of this measurement across subsets of respondents.

Let us partition the respondents so that the first n respondents provide data on both item 1 and item 2, the next m respondents provide data only on item 1, and the last p respondents provide data only on item 2. (There may be still other respondents that provided data on some, if not all, of the other items, but not on items 1 or 2. These will be disregarded in this analysis.)

Let us denote by x_i the observed measurement for respondent i (i = 1, 2, …, n), by y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m), and by z_i the observed measurement for respondent i (i = n+m+1, n+m+2, …, n+m+p). (I assign each of these measurements different letter names for clarity of exposition; the data are really a set of n+m+p observations.)

UNWEIGHTED DATA

The mean of the measurements for that attribute for those responding to item 1 is given by

$$\bar{X}_1 = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{n+m}$$

and the mean for that attribute for those responding to item 2 is given by

$$\bar{X}_2 = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+m+1}^{n+m+p} z_i}{n+p}$$

The difference of the two means is given by

$$\bar{X}_1 - \bar{X}_2 = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{n+m} - \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+m+1}^{n+m+p} z_i}{n+p}$$

$$= \left(\frac{1}{n+m} - \frac{1}{n+p}\right) n\bar{x} + \left(\frac{1}{n+m}\right) m\bar{y} - \left(\frac{1}{n+p}\right) p\bar{z}$$

where $\bar{x}$ is the mean of the measurements among those who were positive on both item 1 and item 2, $\bar{y}$ is the mean of the measurements among those who were positive only on item 1, and $\bar{z}$ is the mean of the measurements among those who were positive only on item 2.

Therefore the variance of the difference of the two means is given by

$$(\frac{1}{n+m}-\frac{1}{n+p})^2 n\sigma_x^2+(\frac{1}{n+m})^2 m\sigma_y^2+(\frac{1}{n+p})^2 p\sigma_z^2$$

The estimate of the variance of the difference of the two means is given by

$$s_d^2=(\frac{1}{n+m}-\frac{1}{n+p})^2 n\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}+(\frac{1}{n+m})^2 m\frac{\sum_{i=n+1}^{n+m}(y_i-\bar{y})^2}{m-1}+(\frac{1}{n+p})^2 p\frac{\sum_{i=n+m+1}^{n+m+p}(z_i-\bar{z})^2}{p-1}$$

The t-statistic for testing the difference of means is given by

$$t=\frac{\bar{X}_1-\bar{X}_2}{s_d}$$

The computation of the number of degrees of freedom is based on a generalization of the Satterthwaite formula, and is given by

$$df_S=\frac{\left[\frac{s_x^2}{n}+\frac{s_y^2}{m}+\frac{s_z^2}{p}\right]^2}{\frac{\left[\frac{s_x^2}{n}\right]^2}{n-1}+\frac{\left[\frac{s_y^2}{m}\right]^2}{m-1}+\frac{\left[\frac{s_z^2}{p}\right]^2}{p-1}}$$

where

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s_y^2 = \frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{m-1}$$

$$s_z^2 = \frac{\sum_{i=n+m+1}^{n+m+p}(z_i - \bar{z})^2}{p-1}$$

When m ≤ 1 then s_y^2 is 0 and the second term in the expression for s_d^2 is eliminated. When p ≤ 1 then s_z^2 is 0 and the third term in the expression for s_d^2 is eliminated. When both m and p are equal to 0, i.e., when there is total overlap, this test reduces to the dependent paired t test. When n=0, i.e., when there is no overlap, this test reduces to the independent t test with unequal variances.

Part-Whole Comparisons

Suppose we wanted to compare the mean of a respondent's attribute (e.g., age) for those responding to item 1 (e.g., drank Coke) with the mean of that attribute for those responding to the questionnaire. Here we deal with a single measurement and compare averages of this measurement between a subset of respondents and all respondents.

Let us partition the respondents so that the first n respondents provide data on both item 1 and at least one other item and the last m respondents provide data only on some other item.

Let us denote by x_i the observed measurement for respondent i (i = 1, 2, …, n) and by y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m). (I assign each of these measurements different letter names for clarity of exposition; the data are really a set of n+m observations.)

The mean of the measurements for that attribute for those responding to item 1 is given by

$$\bar{X}_1 = \frac{\sum_{i=1}^{n} x_i}{n}$$

and the mean for that attribute for those responding to all the items is given by

$$\bar{X}_T = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{n+m}$$

The difference of the two means is given by

$$\bar{X}_1 - \bar{X}_T = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{n+m}$$
$$= (\frac{1}{n} - \frac{1}{n+m})n\bar{x} - (\frac{1}{n+m})m\bar{y}$$

where $\bar{x}$ is the mean of the measurements among those who were positive on item 1 and $\bar{y}$ is the mean of the measurements among those who were positive only on items other than item 1.

Therefore the variance of the difference of the two means is given by

$$(\frac{1}{n} - \frac{1}{n+m})^2 n\sigma_x^2 + (\frac{1}{n+m})^2 m\sigma_y^2$$

The estimate of the variance of the difference of the two means is given by

$$s_d^2 = (\frac{1}{n} - \frac{1}{n+m})^2 n\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} + (\frac{1}{n+m})^2 m\frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{m-1}$$

The t-statistic for testing the difference of means is given by

$$t = \frac{\bar{X}_1 - \bar{X}_T}{s_d}$$

The computation of the number of degrees of freedom is based on the Satterthwaite formula, and is given by

$$df_S = \frac{(n-1)(m-1)}{(n-1)(1-g)^2 + (m-1)g^2}$$

where

$$g = \frac{s_x^2 / n}{s_x^2 / n + s_y^2 / m}$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s_y^2 = \frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{m-1}$$

SINGLY WEIGHTED DATA

When the data are weighted, then

$$\bar{X}_{1w} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{m} w_i}$$

and

$$\bar{X}_{2w} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}$$

The difference of the two means is given by

$$\bar{X}_{1w} - \bar{X}_{2w} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i} - \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}$$

$$= \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i} - \frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}\right)\sum_{i=1}^{n} w_i x_i + \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}\right)\sum_{i=n+1}^{n+m} w_i y_i - \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}\right)\sum_{i=n+m+1}^{n} w_i z_i$$

$$= \left(\frac{\sum_{i=n+m+1}^{n+m+p} w_i - \sum_{i=n+1}^{n+m} w_i}{\left[\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i\right]\left[\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i\right]}\right)\sum_{i=1}^{n} w_i x_i + \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}\right)\sum_{i=n+1}^{n+m} w_i y_i - \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}\right)\sum_{i=n+m+1}^{n} w_i z_i$$

Let f_x be the sum of the weights for the x's, f_y be the sum of the weights for the y's, and f_z be the sum of the weights for the z's. Then the variance of the difference of the two means is given by

$$\left(\frac{f_z - f_y}{[f_x + f_y][f_x + f_z]}\right)^2 \left(\sum_{i=1}^{n} w_i^2\right)\sigma_x^2 + \left(\frac{1}{f_x + f_f}\right)^2 \left(\sum_{i=n+1}^{n+m} w_i^2\right)\sigma_y^2 + \left(\frac{1}{f_x + f_z}\right)^2 \left(\sum_{i=n+m+1}^{n} w_i^2\right)\sigma_z^2$$

The estimate of the variance of the difference of the two means is given by

$$s_d^2 = \left(\frac{f_z - f_y}{[f_x + f_y][f_x + f_z]}\right)^2 \left(\sum_{i=1}^{n} w_i^2\right) \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} + \left(\frac{1}{f_x + f_f}\right)^2 \left(\sum_{i=n+1}^{n+m} w_i^2\right) \frac{\sum_{i=n+1}^{n+m} (y_i - \bar{y})^2}{m-1} + \left(\frac{1}{f_x + f_z}\right)^2 \left(\sum_{i=n+m+1}^{n} w_i^2\right) \frac{\sum_{i=n+m+1}^{n+m+p} (z_i - \bar{z})^2}{p-1}$$

The t-statistic for testing the difference of means is given by

$$t = \frac{\bar{X}_{1w} - \bar{X}_{2w}}{s_d}$$

The computation of the number of degrees of freedom is based on a generalization of the Satterthwaite formula, and is given by

$$df_S = \frac{\left[\frac{s_x^2}{n} + \frac{s_y^2}{m} + \frac{s_z^2}{p}\right]^2}{\frac{\left[\frac{s_x^2}{n}\right]^2}{n-1} + \frac{\left[\frac{s_y^2}{m}\right]^2}{m-1} + \frac{\left[\frac{s_z^2}{p}\right]^2}{p-1}}$$

where

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$s_y^2 = \frac{\sum_{i=n+1}^{n+m} (y_i - \bar{y})^2}{m-1}$$

$$s_z^2 = \frac{\sum_{i=n+m+1}^{n+m+p} (z_i - \bar{z})^2}{p-1}$$

When m ≤ 1 then s_y^2 is 0 and the second term in the expression for s_d^2 is eliminated. When p ≤ 1 then s_z^2 is 0 and the third term in the expression for s_d^2 is eliminated. When both m and p are equal to 0, i.e., when there is total overlap, this test reduces to the dependent paired t test. When n=0, i.e., when there is no overlap, this test reduces to the independent t test with unequal variances.

Part-Whole Comparisons

When the data are weighted, then

$$\bar{X}_{1w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

and

$$\bar{X}_{Tw} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}$$

The difference of the two means is given by

$$\bar{X}_{1w} - \bar{X}_{Tw} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} - \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}$$

$$= \left(\frac{1}{\sum_{i=1}^{n} w_i} - \frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}\right)\sum_{i=1}^{n} w_i x_i - \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+m}^{n+m} w_i}\right) \sum_{i=n+m+1}^{n} w_i y_i$$

$$= \frac{\sum_{i=n+1}^{n+m} w_i}{\left[\sum_{i=1}^{n} w_i\right]\left[\sum_{i=1}^{n+m} w_i\right]} \sum_{i=1}^{n} w_i x_i - \frac{1}{\sum_{i=1}^{n+m} w_i} \sum_{i=n+1}^{n+n} w_i y_i$$

Let f_x be the sum of the weights for the x's, f_y be the sum of the weights for the y's, and $f=f_x+f_y$ be the sum of the weights for all the observations. Then the variance of the difference of the two means is given by

$$\frac{f_y^2}{f^2 f_x^2}\left(\sum_{i=1}^{n} w_i^2\right)\sigma_x^2 + \frac{1}{f^2}\left(\sum_{i=n+1}^{n+m} w_i^2\right)\sigma_y^2$$

The estimate of the variance of the difference of the two means is given by

$$s_d^2 = \frac{f_y}{f^2 f_x^2}\left(\sum_{i=1}^{n} w_i^2\right)\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} + \frac{1}{f^2}\left(\sum_{i=n+1}^{n+m} w_i^2\right)\frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{m-1}$$

The t-statistic for testing the difference of means is given by

$$t = \frac{\bar{X}_{1w} - \bar{X}_{Tw}}{s_d}$$

The computation of the number of degrees of freedom is based on the Satterthwaite formula, and is given by

$$df_S = \frac{(n-1)(m-1)}{(n-1)(1-g)^2 + (m-1)g^2}$$

where

$$g = \frac{s_x^2 / n}{s_x^2 / n + s_y^2 / m}$$

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$s_y^2 = \frac{\sum_{i=n+1}^{n+m} (y_i - \bar{y})^2}{m-1}$$

MULTIPLY WEIGHTED DATA

When the data are weighted, with two separate weights applied to the x_is, where w_{i1} is used for the first weighted mean and w_{i2} is used for the second mean, then

$$\bar{X}_{1w} = \frac{\sum_{i=1}^{n} w_{i1} x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{n} w_{i1} + \sum_{i=n+1}^{m} w_i}$$

and

$$\bar{X}_{2w} = \frac{\sum_{i=1}^{n} w_{i2} x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{\sum_{i=1}^{n} w_{i2} + \sum_{i=n+m+1}^{n+m+p} w_i}$$

Let f_{x1} be the sum of the weights for the x's using weight 1, f_{x2} be the sum of the weights for the x's using weight 2, f_y be the sum of the weights for the y's, and f_z be the sum of the weights for the z's. The difference of the two means is given by

$$\bar{X}_{1w} - \bar{X}_{2w} = \frac{\sum_{i=1}^{n} w_{i1} x_i + \sum_{i=n+1}^{n+m} w_i y_i}{f_{x1} + f_y} - \frac{\sum_{i=1}^{n} w_{i2} x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{f_{x2} + f_z}$$

$$= \frac{\sum_{i=1}^{n} [(f_{x2} + f_z) w_{i1} - (f_{x1} + f_y) w_{i2}] x_i}{(f_{x1} + f_y)(f_{x2} + f_z)} + \frac{\sum_{i=n+1}^{n+m} w_i y_i}{f_{x1} + f_y} - \frac{\sum_{i=n+m+1}^{n} w_i z_i}{f_{x2} + f_z}$$

Then the variance of the difference of the two means is given by

$$\frac{\sum_{i=1}^{n} [(f_{x2} + f_z) w_{i1} - (f_{x1} + f_y) w_{i2})]^2 \sigma_x^2}{(f_{x1} + f_y)^2 (f_{x2} + f_z)^2} + \frac{\sum_{i=n+1}^{n+m} w_i^2 \sigma_y^2}{(f_{x1} + f_y)^2} + \frac{\sum_{i=n+m+1}^{n} w_i^2 \sigma_z^2}{(f_{x2} + f_z)^2}$$

The estimate of the variance of the difference of the two means is given by

$$s_d^2 = \frac{\sum_{i=1}^{n} [(f_{x2} + f_z) w_{i1} - (f_{x1} + f_y) w_{i2})]^2}{(f_{x1} + f_y)^2 (f_{x2} + f_z)^2} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} + \frac{\sum_{i=n+1}^{n+m} w_i^2}{(f_{x1} + f_y)^2} \frac{\sum_{i=n+1}^{n+m} (y_i - \bar{y})^2}{m-1} + \frac{\sum_{i=n+m+1}^{n} w_i^2}{(f_{x2} + f_z)^2} \frac{\sum_{i=n+m+1}^{n+m+p} (z_i - \bar{z})^2}{p-1}$$

The t-statistic for testing the difference of means is given by

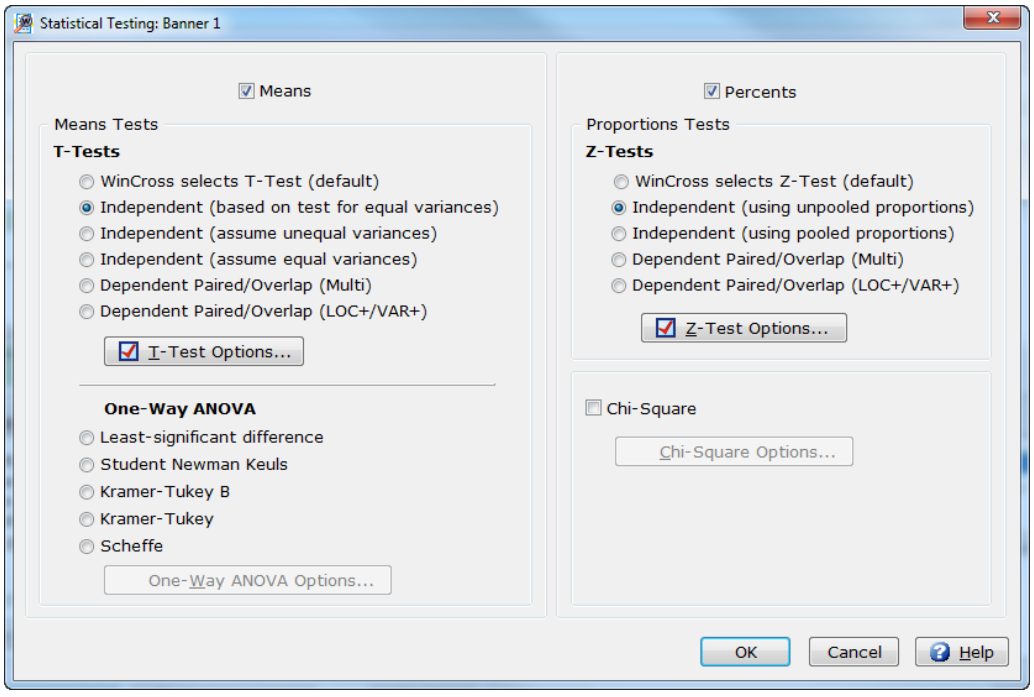
$$t = \frac{\bar{X}_{1w} - \bar{X}_{2w}}{s_d}$$

Z-TESTS - INDEPENDENT

General Notation

We consider here the situation in which we have proportions from two populations, where $\mathbf{n_1}$ is the number of observations in data set 1, $\mathbf{n_2}$ is the number of observations in data set 2, and the data are drawn independently from each of the populations. The proportions from each the two data sets will be designated as $\mathbf{p_1}$ and $\mathbf{p_2}$. The object of this z-test is to test whether the proportions in the two populations from which the data were drawn are different.

WinCross gives the user the option to either estimate the common proportion (when the null hypothesis of no difference in population proportions is true) by pooling the separate sample proportions or to use each of the sample proportions separately. For reasons which will be explained later, we recommend the latter approach. This approach is implemented by selecting the **Independent (using unpooled proportions)** option. If one wants to pool the two proportions and use that test, one selects the **Independent (using pooled proportions)** option.



UNWEIGHTED DATA

Using unpooled proportions

The z statistic is given by

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

This statistic has a standard normal distribution even when the null hypothesis is false.

Using pooled proportions

When the null hypothesis that the two population proportions are equal is true, then one could create a pooled estimate of the common proportion, namely

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2},$$

next estimate the variance of p_1-p_2 by

$$\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2}),$$

and finally calculate

$$z^* = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}.$$

This statistic has a standard normal distribution only when the null hypothesis is true. Though some statistics textbooks recommend this latter test statistic, using the argument that the denominator of z* is a more accurate estimate of the standard deviation of the numerator than is the denominator of z. This argument is specious. The null hypothesis characteristics of the two tests are identical, and the z statistic using unpooled proportions is the more powerful test. Details about this may be found in the paper "A Comparison of Two Tests for Equality of Two Proportions" by Keith R. Eberhardt and Michael A. Fligner which appeared on pages 151-5 of Volume 31, Number 4 (November 1977) of the American Statistician.

Technical comment:

Testing for Equality of Two Multinomial Proportions

Given a sample of size n, and sample counts n_1, n_2, ...,n_m in m categories (with n_1+ n_2+ ...+n_m= n), one would like to test whether the sample counts in two of the categories, say i and j, are significantly different. We assume that the items in the sample are independently drawn from a multinomial population, with P_k denoting the probability that a randomly selected item comes from category k, k = 1, 2, ..., m (where P_1+ P_2+ ...+P_m = 1). The null hypothesis being tested is that $P_i = P_j$.

Though this hypothesis being tested looks in form like the test situation considered in this section, it is NOT the same. First of all, the independent z-test situation considered in this section is typically set up to test equality of proportions from pairs of columns, whereas in this note we are considering testing equality of proportions from pairs of rows. But the main reason it is not the same is that the observations on P_i are not independent of the observations on P_j, because the higher the estimate of P_i the lower will be the estimate of P_j (because the sum of the estimates of the Ps must add to 1).

So how does one set up the test of this hypothesis? Let $p_i=n_i/n$ and $p_j=n_j/n$ be the estimates of P_i and P_j based on the sample. The test statistic will be based on $p_i - p_j$. The variance of p_i is $P_i(1-P_i)/n$, the variance of p_j is $P_j(1-P_j)/n$, and the covariance of p_i and p_j is $-P_i P_j/n$. Consequently, the variance of $p_i - p_j$ is given by

$$V=P_i(1-P_i)/n + P_j(1-P_j)/n +2 P_i P_j/n.$$

As the P's are unknown, V is estimated by replacing the P's by their sample estimates, the corresponding p's.

From these results we can construct a z-score to test the null hypothesis, namely as the test statistic for testing the null hypothesis that $P_i = P_j$.

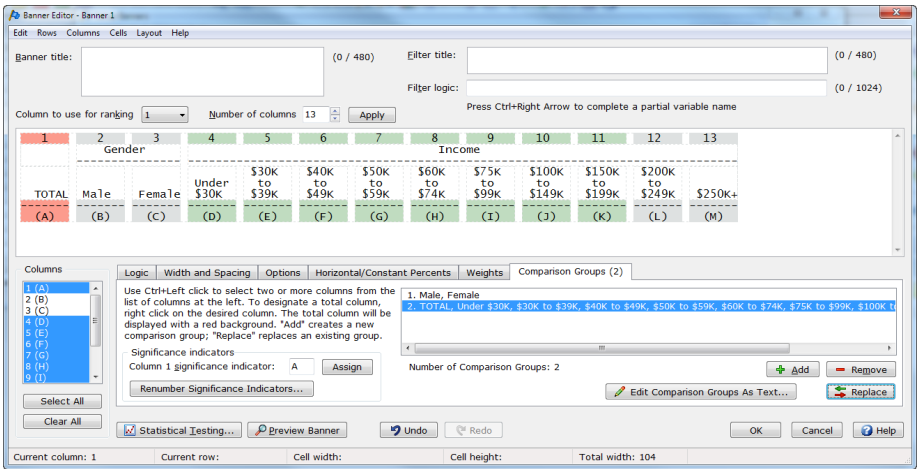
$$z = \frac{p_i - p_j}{\sqrt{\frac{p_i(1-p_i) + p_j(1-p_j) + 2p_i p_j}{n}}}$$

Notice that the denominator is larger than the z-statistic for comparison of independent proportions. Therefore, if one uses (incorrectly) the z-statistic for comparison of independent proportions one will be calculating a smaller-than-appropriate test statistic and erroneously saying that the two proportions are not significantly different when in fact they are.

WinCross does not have a facility for performing this test. However, The Analytical Group provides a facility for doing so, via the Quick Tools program that can be found on our website: www.analyticalgroup.com

Part-Whole Comparisons

One sometimes wants to compare the proportion p_1 of a subsample (e.g., a sample from division of a company) with the proportion p of the full sample (e.g., a sample from the entire company). These proportions are not independent, and so a special statistical procedure is necessary to implement this comparison. In particular, one has to designate which column of the table contains the totals. WinCross is told that one of the columns being used in a statistical test is a Total column by right-clicking on that column in the **Banner Editor**, as in this example:



Let m be the sample size of the subsample and n be the sample size of the full sample.

Since the null hypothesis is that the two proportions are equal, the proper z statistic for testing whether the subpopulation proportion differs from the population proportion, when using "pooled proportions" is

$$z = \frac{p_1 - p}{\sqrt{p(1-p)\left(\frac{1}{m} - \frac{1}{n}\right)}}$$

By contrast, if one erroneously used the independent t test one would calculate

$$z = \frac{p_1 - p}{\sqrt{p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

The denominator of this z statistic is larger than that of the correct z statistic, so that one will be calculating a smaller-than-appropriate test statistic and erroneously saying that the two proportions are not significantly different when in fact they are.

However, using the same rationale as given above in the Technical Comment: A Note on "Pooled Proportions," WinCross instead uses the statistic

$$z^* = \frac{p_1 - p}{\sqrt{\frac{(n-m)^2\, p_1(1-p_1)}{mn^2} + \frac{(n-m)\, p_{\sim 1}(1-p_{\sim 1})}{n^2}}}$$

where $p_{\sim 1}$ is the proportion of the complementary n-m subsample of the full sample.

As noted in the document A NOTE ON SPURIOUS SIGNIFICANCE on our web site, there is the possibility of spuriously finding "significant" differences due only because of the degree of overlap of the part to the whole, WinCross has adopted the safeguard of declaring all such differences not significant if the fraction of the part to the whole is less than 5% or greater than 95%.

SINGLY and MULTIPLY WEIGHTED DATA

General Notation

We consider here the situation in which we have data from two populations, where $\mathbf{n_1}$ is the number of observations in data set 1, $\mathbf{n_2}$ is the number of observations in data set 2, and the data are drawn independently from each of the populations. The proportions from each of the two data sets will be designated as $\mathbf{p_1}$ and $\mathbf{p_2}$. The weighted proportions of the two data sets will be designated as $\mathbf{p_{1w}}$ and $\mathbf{p_{2w}}$. $\bar{x}_{2w}$. These proportions may be calculated using a single weight for the observations from the two populations or separate weights applied to the data from each of the populations. The unweighted variances of the two data sets are, respectively, $\mathbf{p_1(1- p_1)}$ and $\mathbf{p_2(1- p_2)}$.

Using unpooled proportions

The z statistic is given by

$$z = \frac{p_{1w} - p_{2w}}{\sqrt{\frac{p_1(1-p_1)}{e_1} + \frac{p_2(1-p_2)}{e_2}}}$$

where e_1 and e_2 are the effective sample sizes of the two samples.

If one is performing a part-whole comparison with weighted data, the z statistic is given by

$$z^* = \frac{p_{1w} - p_w}{\sqrt{\frac{(e-e_1)^2\, p_1(1-p_)}{e_1 e^2} + \frac{(e-e_1)\, p_{\sim 1}(1-p_{\sim 1})}{e^2}}}$$

where $\mathbf{p_{1w}}$ is the weighted proportion of the subsample, $\mathbf{p_w}$ is the weighted proportion of the whole sample, $\mathbf{p_1}$ is the unweighted proportion of the subsample, $\mathbf{p_{\sim 1}}$ is the unweighted proportion of the complement of the subsample, e_1 is the effective sample size of the subsample, and e is the effective sample size of the whole sample.

Using pooled proportions

The z statistic is given by

$$z^* = \frac{p_{1w} - p_{2w}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{e_1} + \frac{1}{e_2})}}$$

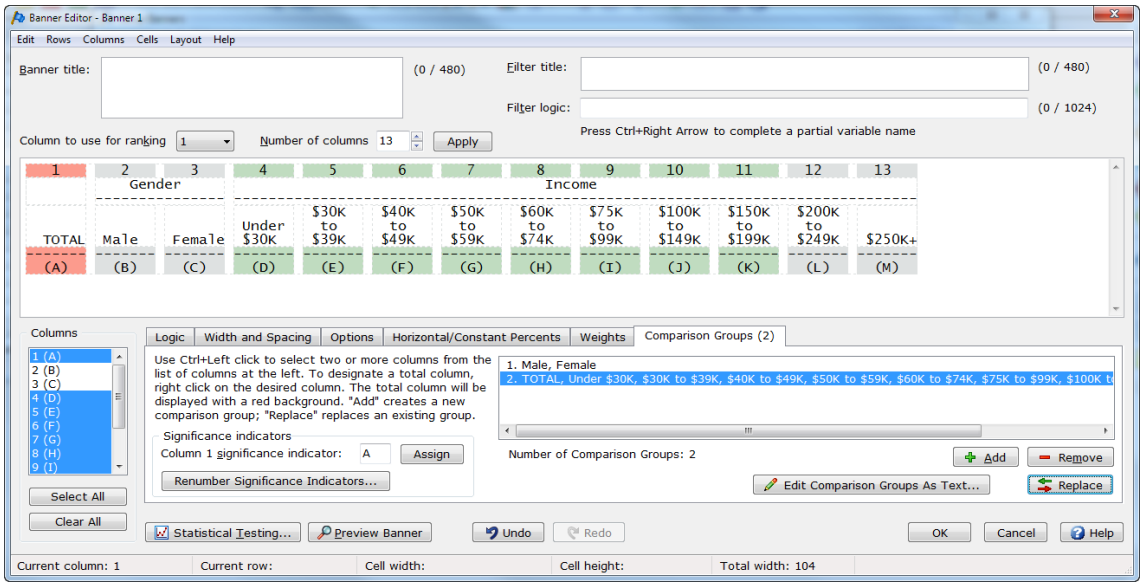
where

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

is the unweighted pooled proportion.

Part-Whole Comparisons

One sometimes wants to compare the weighted proportion p_{1w} of a subsample (e.g., a sample from division of a company) with the weighted proportion p_w of the full sample (e.g., a sample from the entire company). ***WinCross only applies the part-whole comparison test when a single weight is applied to all of the observations***. These proportions are not independent, and so a special statistical procedure is necessary to implement this comparison. In particular, one has to designate which column of the table contains the totals. WinCross is told that one of the columns being used in a statistical test is a Total column by right-clicking on that column in the **Banner Editor**, as in this example:



Let e_1 be the effective sample size of the subsample and e be the effective sample size of the full sample.

Since the null hypothesis is that the two proportions are equal, the proper z statistic for testing whether the subpopulation proportion differs from the population proportion, when using "pooled proportions" is

$$z = \frac{p_{1w} - p_w}{\sqrt{p(1-p)(\frac{1}{e_1} - \frac{1}{e})}}$$

where p is the unpooled proportion in the full sample.

However, using the same rationale as given above in the Technical Comment: A Note on "Pooled Proportions," we recommend instead the statistic

$$z^* = \frac{p_{1w} - p_w}{\sqrt{\frac{(e-e_1)^2\, p_1(1-p_1)}{e_1 e^2} + \frac{(e-e_1)\, p_{\sim 1}(1-p_{\sim 1})}{e^2}}}$$

where $p_{\sim 1}$ is the proportion of the complementary n-m subsample of the full sample.

Z-TESTS - DEPENDENT PAIRED/OVERLAP (LOC+/VAR+)

General Notation

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_2 independent observations $x_{21},...,x_{2n_2}$ from population 2, where each observation can take on only the values of 0 or 1 (e.g., an answer to a question as to whether the respondent liked or disliked a product). Suppose further that the first n_0 observations from the two populations are paired (for example, population 1 relates to a "treatment," population 2 relates to a "control," and the first n_0 observations are taken from the same respondent; for another example, population 1 relates to Coke, population 2 relates to Pepsi, and the first n_0 pairs of responses are taken from the same respondent).

The two sample proportions are

$$p_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad p_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

The two sample variances are given by

$$s_1^2 = p_1(1-p_1), \quad s_2^2 = p_2(1-p_2)$$

The object of this z-test is to test whether the proportions in the two populations from which the data were drawn are different.

UNWEIGHTED DATA

z-Test for Proportions with Partial Pairing

As with the two sample t-test for comparison of partially paired means, the sample covariance between the two sets of paired observations is given by

$$c = \frac{\sum_{i=1}^{n_0}(x_{1i} - p_{10})(x_{2i} - p_{20})}{n_0}$$

$$= \frac{\sum_{i=1}^{n_0} x_{1i}x_{2i} - n_0 p_{10} p_{20}}{n_0}$$

where p_{10} is the proportion of 1's in the first n_0 observations on population 1 and p_{20} is the proportion of 1's in the first n_0 observations on population 2. But

$$\frac{\sum_{i=1}^{n_0} x_{1i}x_{2i}}{n_0} = p_{120},$$

the proportion of first n_0 observations that are 1 in both population 1 and 2. Consequently, the sample covariance simplifies to

$$c = p_{120} - p_{10}p_{20}$$

The variance of the difference between the two sample proportions is estimated by

$$s_d^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - \frac{2n_0c}{n_1n_2}$$

The z-statistic to test the difference between the two proportions is given by

$$z = \frac{p_1 - p_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - \frac{2n_0c}{n_1n_2}}}$$

As noted in the document A NOTE ON SPURIOUS SIGNIFICANCE on our web site, there is the possibility of spuriously finding "significant" differences due only because of the degree of overlap of the two samples. WinCross has adopted the safeguard of declaring all such differences not significant if a factor, described in that document, based on the fraction of unique observations from population 1 and from population 2 is less than 5%.

Technical Comment:

A Note on Perfect Pairing

In the case where $n_1 = n_2 = n_0 = n$, say, i.e., when all the observations are paired, all these computations simplify considerably. First of all, the estimate of the variance of the difference between the two sample proportions simplifies to

$$s_d^2 = \frac{p_1(1-p_1) + p_2(1-p_2) - 2(p_{12} - p_1p_2)}{n}$$
$$= \frac{p_1 + p_2 - 2p_{12} - (p_1 - p_2)^2}{n}$$

Moreover, there is no need to calculate p_{12}, for, letting $d_i = x_{1i} - x_{2i}$, we see that

$$p_1 - p_2 = \bar{d} = \frac{\sum_{i=1}^{n} d_i}{n},$$

the proportion of (1,0) pairs minus the proportion of (0,1) pairs. Given this, the standard deviation of the differences between the paired observations can be calculated by

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n}}$$

so that the z-statistic to test the difference between the two proportions is given by

$$z = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Part-Whole Comparisons

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_j independent observations $x_{j1},...,x_{jn_j}$ from population j, j=2,…,m, where each observation can take on only the values of 0 or 1 (e.g., an answer to a question as to whether the respondent liked or disliked a product). For each of the m-1 pairs of observations (x_{1i}, x_{ji}) i=1,..., n, j=2, …, m there are n_{0j} that are paired (for example, population 1 is the liking or disliking of Coke, population 2 is the liking or disliking of Pepsi, population 3 is the liking or disliking of Seven-Up, and there are n_{02} sets of ratings from the same respondent for Coke and Pepsi and n_{03} sets of ratings from the same respondent for Coke and Seven-Up). We want to compare the proportion of 1's in population 1 (e.g., the proportion who like Coke) with the proportion of 1's across all m populations.

The two proportions are

$$p_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad p_T = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{ji}}{\sum_{j=1}^{m} n_j}$$

The variance of the difference between the two sample proportions is given by

$$\begin{aligned} & Vp_1 + Vp_T - 2Cov(p_1, p_T) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_T^2}{n_T} - \frac{2}{n_1 n_T}\sum_{j=1}^{m} Cov(\sum_{i=1}^{n_1} x_{1i}, \sum_{i=1}^{n_j} x_{ji}) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_T^2}{n_T} - \frac{2\sum_{j=1}^{m} n_{10j} Cov(x_1, x_j)}{n_1 n_T} \end{aligned}$$

The two sample variances are given by

$$s_1^2 = p_1(1-p_1), \; s_T^2 = p_T(1-p_T)$$

For each of the m-1 pairs of observations (x_{1i}, x_{ji}) i=1,..., n, j=2, …, m are correlated, we must calculate the covariance between the sample means as part of the standard error computation. WinCross calculates the sample covariance between the two sets of paired observations as

$$c_j = \frac{\sum_{i=1}^{n_{0j}} (x_{1i} - p_{10j})(x_{ji} - p_{j01})}{n_{10j} - 1}$$

where p_{10j} is the proportion of 1's in item 1 and p_{j01} is the proportion of 1's in item j among the n_{10j} observations from respondents who answered both item 1 and item j

The variance of the difference between the two sample proportions is estimated by

$$\frac{s_1^2}{n_1} + \frac{s_T^2}{n_T} - \frac{2c}{n_1 n_T}$$

where

$$c = \sum_{j=1}^{m} n_{10j} c_j$$

The z-statistic to test the difference between the two proportions is given by

$$z = \frac{p_1 - p_T}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_T^2}{n_T} - \frac{2c}{n_1 n_T}}}$$

SINGLY WEIGHTED DATA

General Notation

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_2 independent observations $x_{21},...,x_{2n_2}$ from population 2, where each observation can take on only the values of 0 or 1.. Suppose further that the first n_0 observations from the two populations are paired (e.g., population 1 is a "treatment", population 2 is a "control," and the first n_0 observations are taken from the same respondent). Finally, suppose that each of the respondents has an associated weight, with $w_{11},...,w_{1n_1}$ the weights for the respondents from population 1, $w_{21},...,w_{2n_2}$ the weights for the respondents from population 2, and where the weights applied to each of the observations on the first n_0 respondents are identical for both observations, i.e., $w_{11} = w_{21} = w_1,...,w_{1n_0} = w_{2n_0} = w_{n_0}$.

The two weighted sample proportions are

$$p_{1w} = \frac{\sum_{i=1}^{n_1} w_{1i} x_{1i}}{\sum_{i=1}^{n_1} w_{1i}}, \quad p_{2w} = \frac{\sum_{i=1}^{n_2} w_{2i} x_{2i}}{\sum_{i=1}^{n_2} w_{2i}}$$

The two unweighted sample proportions are

$$p_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad p_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

The two unweighted sample variances are given by

$$s_1^2 = p_1(1-p_1), \quad s_2^2 = p_2(1-p_2)$$

z-Test for Proportions with Partial Pairing

The unweighted sample covariance between the two sets of paired observations is given by

$$c = \frac{\sum_{i=1}^{n_0} (x_{1i} - p_{10})(x_{2i} - p_{20})}{n_0}$$

$$= \frac{\sum_{i=1}^{n_0} x_{1i} x_{2i} - n_0 p_{10} p_{20}}{n_0}$$

where p_{10} is the proportion of 1's in the first n_0 observations on population 1 and p_{20} is the proportion of 1's in the first n_0 observations on population 2. But

$$\frac{\sum_{i=1}^{n_0} x_{1i} x_{2i}}{n_0} = p_{120},$$

the proportion of first n_0 observations that are 1 in both population 1 and 2. Consequently, the sample covariance simplifies to

$$c = p_{120} - p_{10} p_{20}$$

The best estimate of the variance of the difference between the two sample weighted means is given by

$$\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0 c}{e_1 e_2},$$

where e_1 and e_2 are the effective sample sizes for the samples from populations 1 and 2,

$$e_1 = \frac{(\sum_{i=1}^{n_1} w_{1i})^2}{\sum_{i=1}^{n_1} w_{1i}^2}, \quad e_2 = \frac{(\sum_{i=1}^{n_2} w_{2i})^2}{\sum_{i=1}^{n_2} w_{2i}^2}$$

and e_0 is the effective sample size for the observations common to populations 1 and 2,

$$e_0 = \frac{(\sum_{i=1}^{n_0} w_i)^2}{\sum_{i=1}^{n_0} w_i^2}$$

The z-statistic to test the difference between the two weighted proportions is given by

$$z = \frac{p_{1w} - p_{2w}}{\sqrt{\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0 c}{e_1 e_2}}}$$

Technical Comment:

A Note on Perfect Pairing

In the case where $n_1 = n_2 = n_0 = n$, say, i.e., when all the observations are paired, all these computations simplify considerably. Letting $d_i = x_{1i} - x_{2i}$, we see that

$$p_{1w} - p_{2w} = \bar{d}_w = \frac{\sum_{i=1}^{n} w_i d_i}{\sum_{i=1}^{n} w_i},$$

Given this, the variance of $\bar{d}_w$ is just the unweighted variance of d divided by the effective sample size

$$e = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$$

The unweighted variance of d can be calculated by

$$s_d^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n}$$

so that the z-statistic to test the difference between the two proportions is given by

$$z = \frac{\bar{d}_w}{s_d / \sqrt{e}}$$

Part-Whole Comparisons

Suppose we have n_1 independent observations $x_{11},...,x_{1n_1}$ from population 1 and n_j independent observations $x_{j1},...,x_{jnj_2}$ from population j, j=2,…,m, where each observation can take on only the values of 0 or 1. Suppose further that the first n_0 observations from each of the populations are paired (e.g., the first n_0 observations are taken from the same respondent). Finally, suppose that each of the respondents has an associated weight, with $w_1,...,w_n$ and where the weights applied to each of the

observations on the first n_0 respondents are identical for all the observations, i.e., $w_{11} = w_{21} = w_1, ..., w_{1n_0} = w_{2n_0} = w_{n_0}$.

The two sample proportions are

$$p_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \; p_T = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{ji}}{\sum_{j=1}^{m} n_j}$$

The two weighted sample proportions are

$$p_{1w} = \frac{\sum_{i=1}^{n_1} w_{i1} x_{1i}}{\sum_{i=1}^{n_1} w_{1i}}, \; p_{Tw} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} w_i x_{ji}}{\sum_{j=1}^{m} \sum_{i=1}^{n_j} w_i}$$

The variance of the difference between the two sample means is given by

$$\begin{aligned} & Vp_{1w} + Vp_{Tw} - 2Cov(p_{1w}, p_{Tw}) \\ &= \frac{\sigma_1^2}{e_1} + \frac{\sigma_T^2}{e_T} - \frac{2}{(\sum_{j=1}^{n} w_{1j})(\sum_{j=1}^{m} \sum_{i=1}^{n_j} w_i)} \sum_{j=1}^{m} Cov(\sum_{i=1}^{n_1} w_{1i} x_{1i}, \sum_{i=1}^{n_j} w_{ji} x_{ji}) \\ &= \frac{\sigma_1^2}{e_1} + \frac{\sigma_T^2}{e_T} - \frac{2\sum_{j=1}^{m} Cov(x_1, x_j) \sum_{i=1}^{n_j} w_{10ji}^2}{(\sum_{j=1}^{n} w_{1j})(\sum_{j=1}^{m} \sum_{i=1}^{n_j} w_i)} \end{aligned}$$

where w_{10ji}^2 is the square of the weight for the i-th respondent who answered both questions 1 and j.

The two unweighted sample variances are given by

$$s_1^2 = p_1(1 - p_1), \; s_T^2 = p_T(1 - p_T)$$

As above, we use the unweighted sample covariance in calculating the variance of the difference between the two means. This is given by

$$c_j = \frac{\sum_{i=1}^{n_{0j}} (x_{1i} - p_{10j})(x_{ji} - p_{j01})}{n_{10j} - 1}$$

where p_{10j} is the proportion of 1's in item 1 and p_{j01} is the proportion of 1's in item j among the n_{10j} observations from respondents who answered both item 1 and item j

The best estimate of the variance of the difference between the two sample weighted proportions is given by

$$\frac{s_1^2}{e_1} + \frac{s_T^2}{e_T} - \frac{2c}{e_1 e_T},$$

where

$$c = \sum_{j=1}^{m} c_j \sum_{i=1}^{n_j} w_{10\,ji}^2$$

and where e_1 and e_T are the effective sample sizes for the samples from populations 1 and the set of m populations, namely

$$e_1 = \frac{(\sum_{i=1}^{n_1} w_{1i})^2}{\sum_{i=1}^{n_1} w_{1i}^2}, \quad e_T = \frac{(\sum_{j=1}^{m} \sum_{i=1}^{n_2} w_i)^2}{\sum_{j=1}^{m} \sum_{i=1}^{n_2} w_i^2}$$

The z-statistic to test the difference between the two means is given by

$$z = \frac{p_{1w} - p_{Tw}}{\sqrt{\frac{s_1^2}{e_1} + \frac{s_T^2}{e_T} - \frac{2c}{e_1 e_T}}}$$

MULTIPLY WEIGHTED DATA

z-Test for Proportions with Partial Pairing

The two weighted sample proportions are

$$p_{1w} = \frac{\sum_{i=1}^{n_1} w_{1i} x_{1i}}{\sum_{i=1}^{n_1} w_{1i}}, \quad p_{2w} = \frac{\sum_{i=1}^{n_2} w_{2i} x_{2i}}{\sum_{i=1}^{n_2} w_{2i}}$$

The two unweighted sample proportions are

$$p_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \quad p_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

The two unweighted sample variances are given by

$$s_1^2 = p_1(1-p_1), \quad s_2^2 = p_2(1-p_2)$$

The unweighted sample covariance between the two sets of paired observations is given by

$$c = \frac{\sum_{i=1}^{n_0}(x_{1i} - p_{10})(x_{2i} - p_{20})}{n_0}$$

$$= \frac{\sum_{i=1}^{n_0} x_{1i}x_{2i} - n_0 p_{10} p_{20}}{n_0}$$

where p_{10} is the proportion of 1's in the first n_0 observations on population 1 and p_{20} is the proportion of 1's in the first n_0 observations on population 2. But

$$\frac{\sum_{i=1}^{n_0} x_{1i}x_{2i}}{n_0} = p_{120},$$

the proportion of first n_0 observations that are 1 in both population 1 and 2. Consequently, the sample covariance simplifies to

$$c = p_{120} - p_{10}p_{20}$$

The best estimate of the variance of the difference between the two sample weighted means is given by

$$\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0 c}{e_1 e_2},$$

where e_1 and e_2 are the effective sample sizes for the samples from populations 1 and 2,

$$e_1 = \frac{(\sum_{i=1}^{n_1} w_{1i})^2}{\sum_{i=1}^{n_1} w_{1i}^2}, \quad e_2 = \frac{(\sum_{i=1}^{n_2} w_{2i})^2}{\sum_{i=1}^{n_2} w_{2i}^2}$$

and e_0 is given by

$$e_0 = \frac{\sum_{i=1}^{n_1} w_{1i} \sum_{i=1}^{n_2} w_{2i}}{\sum_{i=1}^{n_0} w_{1i} w_{2i}}$$

The z-statistic to test the difference between the two weighted proportions is given by

$$z = \frac{p_{1w} - p_{2w}}{\sqrt{\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0 c}{e_1 e_2}}}$$

Z-TESTS - DEPENDENT PAIRED/OVERLAP (MULTI)

General notation

Suppose we wanted to compare the proportion of respondents who had a particular attribute (e.g., scored a new product as "favorable") for those responding to item 1 (e.g., drank Coke) with the proportion of respondents who had that particular attribute for those responding to item 2 (e.g., drank Pepsi). Here we deal with a single dichotomous attribute, i.e., an attribute that can take on a value of 1 if present and 0 if absent, and compare proportions who had that attribute across subsets of respondents.

Let us partition the respondents so that the first n respondents provide data on both item 1 and item 2, the next m respondents provide data only on item 1, and the last p respondents provide data only on item 2. (There may be still other respondents that provided data on some, if not all, of the other items, but not on items 1 or 2. These will be disregarded in this analysis.)

Let us denote by x_i the observed attribute value for respondent i (i = 1, 2, …, n), by y_i the observed attribute value for respondent i (i = n+1, n+2, …, n+m), and by z_i the observed attribute value for respondent i (i = n+m+1, n+m+2, …, n+m+p). (I assign each of these attribute values different letter names for clarity of exposition; the data are really a set of n+m+p observations.)

UNWEIGHTED DATA

The proportion of the sample with the attribute under consideration for those responding to item 1 is given by

$$q_1 = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{n+m}$$

and the proportion for that attribute for those responding to item 2 is given by

$$q_2 = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+m+1}^{n+m+p} z_i}{n+p}$$

The difference of the two proportions is given by

$$q_1 - q_2 = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{n+m} - \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+m+1}^{n+m+p} z_i}{n+p}$$

$$= \left(\frac{1}{n+m} - \frac{1}{n+p}\right) n q_x + \left(\frac{1}{n+m}\right) m q_y - \left(\frac{1}{n+p}\right) p q_z$$

where q_x is the proportion with the attribute among those who were positive on both item 1 and item 2, q_y is the proportion with the attribute among those who were positive only on item 1, and q_z is the proportion with the attribute among those who were positive only on item 2.

The variance of the difference of the two proportions is therefore estimated by

$$s_d^2 = (\frac{1}{n+m} - \frac{1}{n+p})^2 nq_x(1-q_x) + (\frac{1}{n+m})^2 mq_y(1-q_y) + (\frac{1}{n+p})^2 pq_z(1-q_z)$$

The z-statistic for testing the difference of proportions is given by

$$z = \frac{q_1 - q_2}{s_d}$$

Part-Whole Comparisons

Suppose we wanted to compare the proportion of respondents with a given attribute (e.g., males) on for those responding to item 1 (e.g., drank Coke) with the proportion of respondents with that attribute for those responding to the questionnaire. Here we deal with a single measurement and compare averages of this measurement between a subset of respondents and all respondents.

Let us partition the respondents so that the first n respondents provide data on both item 1 and at least one other item and the last m respondents provide data only on some other item. Let us denote by x_i the observed measurement for respondent i (i = 1, 2, …, n) and by y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m). (I assign each of these measurements different letter names for clarity of exposition; the data are really a set of n+m observations.) We consider here that the x's and y's are either 0s or 1s.

The proportion of those responding to item 1 with that attribute is given by

$$p_1 = \frac{\sum_{i=1}^{n} x_i}{n}$$

and the proportion with that attribute for those responding to all the items is given by

$$p_T = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{n+m}$$

The difference of the two proportions is given by

$$p_1 - p_T = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{n+m}$$
$$= (\frac{1}{n} - \frac{1}{n+m})n\bar{x} - (\frac{1}{n+m})m\bar{y}$$

where $\bar{x}$ is the proportion of the respondents among those who were positive on item 1 and $\bar{y}$ is the proportion of the respondents among those who were positive only on items other than item 1.

Therefore the variance of the difference of the two proportions is given by

$$(\frac{1}{n} - \frac{1}{n+m})^2 n\sigma_x^2 + (\frac{1}{n+m})^2 m\sigma_y^2$$

The estimate of the variance of the difference of the two proportions is given by

$$s_d^2 = (\frac{1}{n} - \frac{1}{n+m})^2 np_x(1-p_x) + (\frac{1}{n+m})^2 mp_y(1-p_y)$$

The t-statistic for testing the difference of means is given by

$$t = \frac{\bar{X}_1 - \bar{X}_T}{s_d}$$

SINGLY WEIGHTED DATA

When the data are weighted, then

$$q_{1w} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{m} w_i}$$

and

$$q_{2w} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}$$

The difference of the two proportions is given by

$$q_{1w} - q_{2w} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i} - \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}$$

$$= \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i} - \frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}\right)\sum_{i=1}^{n} w_i x_i + \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}\right)\sum_{i=n+1}^{n+m} w_i y_i - \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}\right)\sum_{i=n+m+1}^{n} w_i z_i$$

$$= \left(\frac{\sum_{i=n+m+1}^{n+m+p} w_i - \sum_{i=n+1}^{n+m} w_i}{\left[\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i\right]\left[\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i\right]}\right)\sum_{i=1}^{n} w_i x_i + \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}\right)\sum_{i=n+1}^{n+m} w_i y_i - \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+m+1}^{n+m+p} w_i}\right)\sum_{i=n+m+1}^{n} w_i z_i$$

Let f_x be the sum of the weights for the x's, f_y be the sum of the weights for the y's, and f_z be the sum of the weights for the z's. Then the variance of the difference of the two means is given by

$$\left(\frac{f_z - f_y}{[f_x + f_y][f_x + f_z]}\right)^2 \left(\sum_{i=1}^{n} w_i^2\right)\sigma_x^2 + \left(\frac{1}{f_x + f_y}\right)^2 \left(\sum_{i=n+1}^{n+m} w_i^2\right)\sigma_y^2 + \left(\frac{1}{f_x + f_z}\right)^2 \left(\sum_{i=n+m+1}^{n} w_i^2\right)\sigma_z^2$$

The estimate of the variance of the difference of the two means is given by

$$s_d^2 = \left(\frac{f_z - f_y}{[f_x + f_y][f_x + f_z]}\right)^2 \left(\sum_{i=1}^{n} w_i^2\right)q_x(1-q_x) + \left(\frac{1}{f_x + f_y}\right)^2 \left(\sum_{i=n+1}^{n+m} w_i^2\right)q_y(1-q_y) + \left(\frac{1}{f_x + f_z}\right)^2 \left(\sum_{i=n+m+1}^{n} w_i^2\right)q_z(1-q_z)$$

where q_x is the proportion with the attribute among those who were positive on both item 1 and item 2, q_y is the proportion with the attribute among those who were positive only on item 1, and q_z is the proportion with the attribute among those who were positive only on item 2.

The z-statistic for testing the difference of proportions is given by

$$z = \frac{q_1 - q_2}{s_d}$$

Part-Whole Comparisons

When the data are weighted, then

$$p_{1w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

and

$$p_{Tw} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}$$

The difference of the two weighted proportions is given by

$$p_{1w} - p_{Tw} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} - \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}$$

$$= \left(\frac{1}{\sum_{i=1}^{n} w_i} - \frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+1}^{n+m} w_i}\right)\sum_{i=1}^{n} w_i x_i - \left(\frac{1}{\sum_{i=1}^{n} w_i + \sum_{i=n+m}^{n+m} w_i}\right) \sum_{i=n+m+1}^{n} w_i y_i$$

$$= \frac{\sum_{i=n+1}^{n+m} w_i}{[\sum_{i=1}^{n} w_i][\sum_{i=1}^{n+m} w_i]} \sum_{i=1}^{n} w_i x_i - \frac{1}{\sum_{i=1}^{n+m} w_i} \sum_{i=n+1}^{n+n} w_i y_i$$

Let f_x be the sum of the weights for the x's, f_y be the sum of the weights for the y's, and $f=f_x+f_y$ be the sum of the weights for all the observations. Then the variance of the difference of the two weighted proportions is given by

$$\frac{f_y^2}{f^2 f_x^2}\left(\sum_{i=1}^{n} w_i^2\right)\sigma_x^2 + \frac{1}{f^2}\left(\sum_{i=n+1}^{n+m} w_i^2\right)\sigma_y^2$$

The estimate of the variance of the difference of the two means is given by

$$s_d^2 = \frac{f_y}{f^2 f_x^2}\left(\sum_{i=1}^{n} w_i^2\right) p_x(1-p_x) + \frac{1}{f^2}\left(\sum_{i=n+1}^{n+m} w_i^2\right) p_y(1-p_y)$$

The z-statistic for testing the difference of means is given by

$$z = \frac{p_{1w} - p_{Tw}}{s_d}$$

MULTIPLY WEIGHTED DATA

The test takes on the same form as the t test for means, except that in this case the x's are either 0 or 1, the proportions are

$$p_{1w} = \frac{\sum_{i=1}^{n} w_{i1} x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^{n} w_{i1} + \sum_{i=n+1}^{m} w_i}$$

and

$$p_{2w} = \frac{\sum_{i=1}^{n} w_{i2} x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{\sum_{i=1}^{n} w_{i2} + \sum_{i=n+m+1}^{n+m+p} w_i}$$

and the variance of the difference of the two proportions is given by

$$\frac{\sum_{i=1}^{n} [(f_{x2} + f_z) w_{i1} - (f_{x1} + f_y) w_{i2})]^2 \sigma_x^2}{(f_{x1} + f_y)^2 (f_{x2} + f_z)^2} + \frac{\sum_{i=n+1}^{n+m} w_i^2 \sigma_y^2}{(f_{x1} + f_y)^2} + \frac{\sum_{i=n+m+1}^{n} w_i^2 \sigma_z^2}{(f_{x2} + f_z)^2}$$

In this case the variance is estimated by

$$s_d^2 = \frac{\sum_{i=1}^{n} [(f_{x2} + f_z) w_{i1} - (f_{x1} + f_y) w_{i2})]^2}{(f_{x1} + f_y)^2 (f_{x2} + f_z)^2} p_x (1 - p_x) + \frac{\sum_{i=n+1}^{n+m} w_i^2}{(f_{x1} + f_y)^2} p_y (1 - p_y) + \frac{\sum_{i=n+m+1}^{n} w_i^2}{(f_{x2} + f_z)^2} p_z (1 - p_z)$$

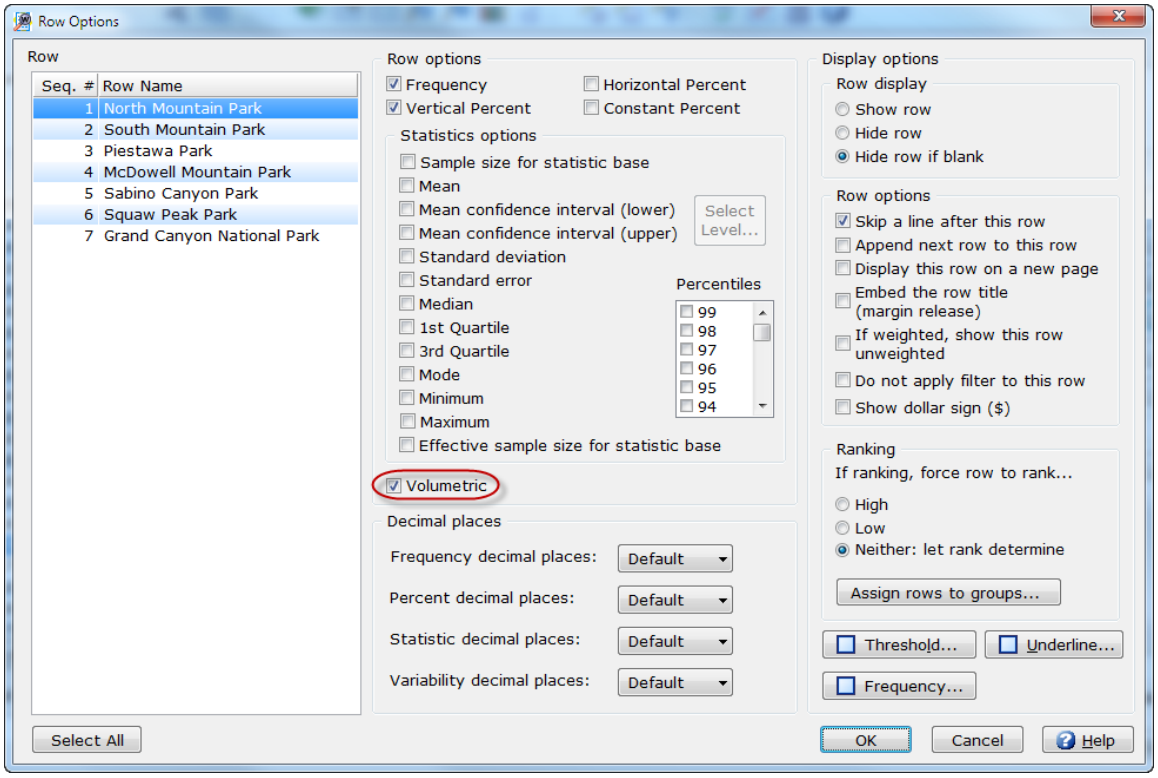
where p_x, p_y, and p_z are the unweighted proportions based on the x's, y's, and z's respectively.

The z-statistic for testing the difference of proportions is given by

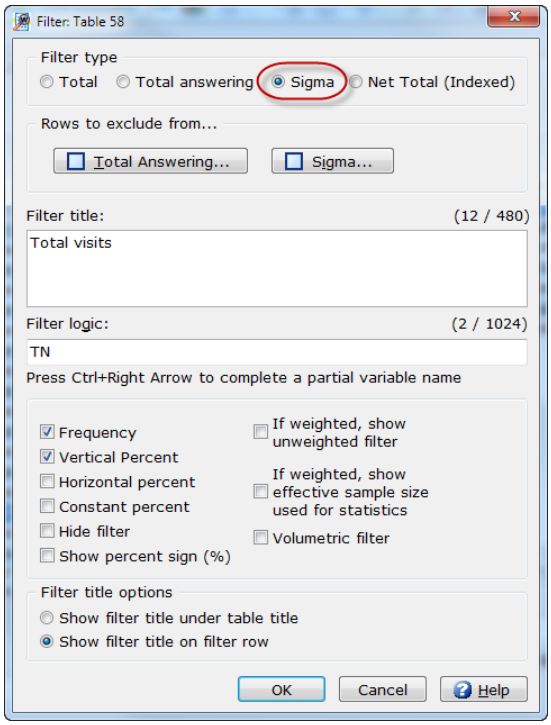
$$t = \frac{p_{1w} - p_{2w}}{s_d}$$

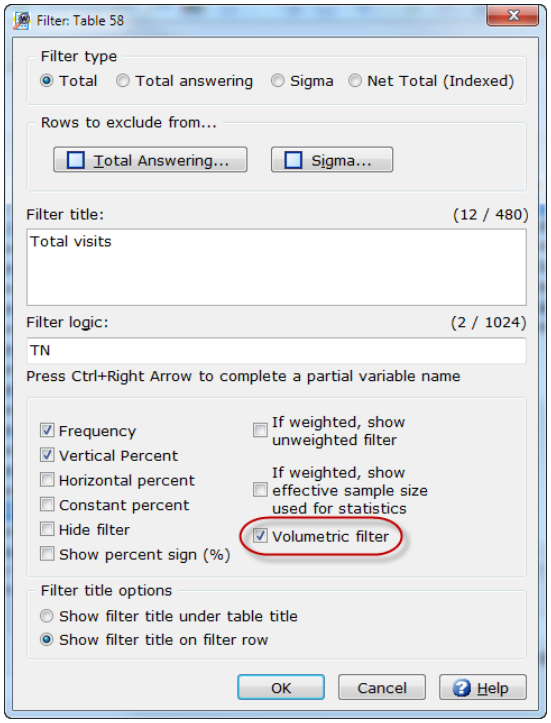
COMPARING VOLUMETRIC PERCENTAGES

WinCross provides the user with the ability to perform significance tests of differences of percentages calculated from volumetric data. One way to indicate that the row percentages are based on volumetric data, rather than on frequency data, is by selecting the **Volumetric** option on the **Row Options** dialog:



There are two other ways of indicating that the row percentages are based on volumetric data, rather than on frequency data. One is by selecting the **Volumetric filter** option on the **Filter** dialog for the table and the other is by selecting the **Sigma** option on the **Filter** dialog for the table, as seen by the following:





Finally, row percentages may be volumetric if they are generated from data calculated using WinCross's COUNT feature.

DEPENDENT PAIRED/OVERLAP (LOC+/VAR+) UNWEIGHTED DATA

Suppose we wanted to compare the percent that respondents with a given attribute contribute to a total of all respondents on that attribute. For example, suppose column 1 records the number of bottles of Coke consumed at different occasions during the week, column 2 records the number of bottles of Pepsi consumed at different occasions during the week, the total row contains the total consumption of soft drinks in the respective columns, and row 1 contains the consumption of the soft drinks at breakfast. The percentages in question here are the percentage of the total Coke consumption that is done at breakfast and the percentage of total Pepsi consumption that is done at breakfast. The possible paired/overlap situation is that there are respondents who consumed both Coke and Pepsi at breakfast during the week.

	Volume of soft drinks consumed		
	Coke	Pepsi	Sprite
Total	5539	2842	3002
	100.0%	100.0%	100.0%
breakfast	850	438	491
	15.3%	15.4%	16.4%
lunch	1424	714	785
	25.7%	25.1%	26.1%
dinner	2094	998	1084
	37.8%	35.1%	36.1%
other	1171	692	642
	21.1%	24.3%	21.3%

In this example we compare 15.3% with 15.4%.

Let us begin with the attribute measures that make up the numerator of the percentage. Let us partition the respondents so that the first n respondents provide data for both columns 1 and 2, the next m respondents provide data only for column 1 and the last p respondents provide data only for column 2. (There may be still other respondents that provided data on some, if not all, of the other banner items, but not on items 1 or 2. These will be disregarded in this analysis.)

Let us denote by x_{1i} the observed measurement for column 1 for respondent i (i = 1, 2, …, n), by x_{2i} the observed measurement for column 2 for respondent i (i = 1, 2, …, n), by y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m), and by z_i the observed measurement for respondent i (i = n+m+1, n+m+2, …, n+m+p). (I assign each of these measurements different letter names for clarity of exposition; the data are really a set of 2n+m+p observations.)

The total of the measurements for that attribute for those responding to column 1 is given by

$$X_1^+ = \sum_{i=1}^{n} x_{1i} + \sum_{i=n+1}^{n+m} y_i$$

and the total of the measurements for that attribute for those responding to column 2 is given by

$$X_2^+ = \sum_{i=1}^{n} x_{2i} + \sum_{i=n+m+1}^{n+m+p} z_i$$

Let X_1 be the total of the measurements for those responding to column 1 across all attributes and X_2 be the total of the measurements for those responding to column 2 across all attributes. Then the percentages under consideration are

$$p_1 = \frac{X_1^+}{X_1}, p_2 = \frac{X_2^+}{X_2}$$

The difference of the two percentages is given by

$$d = p_1 - p_2 = \frac{\sum_{i=1}^{n} x_{1i} + \sum_{i=n+1}^{n+m} y_i}{X_1} - \frac{\sum_{i=1}^{n} x_{2i} + \sum_{i=n+m+1}^{n+m+p} z_i}{X_2}$$
$$= (\frac{n\bar{x}_1}{X_1} - \frac{n\bar{x}_2}{X_2}) + (\frac{1}{X_1})m\bar{y} - (\frac{1}{X_2})p\bar{z}$$

where $\bar{x}_j$ is the mean of the measurements for column j (j=1,2) among those who qualified for both columns 1 and 2, $\bar{y}$ is the mean of the measurements among those who qualified only for column 1, and $\bar{z}$ is the mean of the measurements among those who qualified only for column 2.

Therefore the variance of the difference of the two percentages, conditional on the totals X_1 and X_2, is given by

$$n(\frac{\sigma_{x1}^2}{X_1^2} + \frac{\sigma_{x2}^2}{X_2^2} - \frac{2\rho\sigma_{x1}\sigma_{x2}}{X_1X_2}) + (\frac{1}{X_1})^2 m\sigma_y^2 + (\frac{1}{X_2})^2 p\sigma_z^2$$

where σ_{x1}^2 is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2, σ_{x2}^2 is the variance of the measurements in column 2 of those respondents who qualified for both columns 1 and 2, r is the correlation between the measurements in column 1 and column 2 of those respondents who qualified for both columns 1 and 2, σ_y^2 is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and σ_z^2 is the variance of the measurements in column 2 of those respondents who only qualified for column 2.

The estimate of the variance of the difference of the two percentages is given by

$$s_d^2 = n[\frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}{(n-1)X_1^2} + \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}{(n-1)X_2^2} - \frac{2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{(n-1)X_1X_2}] + m\frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{(m-1)X_1^2} + p\frac{\sum_{i=n+m+1}^{n+m+p}(z_i - \bar{z})^2}{(p-1)X_2^2}$$

SINGLY WEIGHTED DATA

Let us denote by x_{1i} the observed measurement for column 1 for respondent i (i = 1, 2, …, n), by x_{2i} the observed measurement for column 2 for respondent i (i = 1, 2, …, n), by y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m), and by z_i the observed measurement for respondent i (i = n+m+1, n+m+2, …, n+m+p). (I assign each of these measurements different letter names for clarity of exposition; the data are really a set of 2n+m+p observations.)

The weighted total of the measurements for that attribute for those responding to column 1 is given by

$$X_{1w}^{+} = \sum_{i=1}^{n} w_i x_{1i} + \sum_{i=n+1}^{n+m} w_i y_i$$

and the weighted total of the measurements for that attribute for those responding to column 2 is given by

$$X_{2w}^{+} = \sum_{i=1}^{n} w_{2i} x_{2i} + \sum_{i=n+m+1}^{n+m+p} w_{2i} z_i$$

Let X_{1w} be the weighted total of the measurements for those responding to column 1 across all attributes and X_{2w} be the weighted total of the measurements for those responding to column 2 across all attributes. Then the percentages under consideration are

$$p_{1w} = \frac{X_{1w}^{+}}{X_{1w}}, p_2 = \frac{X_{2w}^{+}}{X_{2w}}$$

The difference of the two percentages is given by

$$d = p_{1w} - p_{2w} = \frac{\sum_{i=1}^{n} w_i x_{1i} + \sum_{i=n+1}^{n+m} w_i y_i}{X_{1w}} - \frac{\sum_{i=1}^{n} w_i x_{2i} + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{X_{2w}}$$

Therefore the variance of the difference of the two percentages, conditional on the totals X_{1w} and X_{2w}, is given by

$$(\frac{\sigma_{x1}^2}{X_{1w}^2} + \frac{\sigma_{x2}^2}{X_{2w}^2} - \frac{2\rho\sigma_{x1}\sigma_{x2}}{X_{1w}X_{2w}})\sum_{i=1}^{n} w_i^2 + (\frac{1}{X_{1w}})^2 \sigma_y^2 \sum_{i=n+1}^{n+m} w_i^2 + (\frac{1}{X_{2w}})^2 \sigma_z^2 \sum_{i=n+m+1}^{n+m+p} w_i^2$$

where σ_{x1}^2 is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2, σ_{x2}^2 is the variance of the measurements in column 2 of those respondents who qualified for both columns 1 and 2, r is the correlation between the measurements in column 1 and column 2 of those respondents who qualified for both

columns 1 and 2, σ_y^2 is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and σ_z^2 is the variance of the measurements in column 2 of those respondents who only qualified for column 2.

The estimate of the variance of the difference of the two percentages is given by

$$s_d^2 = \sum_{i=1}^{n} w_i^2 [\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}{(n-1)X_{1w}^2} + \frac{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}{(n-1)X_{2w}^2} - \frac{2\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{(n-1)X_{1w}X_{2w}}] + \sum_{i=n+1}^{n+m} w_i^2 \frac{\sum_{i=n+1}^{n+m}(y_i-\bar{y})^2}{(m-1)X_{1w}^2} + \sum_{i=n+m+1}^{n+m+p} w_i^2 \frac{\sum_{i=n+m+1}^{n+m+p}(z_i-\bar{z})^2}{(p-1)X_{2w}^2}$$

MULTIPLY WEIGHTED DATA

Let us denote by x_{1i} the observed measurement for column 1 for respondent i (i = 1, 2, …, n), by x_{2i} the observed measurement for column 2 for respondent i (i = 1, 2, …, n), by y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m), and by z_i the observed measurement for respondent i (i = n+m+1, n+m+2, …, n+m+p). (I assign each of these measurements different letter names for clarity of exposition; the data are really a set of 2n+m+p observations.)

The weighted total of the measurements for that attribute for those responding to column 1 is given by

$$X_{1w}^{+} = \sum_{i=1}^{n} w_{1i} x_{1i} + \sum_{i=n+1}^{n+m} w_{1i} y_i$$

and the weighted total of the measurements for that attribute for those responding to column 2 is given by

$$X_{2w}^{+} = \sum_{i=1}^{n} w_{2i} x_{2i} + \sum_{i=n+m+1}^{n+m+p} w_{2i} z_i$$

Let X_{1w} be the weighted total of the measurements for those responding to column 1 across all attributes and X_{2w} be the weighted total of the measurements for those responding to column 2 across all attributes. Then the percentages under consideration are

$$p_{1w} = \frac{X_{1w}^{+}}{X_{1w}}, p_2 = \frac{X_{2w}^{+}}{X_{2w}}$$

The difference of the two percentages is given by

$$d = p_{1w} - p_{2w} = \frac{\sum_{i=1}^{n} w_{1i} x_{1i} + \sum_{i=n+1}^{n+m} w_{1i} y_i}{X_{1w}} - \frac{\sum_{i=1}^{n} w_{2i} x_{2i} + \sum_{i=n+m+1}^{n+m+p} w_{2i} z_i}{X_{2w}}$$

Therefore the variance of the difference of the two percentages, conditional on the totals X_{1w} and X_{2w}, is given by

$$\frac{\sigma_{x1}^2\sum_{i=1}^{n}w_{1i}^2}{X_{1w}^2}+\frac{\sigma_{x2}^2\sum_{i=1}^{n}w_{2i}^2}{X_{2w}^2}-\frac{2\rho\sigma_{x1}\sigma_{x2}\sum_{i=1}^{n}w_{1i}w_{2i}}{X_{1w}X_{2w}}+(\frac{1}{X_{1w}})^2\sigma_y^2\sum_{i=n+1}^{n+m}w_{1i}^2+(\frac{1}{X_{2w}})^2\sigma_z^2\sum_{i=n+m+1}^{n+m+p}w_{2i}^2$$

where σ_{x1}^2 is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2, σ_{x2}^2 is the variance of the measurements in column 2 of those respondents who qualified for both columns 1 and 2, r is the correlation between the measurements in column 1 and column 2 of those respondents who qualified for both columns 1 and 2, σ_y^2 is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and σ_z^2 is the variance of the measurements in column 2 of those respondents who only qualified for column 2.

The estimate of the variance of the difference of the two percentages is given by

$$s_d^2=\sum_{i=1}^{n}w_{1i}^2\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}{(n-1)X_{1w}^2}+\sum_{i=1}^{n}w_{2i}^2\frac{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}{(n-1)X_{2w}^2}-2\sum_{i=1}^{n}w_{1i}w_{2i}\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{(n-1)X_{1w}X_{2w}}$$
$$+\sum_{i=n+1}^{n+m}w_{1i}^2\frac{\sum_{i=n+1}^{n+m}(y_i-\bar{y})^2}{(m-1)X_{1w}^2}+\sum_{i=n+m+1}^{n+m+p}w_{2i}^2\frac{\sum_{i=n+m+1}^{n+m+p}(z_i-\bar{z})^2}{(p-1)X_{2w}^2}$$

DEPENDENT PAIRED/OVERLAP (MULTI) UNWEIGHTED DATA

Suppose we wanted to compare the percent that respondents with a given attribute contribute to a total of all respondents on that attribute. For example, suppose column 1 records the number of oil changes per year by people who have ever owned a Ford, column 2 records the number of oil changes per year by people who have ever owned a Chevy, the total row contains the total number of oil changes per year by people based on the respective column designations, and row 1 contains the number of oil changes per year performed at a dealer for each of the column designations. The percentages in question here are the percentages that oil changes at the dealer make up of the total number of oil changes made by Ford owners and by Chevy owners. Here is what such a table would look like:

	Number of oil changes per year by respondents who have ever owned a Ford	Chevy	VW
Total	1715	2169	1115
	100.0%	100.0%	100.0%
At dealer	822	1071	540
	47.9%	49.4%	48.4%
At garage	609	756	392
	35.5%	34.9%	35.2%
Elsewhere	284	342	183
	16.6%	15.8%	16.4%

So we want to compare 47.9% with 49.4%.

Let us begin with the attribute measures that make up the numerator of the percentage. Let us partition the respondents so that the first n respondents provide data for both columns 1 and 2 (in this example, owned both a Ford and a Chevy), the next m respondents provide data only for column 1 (in this example, owned a Ford but not a Chevy), and the last p respondents provide data only for column 2 (in this example, owned a Chevy but not a Ford). (There may be still other respondents that provided data on some, if not all, of the other banner items, but not on items 1 or 2. These will be disregarded in this analysis.)

Let us denote by x_i the observed measurement for both columns 1 and 2 for respondent i (i = 1, 2, …, n), y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m), and by z_i the observed measurement for respondent i (i = n+m+1, n+m+2, …, n+m+p). (I assign each of these measurements different letter names for clarity of exposition; the data are really a set of n+m+p observations.)

The total of the measurements for that attribute for those responding to column 1 is given by

$$X_1^+ = \sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i$$

and the total of the measurements for that attribute for those responding to column 2 is given by

$$X_2^+ = \sum_{i=1}^{n} x_i + \sum_{i=n+m+1}^{n+m+p} z_i$$

(In this example, $X_1^+ = 822$ and $X_2^+ = 1071$). Let X_1 be the total of the measurements for those responding to column 1 across all attributes (in this example, the total number of oil changes from respondents who ever owned a Ford, $X_1 = 1715$) and X_2 be the total of the measurements for those responding to column 2 across all attributes (in this example, the

total number of oil changes from respondents who ever owned a Chevy, X_2=2169). Then the percentages under consideration are

$$p_1 = \frac{X_1^+}{X_1}, p_2 = \frac{X_2^+}{X_2}$$

The difference of the two percentages is given by

$$d = p_1 - p_2 = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{X_1} - \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+m+1}^{n+m+p} z_i}{X_2}$$
$$= (\frac{1}{X_1} - \frac{1}{X_2})n\bar{x} + (\frac{1}{X_1})m\bar{y} - (\frac{1}{X_2})p\bar{z}$$

where $\bar{x}$ is the mean of the measurements for column 1 among those who qualified for both columns 1 and 2, $\bar{y}$ is the mean of the measurements among those who qualified only for column 1, and $\bar{z}$ is the mean of the measurements among those who qualified only for column 2.
Therefore the variance of the difference of the two percentages, conditional on the totals X_1 and X_2, is given by

$$(\frac{1}{X_1} - \frac{1}{X_2})^2 n\sigma_x^2 + (\frac{1}{X_1})^2 m\sigma_y^2 + (\frac{1}{X_2})^2 p\sigma_z^2$$

where σ_x^2 is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2, σ_y^2 is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and σ_z^2 is the variance of the measurements in column 2 of those respondents who only qualified for column 2,

The estimate of the variance of the difference of the two percentages is given by

$$s_d^2 = n[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}(\frac{1}{X_1} - \frac{1}{X_2})^2] + m\frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{(m-1)X_1^2} + p\frac{\sum_{i=n+m+1}^{n+m+p}(z_i - \bar{z})^2}{(p-1)X_2^2}$$

SINGLY WEIGHTED DATA

Let us denote by x_i the observed measurement for both columns 1 and 2 for respondent i (i = 1, 2, …, n), y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m), and by z_i the observed measurement for respondent i (i = n+m+1, n+m+2, …, n+m+p). (I

assign each of these measurements different letter names for clarity of exposition; the data are really a set of n+m+p observations.)

The weighted total of the measurements for that attribute for those responding to column 1 is given by

$$X_{1w}^{+} = \sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i$$

and the total of the measurements for that attribute for those responding to column 2 is given by

$$X_{2w}^{+} = \sum_{i=1}^{n} w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i$$

Let X_{1w} be the weighted total of the measurements for those responding to column 1 across all attributes (e.g., the total Coke consumption respondents of all ages who ever ate at the Ritz-Carleton) and X_{2w} be the weighted total of the measurements for those responding to column 2 across all attributes (e.g., the total Coke consumption respondents of all ages who ever ate at the Four Seasons). Then the percentages under consideration are

$$p_{1w} = \frac{X_{1w}^{+}}{X_{1w}}, p_{2w} = \frac{X_{2w}^{+}}{X_{2w}}$$

The difference of the two percentages is given by

$$d = p_{1w} - p_{2w} = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{X_{1w}} - \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{X_{2w}}$$

Therefore the variance of the difference of the two percentages, conditional on the weighted totals X_{1w} and X_{2w}, is given by

$$(\frac{1}{X_{1w}} - \frac{1}{X_{2w}})^2 \sigma_x^2 \sum_{i=1}^{n} w_i^2 + (\frac{1}{X_{1w}})^2 \sigma_y^2 \sum_{i=n+1}^{n+m} w_i^2 + (\frac{1}{X_{2w}})^2 \sigma_z^2 \sum_{i=n+m+1}^{n+m+p} w_i^2$$

where σ_x^2 is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2, σ_y^2 is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and σ_z^2 is the variance of the measurements in column 2 of those respondents who only qualified for column 2,

The estimate of the variance of the difference of the two percentages is given by

$$s_d^2 = [\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}(\frac{1}{X_{1w}} - \frac{1}{X_{2w}})^2 \sum_{i=1}^{n} w_i^2] + \frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{(m-1)X_{1w}^2} \sum_{i=n+1}^{n+m} w_i^2 + \frac{\sum_{i=n+m+1}^{n+m+p}(z_i - \bar{z})^2}{(p-1)X_{2w}^2} \sum_{i=n+m+1}^{n+m+p} w_i^2$$

MULTIPLY WEIGHTED DATA

Let us denote by x_i the observed measurement for both columns 1 and 2 for respondent i (i = 1, 2, …, n), y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m), and by z_i the observed measurement for respondent i (i = n+m+1, n+m+2, …, n+m+p). (I assign each of these measurements different letter names for clarity of exposition; the data are really a set of n+m+p observations.)

The weighted total of the measurements for that attribute for those responding to column 1 is given by

$$X_{1w}^{+} = \sum_{i=1}^{n} w_{1i} x_i + \sum_{i=n+1}^{n+m} w_{1i} y_i$$

and the total of the measurements for that attribute for those responding to column 2 is given by

$$X_{2w}^{+} = \sum_{i=1}^{n} w_{2i} x_i + \sum_{i=n+m+1}^{n+m+p} w_{2i} z_i$$

Let X_{1w} be the weighted total of the measurements for those responding to column 1 across all attributes (e.g., the total Coke consumption respondents of all ages who ever ate at the Ritz-Carleton) and X_{2w} be the weighted total of the measurements for those responding to column 2 across all attributes (e.g., the total Coke consumption respondents of all ages who ever ate at the Four Seasons). Then the percentages under consideration are

$$p_{1w} = \frac{X_{1w}^{+}}{X_{1w}}, p_{2w} = \frac{X_{2w}^{+}}{X_{2w}}$$

The difference of the two percentages is given by

$$d = p_{1w} - p_{2w} = \frac{\sum_{i=1}^{n} w_{1i} x_i + \sum_{i=n+1}^{n+m} w_{1i} y_i}{X_{1w}} - \frac{\sum_{i=1}^{n} w_{2i} x_i + \sum_{i=n+m+1}^{n+m+p} w_{2i} z_i}{X_{2w}}$$

Therefore the variance of the difference of the two percentages, conditional on the weighted totals X_{1w} and X_{2w}, is given by

$$\sigma_x^2 \sum_{i=1}^{n} \left(\frac{w_{1i}}{X_{1w}} - \frac{w_{2i}}{X_{2w}}\right)^2 + \left(\frac{1}{X_{1w}}\right)^2 \sigma_y^2 \sum_{i=n+1}^{n+m} w_{1i}^2 + \left(\frac{1}{X_{2w}}\right)^2 \sigma_z^2 \sum_{i=n+m+1}^{n+m+p} w_{2i}^2$$

where σ_x^2 is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2, σ_y^2 is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and σ_z^2 is the variance of the measurements in column 2 of those respondents who only qualified for column 2,

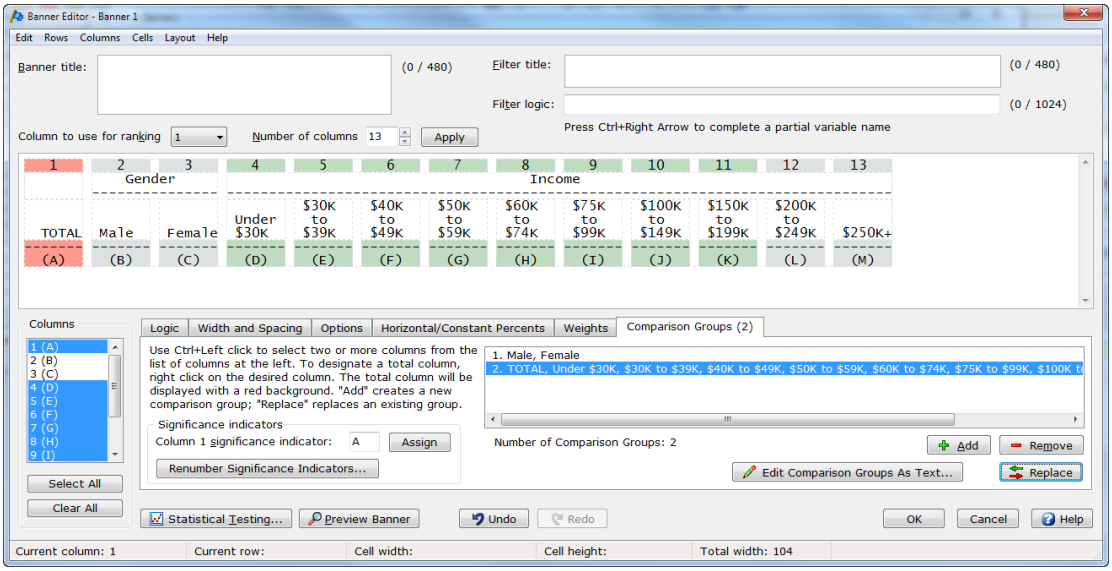
The estimate of the variance of the difference of the two percentages is given by

$$s_d^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)} \sum_{i=1}^{n} \left(\frac{w_{1i}}{X_{1w}} - \frac{w_{2i}}{X_{2w}}\right)^2 + \frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{(m-1)X_{1w}^2} \sum_{i=n+1}^{n+m} w_{1i}^2 + \frac{\sum_{i=n+m+1}^{n+m+p}(z_i - \bar{z})^2}{(p-1)X_{2w}^2} \sum_{i=n+m+1}^{n+m+p} w_{2i}^2$$

COMPARISON WITH TOTAL

Here the situation is compounded by the fact that, when one calculates a percentage based on a total for a row of a table, that total contains the total for the column which is being compared to the total column. There is therefore built in part/whole correlation between the two percentages being compared.

WinCross is told that one of the columns being used in a statistical test is a Total column by right-clicking on that column in the **Banner Editor**, as in this example:



LOC+/VAR+: UNWEIGHTED & SINGLY WEIGHTED

Here we consider the same table used above, except that now we include a Total column, where the Total column reflects all the consumption of soft drinks by all brands at various occasions. We want to compare the percentage of Coke consumption at breakfast with the percentage of all soft drink consumption at breakfast.

	Volume of soft drinks consumed			
	Total	Coke	Pepsi	Sprite
Total	14618	5539	2842	3002
	100.0%	100.0%	100.0%	100.0%
breakfast	2283	850	438	491
	15.6%	15.3%	15.4%	16.4%
lunch	3776	1424	714	785
	25.8%	25.7%	25.1%	26.1%
dinner	5381	2094	998	1084
	36.8%	37.8%	35.1%	36.1%
other	3178	1171	692	642
	21.7%	21.1%	24.3%	21.3%

The percentages to be compared are 15.6% and 15.3%.

To deal with the comparison of a column volumetric percentage with a total volumetric percentage we will need a bit of extra notation. Let n be the number of respondents and c be the number of columns in the table on which the total is based (excluding the total column, which we will refer to as column 0). Define δ_{ji} as 1 if respondent i answered item j and as 0 if respondent i did not answer item j, for i = 1, 2, …, n and j=1, 2, …, c. Let us denote by $x_{ji}\delta_{ji}$ the observed measurement for column j for respondent i. (As you can see, the δ_{ji} are used to keep track of the "no answers" in the data.) The total of the measurements for that attribute for those responding to column 1 is given by

$$X_1^+ = \sum_{i=1}^{n_1} x_{1i}\delta_{1i}$$

and the total of the measurements for that attribute for all respondents is given by

$$X_T^+ = \sum_{i=1}^{n} x_{1i}\delta_{1i} + \sum_{j=2}^{c}\sum_{i=1}^{n} x_{ji}\delta_{ji}$$

For each respondent the total will either be blank (none of the c columns contribute to the total, i.e., the respondent does not qualify for that item) or all of the columns contribute to the total, (even if the entry in any particular column is 0). Therefore in this context the δ_{ji} have the same value for all the columns, so we will designate that common value as the δ_{Ti} .

Let X_1 be the total of the measurements for those responding to column 1 across all attributes and X_T be the total of the measurements for those across columns across all attributes. Then the percentages under consideration are

$$p_1 = \frac{X_1^+}{X_1}, p_T = \frac{X_T^+}{X_T}$$

The difference of the two percentages is given by

$$d = p_1 - p_T = \frac{\sum_{i=1}^{n} x_{1i}\delta_{Ti}}{X_1} - \frac{\sum_{j=1}^{c}\sum_{i=1}^{n} x_{ji}\delta_{Ti}}{X_T}$$
$$= (\frac{1}{X_1} - \frac{1}{X_T})\sum_{i=1}^{n_1} x_{1i}\delta_{Ti} - (\frac{1}{X_T})\sum_{j=2}^{c}\sum_{i=1}^{n} x_{ji}\delta_{Ti}$$

Therefore the variance of the difference of the two percentages, conditional on the totals X_1 and X_T, is given by

$$(\frac{1}{X_1} - \frac{1}{X_T})^2\sigma_1^2\sum_{i=1}^{n}\delta_{Ti} + (\frac{1}{X_T})^2\sum_{j=2}^{c}\sigma_j^2\sum_{i=1}^{n}\delta_{Ti} - 2(\frac{1}{X_T})(\frac{1}{X_1} - \frac{1}{X_T})\sigma_1\sum_{j=2}^{c}\rho_{1j}\sigma_j\sum_{i=1}^{n}\delta_{Ti}\delta_{Ti}$$
$$= n_T[(\frac{1}{X_1} - \frac{1}{X_T})^2\sigma_1^2 + (\frac{1}{X_T})^2\sum_{j=2}^{c}\sigma_j^2 - 2(\frac{1}{X_T})(\frac{1}{X_1} - \frac{1}{X_T})\sigma_1\sum_{j=2}^{c}\rho_{1j}\sigma_j]$$

where n_T is the number of respondents contributing to the total column, σ_j^2 is the variance of the measurements in column j and 2, r_{1j} is the correlation between the measurements in column 1 and column j of those respondents who qualified for both columns 1 and j.

The estimate of the variance of the difference of the two percentages is given by

$$n_T[(\frac{1}{X_1} - \frac{1}{X_T})^2\frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\delta_{Ti}}{n_1 - 1} + (\frac{1}{X_T})^2\frac{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2\delta_{Ti}}{n_j - 1} - 2(\frac{1}{X_T})(\frac{1}{X_1} - \frac{1}{X_T})\sum_{j=2}^{c}\frac{\sum_{i=1}^{n_{1j}}(x_{ji} - \bar{x}_j)(x_{1i} - \bar{x}_1)\delta_{Ti}}{n_{1j} - 1}]$$

When the data are weighted then

$$X_{1w}^+ = \sum_{i=1}^{n_1} x_{1i}w_i\delta_{1i}$$

and the total of the measurements for that attribute for all respondents is given by

$$X_{Tw}^+ = \sum_{i=1}^{n} x_{1i}w_i\delta_{Ti} + \sum_{j=2}^{c}\sum_{i=1}^{n} x_{ji}w_i\delta_{Ti}$$

Let X_{1w} be the weighted total of the measurements for those responding to column 1 across all attributes and X_{Tw} be the total of the measurements for those across columns across all attributes. Then the percentages under consideration are

$$p_{1w} = \frac{X_{1w}^{+}}{X_{1w}}, \; p_{Tw} = \frac{X_{Tw}^{+}}{X_{Tw}}$$

The difference of the two percentages is given by

$$d_w = p_{1w} - p_{Tw} = \frac{\sum_{i=1}^{n} x_{1i} w_i \delta_{Ti}}{X_{1w}} - \frac{\sum_{j=1}^{c} \sum_{i=1}^{n} x_{ji} w_i \delta_{Ti}}{X_{Tw}}$$

$$= (\frac{1}{X_{1w}} - \frac{1}{X_{Tw}}) \sum_{i=1}^{n_1} x_{1i} w_i \delta_{Ti} - (\frac{1}{X_{Tw}}) \sum_{j=2}^{c} \sum_{i=1}^{n} x_{ji} w_i \delta_{Ti}$$

Therefore the variance of the difference of the two percentages, conditional on the totals X_1 and X_T, is given by

$$\sum_{i=1}^{n} w_i^2 \delta_{Ti} [(\frac{1}{X_{1w}} - \frac{1}{X_{Tw}})^2 \sigma_1^2 + (\frac{1}{X_{Tw}})^2 \sum_{j=2}^{c} \sigma_j^2 - 2(\frac{1}{X_{Tw}})(\frac{1}{X_{1w}} - \frac{1}{X_{Tw}}) \sigma_1 \sum_{j=2}^{c} \rho_{1j} \sigma_j]$$

The estimate of the variance of the difference of the two percentages is given by

$$\sum_{i=1}^{n} w_i^2 \delta_{1T} [(\frac{1}{X_{1w}} - \frac{1}{X_{Tw}})^2 \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}{n_1 - 1} + (\frac{1}{X_{Tw}})^2 \sum_{j=2}^{c} \frac{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2}{n_j - 1} - 2(\frac{1}{X_{Tw}})(\frac{1}{X_{1w}} - \frac{1}{X_{Tw}}) \sum_{j=2}^{c} \frac{\sum_{i=1}^{n_{1j}} (x_{ji} - \bar{x}_j)(x_{1i} - \bar{x}_1)}{n_{1j} - 1}]$$

MULTI: UNWEIGHTED & SINGLY WEIGHTED

Here we consider the same table used above, except that now we include a Total column, where the Total column reflects all the oil changes of all respondents, regardless of which car(s) they ever owned. We want to compare the percentage of dealer-performed oil changes against the total number of oil changes made by Ford owners and those made by all car owners.

	Number of oil changes per year by respondents who have ever owned a			
	Total	Ford	Chevy	VW
Total	3893	1715	2169	1115
	100.0%	100.0%	100.0%	100.0%
At dealer	1905	822	1071	540
	48.9%	47.9%	49.4%	48.4%
At garage	1348	609	756	392

	34.6%	35.5%	34.9%	35.2%
Elsewhere	640	284	342	183
	16.4%	16.6%	15.8%	16.4%

The percentages to be compared are 48.9% and 47.9%.

Let us denote by x_i the observed measurement for column 1 for respondent i (i = 1, 2, …, n), y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m). The total of the measurements for that attribute for those responding to column 1 is given by

$$X_1^+ = \sum_{i=1}^{n} x_i$$

and the total of the measurements for that attribute for those responding to the total is given by

$$X_T^+ = \sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i$$

Let X_1 be the total of the measurements for those responding to column 1 across all attributes (e.g., the total oil changes at dealer for Ford owners) and X_T be the weighted total of the measurements for all respondents across all attributes (e.g., the total oil changes at dealer for all respondents). Then the percentages under consideration are

$$p_1 = \frac{X_1^+}{X_1}, p_T = \frac{X_T^+}{X_T}$$

The difference of the two percentages is given by

$$d = p_1 - p_T = \frac{\sum_{i=1}^{n} x_i}{X_1} - \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{X_T}$$

Therefore the variance of the difference of the two percentages, conditional on the totals X_1 and X_T, is given by

$$(\frac{1}{X_1} - \frac{1}{X_T})^2 n\sigma_x^2 + (\frac{1}{X_T})^2 m\sigma_y^2$$

where σ_x^2 is the variance of the measurements in column 1 of those respondents who qualified for column 1 and σ_y^2 is the variance of the measurements in column 1 of those respondents who contributed to the total but did not qualify for column 1.

The estimate of the variance of the difference of the two percentages is given by

$$s_d^2 = n\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}(\frac{1}{X_1} - \frac{1}{X_T})^2 + m\frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{(m-1)X_T^2}$$

If the differences are weighted, then

$$d_w = p_{1w} - p_{Tw} = \frac{\sum_{i=1}^{n} w_i x_i}{X_{1w}} - \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} w_i y_i}{X_{Tw}}$$

where X_{1w} is the weighted total of the measurements for those responding to column 1 across all attributes and X_{Tw} is the weighted total of the measurements for all respondents across all attributes. Then the variance of the difference of the two weighted percentages, conditional on the totals X_{1w} and X_{Tw}, is given by

$$(\frac{1}{X_{1w}} - \frac{1}{X_{Tw}})^2 \sigma_x^2 \sum_{i=1}^{n} w_i^2 + (\frac{1}{X_{Tw}})^2 \sigma_y^2 \sum_{i=n+1}^{n+m} w_i^2$$

The estimate of the variance of the difference of the two weighted percentages is given by

$$s_d^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}(\frac{1}{X_{1w}} - \frac{1}{X_{Tw}})^2 \sum_{i=1}^{n} w_i^2 + \frac{\sum_{i=n+1}^{n+m}(y_i - \bar{y})^2}{(m-1)X_{Tw}^2} \sum_{i=n+1}^{n+m} w_i^2$$

NET PROMOTER SCORE

Consider a multinomial population with k categories whose corresponding means are θ_1, …, θ_k. Assume that we have a random sample of n observations from this population, with observed category proportions p_1, p_2, …, p_k, where

$$\sum_{i=1}^{k} p_i = 1$$

Since the sum of the p's is 1, the p's are correlated. Suppose we wish to test the hypothesis that $\theta_1 = \theta_1$, based on the statistic $p_1 - p_2$. As shown by Wilks[3], the standard deviation of the statistic $p_1 - p_2$ is

$$\sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)+2\theta_1\theta_2}{n}} = \sqrt{\frac{\theta_1+\theta_2-(\theta_1-\theta_2)^2}{n}}$$

Since the θ's are unknown, we estimate them by their sample estimate, and use as a z statistic the quantity

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1+p_2-(p_1-p_2)^2}{n}}}$$

To apply this to NPSs we let k=3, with subscript 1 representing "promoters," subscript 2 representing "detractors," and subscript 3 representing "passives."

To develop the comparable procedure when the data are weighted, assume that the j-th member of the sample is given weight w_j, j=1, … , n, We represent the data as a collection of nk 1's and 0's, with x_{ij} being a 1 if respondent j responded positively to category I and x_{ij} =0 otherwise. Then for example p_1 can be written as

$$p_1 = \frac{\sum_{j=1}^{n} x_{1j}}{n}$$

Weighting

The weighted p_{w1} can be expressed as

$$p_{w1} = \frac{\sum_{j=1}^{n} w_j x_{1j}}{\sum_{j=1}^{n} w_j}$$

The determination of the standard deviation of the statistic $p_{w1} - p_{w2}$ is based on the following.

[3] Wilks, S.S. 1940. Confidence Limits and Critical Differences Between Proportions. Public Opinion Quarterly, 4:2 (June), 332-8.

$$V(p_{w1}-p_{w2})=V[\frac{\sum_{j=1}^{n}w_j x_{1j}}{\sum_{j=1}^{n}w_j}-\frac{\sum_{j=1}^{n}w_j x_{2j}}{\sum_{j=1}^{n}w_j}]$$

$$=\frac{\sum_{j=1}^{n}w_j^2 V(x_{1j}-x_{2j})}{(\sum_{j=1}^{n}w_j)^2}$$

But

$$V(x_{1j}-x_{2j})=V(x_{1j})+V(x_{2j})-2Cov(x_{1j},x_{2j})$$
$$=\theta_1(1-\theta_1)+\theta_2(1-\theta_2)+2\theta_1\theta_2$$

So

$$V(p_{w1}-p_{w2})=\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)+2\theta_1\theta_2}{f}$$

where f is the effective sample size

$$f=\frac{(\sum_{j=1}^{n}w_j)}{\sum_{j=1}^{n}w_j^2}$$

Thus our z statistic is given by

$$z_w=\frac{p_{w1}-p_{w2}}{\sqrt{\frac{p_1+p_2-(p_1-p_2)^2}{f}}}$$

COMPARISON OF NET PROMOTER SCORES DEPENDENT PAIRED/OVERLAP (LOC+/VAR+)

Simple case: sampling from the same multinomial population

Consider a multinomial population with k categories whose corresponding means are θ_1, …, θ_k. Assume that we have two random samples from this population, of sizes n_1 and n_2, respectively. Suppose further that the first n_0 observations from the two populations are paired (for example, population 1 relates to a "treatment," population 2 relates to a "control," and the first n_0 observations are taken from the same respondent; for another example, population 1 relates to Coke, population 2 relates to Pepsi, and the first n_0 pairs of responses are taken from the same respondent).

We represent the data sets as collections of kn_1 and kn_2 1's and 0's, with x_{ija} being a 1 if respondent j of sample a responded positively to category I and x_{ija} =0 otherwise, where a=1, 2. Then for example p_{1a} can be written as

$$p_{1a} = \frac{\sum_{j=1}^{n_a} x_{1ja}}{n_a} = \frac{\sum_{j=1}^{n_0} x_{1ja} + \sum_{j=n_0+1}^{n_a} x_{1ja}}{n_a}$$

Let subscript 1 represent the "promoters" and subscript 2 representing the "detractors." The NPS for sample a is $L_a=p_{1a}-p_{2a}$. So

$$L_a = \frac{\sum_{j=1}^{n_0} x_{1ja} + \sum_{j=n_0+1}^{n_a} x_{1ja} - \sum_{j=1}^{n_0} x_{2ja} - \sum_{j=n_0+1}^{n_a} x_{2ja}}{n_a}$$

To test the equality of the NPSs from the two samples, we must determine the variance of $L_1 - L_2$. But

$$L_1 - L_2 = \frac{\sum_{j=1}^{n_0} x_{1j1} + \sum_{j=n_0+1}^{n_1} x_{1j1} - \sum_{j=1}^{n_0} x_{2j1} - \sum_{j=n_0+1}^{n_1} x_{2j1}}{n_1} - \frac{\sum_{j=1}^{n_0} x_{1j2} + \sum_{j=n_0+1}^{n_2} x_{1j2} - \sum_{j=1}^{n_0} x_{2j2} - \sum_{j=n_0+1}^{n_2} x_{2j2}}{n_2}$$

$$= \frac{n_2 \sum_{j=1}^{n_0} (x_{1j1} - x_{2j1}) - n_1 \sum_{j=1}^{n_0} (x_{1j2} - x_{2j2})}{n_1 n_2} + \frac{\sum_{j=n_0+1}^{n_1} (x_{1j1} - x_{2j1})}{n_1} - \frac{\sum_{j=n_0+1}^{n_2} (x_{1j2} - x_{2j2})}{n_2}$$

Note that the second and third terms of the above are just the NPSs of the unique data from samples 1 and 2. So the variances of those two terms are

$$\frac{\theta_1 + \theta_2 - (\theta_1 - \theta_2)^2}{n_a}$$

Since we have assumed that the same multinomial distribution governs both samples, we can estimate θ_1 by the fraction of "promoters" in both samples (call that fraction p_1) and estimate θ_2 by the fraction of "detractors" in both samples (call that fraction p_2).

The variance of the first term is $\frac{(n_1^2 + n_2^2) n_0 (\theta_1 + \theta_2 - (\theta_1 - \theta_2)^2) - 2 n_0 n_1 n_2 c}{n_1^2 n_2^2}$

where c is the covariance of $(x_{1j1} - x_{2j1})$ with $(x_{1j2} - x_{2j2})$, But each of these quantities is equal to 1 if the j-th respondents is a "promoter" in sample a, 0 if the j-th respondents is a "passive" in sample a, and -1 if the j-th respondents is a "detractor" in sample a, a = 1,2.

Of the n_0 common respondents in both samples 1 and 2, let n_{011} be the number of "promoters" in both samples 1 and 2, n_{022} be the number of "detractors" in both samples 1 and 2, n_{012} be the number of "promoters" in sample 1 and "detractors" in sample 2, and

n_{021} be the number of "promoters" in sample 2 and "detractors" in sample 1. Then p* = $(n_{011} + n_{022} - n_{012} - n_{021})/n_0$ is an unbiased estimate of the expected value of $(x_{1j1} - x_{2j1})(x_{1j2} - x_{2j2})$. The expected value of both $(x_{1j1} - x_{2j1})$ and $(x_{1j2} - x_{2j2})$ is $\theta_1 - \theta_2$, so an unbiased estimate of c is $p^* - (p_1 - p_2)^2$.

Putting all this together, we see that an estimate of the variance of $L_1 - L_2$ is

$$\frac{(n_1^2 + n_2^2)n_0(p_1 + p_2 - (p_1 - p_2)^2) - 2n_0 n_1 n_2(n^* - (p_1 - p_2)^2)}{n_1^2 n_2^2} + \frac{p_1 + p_2 - (p_1 - p_2)^2}{n_1} + \frac{p_1 + p_2 - (p_1 - p_2)^2}{n_2}$$

$$= [p_1 + p_2 - (p_1 - p_2)^2][\frac{(n_1^2 + n_2^2)n_0}{n_1^2 n_2^2} + \frac{1}{n_1} + \frac{1}{n_2}] - \frac{2n_0(p^* - (p_1 - p_2)^2)}{n_1 n_2}$$

Thus our z-statistic for testing the equality of the two NPSs is

$$z = \frac{L_1 - L_2}{\sqrt{[p_1 + p_2 - (p_1 - p_2)^2][\frac{(n_1^2 + n_2^2)n_0}{n_1^2 n_2^2} + \frac{1}{n_1} + \frac{1}{n_2}] - \frac{2n_0(p^* - (p_1 - p_2)^2)}{n_1 n_2}}}$$

Notice that when n_0=0 we are dealing with no overlap and two separate samples from the sample multinomial population, so this reduces to

$$z = \frac{L_1 - L_2}{\sqrt{[p_1 + p_2 - (p_1 - p_2)^2]\frac{[1}{n_1} + \frac{1}{n_2}]}}$$

General case: sampling from separate multinomial populations

The above development was based on the assumption that the same k category multinomial population with means are $\theta_1, \ldots, \theta_k$ govern both samples. This may not be the case even when the NPSs for the two populations are identical. Suppose instead that the a-th sample is governed by parameters $\theta_{1a}, \ldots, \theta_{ka}$, a=1,2. Then for example the population 1 NPS $\lambda_1 = \theta_{11} - \theta_{21}$ can equal the population 2 NPS $\lambda_2 = \theta_{12} - \theta_{22}$ without both $\theta_{11} = \theta_{12}$ and $\theta_{21} = \theta_{22}$.

Though the formula for $L_1 - L_2$ is the same as above, we have to estimate each of the θ_{ia} separately by p_{ia}, the proportion in sample a. The variance of each of the last two terms of $L_1 - L_2$ is estimated by

$$\frac{p_{1a} + p_{2a} - (p_{1a} - p_{2a})^2}{n_a}$$

The variance of the first term is

$$\frac{n_2^2 n_0(\theta_{11}+\theta_{21}-(\theta_{11}-\theta_{21})^2+n_1^2 n_0(\theta_{12}+\theta_{22}-(\theta_{12}-\theta_{22})^2)-2n_0 n_1 n_2 c}{n_1^2 n_2^2}$$

The expected value of $x_{1j1} - x_{2j1}$ is $\theta_{11} - \theta_{21}$ and that of $x_{1j2} - x_{2j2}$ is $\theta_{12} - \theta_{22}$, so an unbiased estimate of c is $p^* - (p_{11} - p_{21})(p_{12} - p_{22})$.

Putting all this together, we see that an estimate of the variance of $L_1 - L_2$ is

$$v = \frac{n_2^2 n_0(p_{11}+p_{21}-(p_{11}-p_{21})^2)+n_1^2 n_0(p_{12}+p_{22}-(p_{12}-p_{22})^2)-2n_0 n_1 n_2(p^*-(p_{11}-p_{21})(p_{12}-p_{22}))}{n_1^2 n_2^2}$$

$$+\frac{p_{11}+p_{21}-(p_{11}-p_{21})^2}{n_1}+\frac{p_{12}+p_{22}-(p_{12}-p_{22})^2}{n_2}$$

and the z-statistic is

$$z = \frac{L_1 - L_2}{\sqrt{v}}$$

Notice that when $n_0=0$ we are dealing with no overlap and two independent samples from separate multinomial populations, so this reduces to the sum of variances of the two NPSs

$$v = \frac{p_{11}+p_{21}-(p_{11}-p_{21})^2}{n_1}+\frac{p_{12}+p_{22}-(p_{12}-p_{22})^2}{n_2}$$

Weighting

The difference of weighted NPSs is

$$L_{w1} - L_{w2} = \frac{\sum_{j=1}^{n_0} w_{j1}x_{1j1} + \sum_{j=n_0+1}^{n_1} w_{j1}x_{1j1} - \sum_{j=1}^{n_0} w_{j1}x_{2j1} - \sum_{j=n_0+1}^{n_1} w_{j1}x_{2j1}}{\sum_{j=1}^{n_1} w_{j1}}$$

$$- \frac{\sum_{j=1}^{n_0} w_{j2}x_{1j2} + \sum_{j=n_0+1}^{n_2} w_{j2}x_{1j2} - \sum_{j=1}^{n_0} w_{j2}x_{2j2} - \sum_{j=n_0+1}^{n_2} w_{j2}x_{2j2}}{\sum_{j=1}^{n_2} w_{j2}}$$

$$= \frac{w_2 \sum_{j=1}^{n_0} (w_{j1}x_{1j1} - w_{j2}x_{2j1}) - w_1 \sum_{j=1}^{n_0} (w_{j2}x_{1j2} - w_{j2}x_{2j2})}{w_1 w_2}$$

$$+ \frac{\sum_{j=n_0+1}^{n_1} (w_{j1}x_{1j1} - w_{j1}x_{2j1})}{w_1} - \frac{\sum_{j=n_0+1}^{n_2} (w_{j2}x_{1j2} - w_{j2}x_{2j2})}{w_2}$$

where

$$w_1 = \sum_{j=1}^{n_1} w_{j1}, w_2 = \sum_{j=1}^{n_2} w_{j2}$$

So the variance of $L_{w1} - L_{w2}$ is

$$v_w = \frac{w_2^2 \sum_{j=1}^{n_0} w_{j1}^2 V(x_{1j1}) + w_{j2}^2 V(x_{2j1}) - 2w_{j1}^2 w_{j2}^2 C(x_{1j1}, x_{2j1})}{w_1^2 w_2^2}$$

$$+ \frac{w_1^2 \sum_{j=1}^{n_0} w_{j2}^2 [V(x_{1j2}) + V(x_{2j2}) - 2C(x_{1j2}, x_{2j2})]}{w_1^2 w_2^2}$$

$$+ \frac{\sum_{j=n_0+1}^{n_1} w_{j1}^2 V(x_{1j1} - x_{2j1})}{w_1^2} - \frac{\sum_{j=n_0+1}^{n_2} w_{j2}^2 V(x_{1j2} - x_{2j2})}{w_2^2}$$

. Each of the last two terms of v_w is estimated by

$$\frac{p_{1a} + p_{2a} - (p_{1a} - p_{2a})^2}{f_a^*}$$

where

$$f_a^* = \frac{w_a^2}{\sum_{j=n_0+1}^{n_a} w_{ja}^2} = \frac{(\sum_{j=1}^{n_1} w_{ja})^2}{\sum_{j=n_0+1}^{n_a} w_{ja}^2}$$

The first term is

$$\frac{w_2^2(\theta_{11} + \theta_{21} - (\theta_{11} - \theta_{21})^2 \sum_{j=1}^{n_0} w_{j1}^2 + w_1^2(\theta_{12} + \theta_{22} - (\theta_{12} - \theta_{22})^2) \sum_{j=1}^{n_0} w_{j2}^2 - 2w_1 w_2 c \sum_{j=1}^{n_0} w_{j1} w_{j2}}{w_1^2 w_2^2}$$

$$= \frac{\theta_{11} + \theta_{21} - (\theta_{11} - \theta_{21})^2 \sum_{j=1}^{n_0} w_{j1}^2}{w_1^2} + \frac{\theta_{12} + \theta_{22} - (\theta_{12} - \theta_{22})^2 \sum_{j=1}^{n_0} w_{j2}^2}{w_2^2} - 2\frac{c \sum_{j=1}^{n_0} w_{j1} w_{j2}}{w_1 w_2}$$

As above, an unbiased estimate of c is p* - ($p_{11} - p_{21}$)($p_{12} - p_{22}$). Let us define f* by the relation

$$\frac{1}{f^*} = \frac{\sum_{j=1}^{n_0} w_{j1} w_{j2}}{w_1 w_2}$$

and f_{a0} as

$$\frac{1}{f_{a0}} = \frac{\sum_{j=1}^{n_0} w_{ja}^2}{w_a^2}$$

Putting all this together, we see that an estimate of the variance of $Lw_1 - Lw_2$ is

$$v_w = \frac{p_{11} + p_{21} - (p_{11} - p_{21})^2}{f_{10}} + \frac{p_{12} + p_{22} - (p_{12} - p_{22})^2}{f_{20}} - \frac{2(p^* - (p_{11} - p_{21})(p_{12} - p_{22}))}{f^*}$$

$$+ \frac{p_{11} + p_{21} - (p_{11} - p_{21})^2}{f_1^*} + \frac{p_{12} + p_{22} - (p_{12} - p_{22})^2}{f_2^*}$$

COMPARISON OF NET PROMOTER SCORES DEPENDENT PAIRED/OVERLAP (MULTI)

Suppose we wanted to compare the NPSs of a sample of respondents for those qualifying for criterion 1 (e.g., drank Coke) with the NPSs of the same sample of respondents for

those qualifying for criterion 2 (e.g., drank Pepsi). Note that a given respondent could qualify for both criterion 1 and criterion 2, so that his response (1 if "promoter", -1 if "detractor", and 0 if "passive") is used in the calculation of both NPSs.

Let us partition the respondents so that the first n_0 respondents qualify for both criteria 1 and 2, the next $n_1 - n_0$ respondents qualify for only criterion 1, and the last $n_2 - n_0$ respondents qualify only for criterion 2. (There may be still other respondents that qualify for some, if not all, of the other criteria, but not criteria 1 or 2. These will be disregarded in this analysis.)

Let subscript 1 represent the "promoters" and subscript 2 representing the "detractors." Let $x_{1ja} = 1$ if respondent j is a "promoter" who qualifies for criterion a, and $x_{2ja} = 1$ if respondent j is a "detractor" who qualifies for criterion a, a = 1, 2. The NPS for criterion a is given by the following:

$$L_a = \frac{\sum_{j=1}^{n_0} x_{1ja} + \sum_{j=n_0+1}^{n_a} x_{1ja} - \sum_{j=1}^{n_0} x_{2ja} - \sum_{j=n_0+1}^{n_a} x_{2ja}}{n_a}$$

so the difference in NPSs is given by

$$L_1 - L_2 = \frac{\sum_{j=1}^{n_0} x_{1j1} + \sum_{j=n_0+1}^{n_a} x_{1j1} - \sum_{j=1}^{n_0} x_{2j1} - \sum_{j=n_0+1}^{n_a} x_{2j1}}{n_1} - \frac{\sum_{j=1}^{n_0} x_{1j2} + \sum_{j=n_0+1}^{n_a} x_{1j2} - \sum_{j=1}^{n_0} x_{2j2} - \sum_{j=n_0+1}^{n_a} x_{2j2}}{n_2}$$

$$= [\frac{1}{n_1} - \frac{1}{n_2}][\sum_{j=1}^{n_0} x_{1j1} - \sum_{j=1}^{n_0} x_{2j1}] + \frac{\sum_{j=n_0+1}^{n_1} x_{1j1} - \sum_{j=n_0+1}^{n_0} x_{2j1}}{n_1} - \frac{\sum_{j=n_0+1}^{n_2} x_{1j2} - \sum_{j=n_0+1}^{n_2} x_{2j2}}{n_2}$$

since for the first n_0 respondents $x_{1j1} \equiv x_{1j2}$ and $x_{2j1} \equiv x_{2j2}$.

In what follows we assume that the underlying multinomial probabilities that generate the score are fixed and do not depend on the criteria under study. Otherwise we would have to assume one multinomial distribution for those qualifying for both items 1 and 2, a second multinomial distribution for those qualifying for item 1 but not item 2, and a third multinomial distribution for those qualifying for item 2 but not item 1. In this case we estimate the proportion of promoters by

$$p_1 = \frac{\sum_{j=1}^{n_0} x_{1j1} + \sum_{j=n_0+1}^{n_1} x_{1j1} + \sum_{j=n_0+1}^{n_2} x_{2j1}}{n_0 + n_1 + n_2}$$

and the proportion of detractors by

$$p_2 = \frac{\sum_{j=1}^{n_0} x_{1j2} + \sum_{j=n_0+1}^{n_1} x_{1j2} + \sum_{j=n_0+1}^{n_2} x_{2j2}}{n_0 + n_1 + n_2}$$

The variance of the difference of the NPSs is estimated by

$$v = [\frac{1}{n_1} - \frac{1}{n_2}]^2 n_0[p_1 + p_2 - (p_1 - p_2)^2] + \frac{1}{n_1^2}(n_1 - n_0)[p_1 + p_2 - (p_1 - p_2)^2]$$
$$+ \frac{1}{n_2^2}(n_2 - n_0)[p_1 + p_2 - (p_1 - p_2)^2]$$
$$= [\frac{1}{n_1} + \frac{1}{n_2} - \frac{2n_0}{n_1 n_2}][p_1 + p_2 - (p_1 - p_2)^2]$$

The z score is therefore

$$z = \frac{L_1 - L_2}{\sqrt{v}}$$

If one does not want to make this simplifying assumption, then one must estimate the proportion of promoters and the proportion of detractors separately for each of three groups: group 0 (those who qualify for both items 1 and 2, group 1 (those who qualify for item 1 and not for item 2), and group 3 (those who qualify for item 2 but not for item 1). These estimates are as follows:

$$p_{10} = \frac{\sum_{j=1}^{n_0} x_{1j1}}{n_0} \quad p_{11} = \frac{\sum_{j=n_0+1}^{n_1} x_{1j1}}{n_1 - n_0} \quad p_{12} = \frac{\sum_{j=n_0}^{n_2} x_{2j1}}{n_2 - n_0}$$
$$p_{20} = \frac{\sum_{j=1}^{n_0} x_{1j2}}{n_0} \quad p_{21} = \frac{\sum_{j=n_0+1}^{n_1} x_{1j2}}{n_1 - n_0} \quad p_{22} = \frac{\sum_{j=n_0}^{n_2} x_{2j2}}{n_2 - n_0}$$

Then the variance of the difference of the NPSs is estimated by

$$v = [\frac{1}{n_1} - \frac{1}{n_2}]^2 n_0[p_{10} + p_{20} - (p_{10} - p_{20})^2] + \frac{1}{n_1^2}(n_1 - n_0)[p_{11} + p_{21} - (p_{11} - p_{21})^2]$$
$$+ \frac{1}{n_2^2}(n_2 - n_0)[p_{12} + p_{22} - (p_{12} - p_{22})^2]$$

Weighting

The difference of weighted NPSs is

$$L_{w1} - L_{w2} = \frac{\sum_{j=1}^{n_0} w_{j1}x_{1j1} + \sum_{j=n_0+1}^{n_1} w_{j1}x_{1j1} - \sum_{j=1}^{n_0} w_{j1}x_{2j1} - \sum_{j=n_0+1}^{n_1} w_{j1}x_{2j1}}{\sum_{j=1}^{n_1} w_{j1}}$$

$$- \frac{\sum_{j=1}^{n_0} w_{j2}x_{1j2} + \sum_{j=n_0+1}^{n_2} w_{j2}x_{1j2} - \sum_{j=1}^{n_0} w_{j2}x_{2j2} - \sum_{j=n_0+1}^{n_2} w_{j2}x_{2j2}}{\sum_{j=1}^{n_2} w_{j2}}$$

$$= \frac{w_2 \sum_{j=1}^{n_0} w_{j1}(x_{1j1} - x_{2j1}) - w_1 \sum_{j=1}^{n_0} w_{j2}(x_{1j2} - x_{2j2})}{w_1 w_2}$$

$$+ \frac{\sum_{j=n_0+1}^{n_1} (w_{j1}x_{1j1} - w_{j1}x_{2j1})}{w_1} - \frac{\sum_{j=n_0+1}^{n_2} (w_{j2}x_{1j2} - w_{j2}x_{2j2})}{w_2}$$

$$= \frac{\sum_{j=1}^{n_0} (w_2 w_{j1} - w_1 w_{j2})(x_{1j1} - x_{2j1})}{w_1 w_2}$$

$$+ \frac{\sum_{j=n_0+1}^{n_1} (w_{j1}x_{1j1} - w_{j1}x_{2j1})}{w_1} - \frac{\sum_{j=n_0+1}^{n_2} (w_{j2}x_{1j2} - w_{j2}x_{2j2})}{w_2}$$

where

$$w_1 = \sum_{j=1}^{n_1} w_{j1}, w_2 = \sum_{j=1}^{n_2} w_{j2}$$

Each of the last two terms of v_w is estimated by

$$\frac{\sum_{j=1}^{n_0}(w_2 w_{j1} - w_1 w_{j2})^2 (p_1 + p_2 - (p_1 - p_2)^2)}{w_1^2 w_2^2}$$

$$+ \frac{\sum_{j=n_0+1}^{n_1} w_{j1}^2 (p_1 + p_2 - (p_1 - p_2)^2)}{w_1^2} + \frac{\sum_{j=n_0+1}^{n_2} w_j^2 (p_1 + p_2 - (p_1 - p_2)^2)}{w_2^2}$$

$$= [\frac{\sum_{j=1}^{n_0}(w_2 w_{j1} - w_1 w_{j2})^2}{w_1^2 w_2^2} + \frac{\sum_{j=n_0+1}^{n_1} w_{j1}^2}{w_1^2} \frac{\sum_{j=n_0+1}^{n_2} w_{j2}^2}{w_2^2}](p_1 + p_2 - (p_1 - p_2)^2)$$

where p_1 and p_2 are defined above.

Let

$$f_a^* = \frac{w_a^2}{\sum_{j=n_0+1}^{n_a} w_{ja}^2} = \frac{(\sum_{j=1}^{n_1} w_{ja})^2}{\sum_{j=n_0+1}^{n_a} w_{ja}^2}$$

and let

$$f_{0a}^* = \frac{w_a^2}{\sum_{j=1}^{n_0} w_{ja}^2} = \frac{(\sum_{j=1}^{n_1} w_{ja})^2}{\sum_{j=1}^{n_a} w_{ja}^2}$$

The estimated variance is

$$v_w == [\frac{1}{f_{01}^*} + \frac{1}{f_{02}^*} - \frac{2\sum_{j=1}^{n_0} w_{j1} w_{j2}}{w_1 w_2} \frac{1}{f_1^*} + \frac{1}{f_2^*}](p_1 + p_2 - (p_1 - p_2)^2)$$

The z score is therefore

$$z_w = \frac{L_{1w} - L_{2w}}{\sqrt{v_w}}$$

If one does not want to make the simplifying assumption that the multinomial populations are identical, then one uses the p_{0i}, p_{11}, p_{21}, p_{02}, p_{12}, p_{22} defined above, to calculate

$$v = \frac{\sum_{j=1}^{n_0}(w_2 w_{j1} - w_1 w_{j2})^2 (p_{01} + p_{02} - (p_{01} - p_{02})^2)}{w_1^2 w_2^2}$$

$$+ \frac{\sum_{j=n_0+1}^{n_1} w_{j1}^2 (p_{11} + p_{12} - (p_{11} - p_{12})^2)}{w_1^2} + \frac{\sum_{j=n_0+1}^{n_2} w_{j2}^2 (p_{21} + p_{22} - (p_{21} - p_{22})^2)}{w_2^2}$$

$$= (p_{01} + p_{02} - (p_{01} - p_{02})^2) \frac{\sum_{j=1}^{n_0} w_2^2 w_{j1} + w_1^2 w_{j1}^2 - 2 w_1 w_2 w_{j1} w_{j2}}{w_1^2 w_2^2}$$

$$+ (p_{11} + p_{12} - (p_{11} - p_{12})^2) \frac{\sum_{j=n_0+1}^{n_1} w_{j1}^2}{w_1^2} + (p_{21} + p_{22} - (p_{21} - p_{22})^2) \frac{\sum_{j=n_0+1}^{n_2} w_{2j}^2}{w_2^2}$$

$$= p_{01} + p_{02} - (p_{01} - p_{02})^2)[\frac{\sum_{j=1}^{n_0} w_{j1}^2}{w_1^2} + \frac{\sum_{j=1}^{n_0} w_{j2}^2}{w_2^3} - 2\frac{\sum_{j=1}^{n_0} w_{j1} w_{j2}}{w_1 w_2}]$$

$$+ (p_{11} + p_{12} - (p_{11} - p_{12})^2) \frac{\sum_{j=n_0+1}^{n_1} w_{j1}^2}{w_1^2} + (p_{21} + p_{22} - (p_{21} - p_{22})^2) \frac{\sum_{j=n_0+1}^{n_2} w_{2j}^2}{w_2^2}$$

ONEWAY ANOVA

General Notation

The one-way analysis of variance (anova) is a statistical procedure to test, based on independent samples from each of m populations, whether the set of m population means are identical or not. When m=2 the appropriate procedure is the t-test, and so the one-way anova is a generalization of this test.

One might ask, "Why not separately test each of the m(m-1)/2 pairs of means using the t-test for each pairing?" The problem with this is that each time one performs a statistical test there is a probability of making the Type I Error of rejecting the null hypothesis of no difference when in fact there is truly a difference between the means. One normally presets this probability (usually referred to as **α**, the level of significance) at some low level, such as 0.05 or 0.01. If one presets this probability at 0.05, then on average one will make a Type I Error once out of every 20 times one performs a significance test. And if one has m=7 populations and performs m(m-1)/2 = 21 t tests then one will on average reject the hypothesis of no difference when in fact there is no difference between the means being compared. Each of the procedures in WinCross under the Oneway anova header is designed to circumvent this problem in a different way. The specifics of the procedures will be presented in turn. But first let us establish some general terminology.

Let $n_1, n_2, \ldots, n_m$ denote the sample sizes from the m populations, and let x_{ij} (i=1, 2, …, m, j= 1, 2, … , n_i) denote the observations. Let $\bar{x}_i$ denote the sample mean of the data from population i, and let $\bar{x}$ denote the mean of all the data, i.e.,

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, \quad \bar{x} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{m} n_i} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i}$$

It makes our exposition of the statistical testing methodology easier if we assume that the indexing of the populations is such that $\bar{x}_1 \le \bar{x}_2 \le \ldots \le \bar{x}_m$. In the first step of each of the procedures, $\bar{x}_1$ is compared with $\bar{x}_m$. If the difference $\bar{x}_m - \bar{x}_1$ is less than an appropriate critical value c_m then we conclude that all the population means are not significantly different, and each of the m means are labeled #1.

Otherwise we can assert that the mean of population m is significantly higher than that of population 1, and we now continue to search to check each of the two subsets of m-1 means, $\bar{x}_2, \bar{x}_3, \ldots, \bar{x}_m$ and $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{m-1}$ to see if they are homogeneous. To check the first subset $\bar{x}_2, \bar{x}_3, \ldots, \bar{x}_m$ we compare $\bar{x}_m - \bar{x}_2$ with an appropriate critical value c_{m-1}, and, if that difference is less than the critical value, then we conclude that the m-1 population means of the first subset are not significantly different, and each of these m-1 means are labeled #1. Otherwise we can assert that the mean of population m is significantly higher than

that of population 2 and we now continue to search to check each of the two subsets of m-2 means, $\bar{x}_1, \bar{x}_2, ..., \bar{x}_{m-2}$ and $\bar{x}_2, \bar{x}_3, ..., \bar{x}_{m-1}$ to see if they are homogeneous.

Similarly, to check the second subset $\bar{x}_1, \bar{x}_2, ..., \bar{x}_{m-1}$ we compare $\bar{x}_{m-1} - \bar{x}_1$ with the same critical value c_{m-1}, and, if that difference is less than the critical value, then we conclude that the m-1 population means of the second subset are not significantly different and each of these m-1 means are labeled #2.

To summarize to this point: If we found no significant difference in the first subset $\bar{x}_2, \bar{x}_3, ..., \bar{x}_m$ and have labeled each of them with a #1 and no significant difference in the second subset, $\bar{x}_1, \bar{x}_2, ..., \bar{x}_{m-1}$, and have labeled each of them with a #2, then x_1 will be labeled #2, x_m will be labeled #1, and each of $\bar{x}_2, \bar{x}_3, ..., \bar{x}_{m-1}$ will be labeled both #1 and #2.

If we did find a significant difference in the first subset $\bar{x}_2, \bar{x}_3, ..., \bar{x}_m$ and no significant difference in the second subset, $\bar{x}_1, \bar{x}_2, ..., \bar{x}_{m-1}$ then the members of the second subset are each labeled with a #1. And now we have to drill down further within $\bar{x}_2, \bar{x}_3, ..., \bar{x}_m$. We can assert that the mean of population m is significantly higher than that of population 2, and we now continue to search to check each of the two subsets of m-2 means, $\bar{x}_3, \bar{x}_2, ..., \bar{x}_m$ and $\bar{x}_2, \bar{x}_3, ..., \bar{x}_{m-1}$ to see if they are homogeneous.

This recursive process continues until we find no significant differences in any of the subsets under consideration.

As was indicated earlier, when m=2 the appropriate procedure is the t-test. If one chooses to perform a one-way anova on two populations and the two means are not significantly different, then both will be labeled #1. If, however, they are significantly different, then the larger mean will be labeled #1 and the smaller mean will be labeled #2.

For example, consider the following comparison of 7 means and the WinCross output from one of the one-way anova procedures:

	(Q)	(R)	(S)	(T)	(U)	(V)	(W)
MEAN	2.31	2.28	2.23	2.57	1.96	2.41	1.42
	#1#2	#1#2	#1#2	#1	#1#2	#1#2	#2

We note that the rank order of the means, in ascending order, is W < U < S < R < Q < V < T. So the recursive algorithm begins with a comparison of the mean of T (2.57) with the mean of W (1.42). It finds that those two means are significantly different. It then looks at the subset of means beginning with that of T and ending with that of U. It finds that the mean of T (257) is not significantly different from that of U (1.96). Then WinCross puts a #1 under the means of this subset, i.e., under the means of U, S, R, Q, V, and T.

WinCross next considers the comparison of the subset of means beginning the mean of V (2.41), the next smaller mean to that of T, and ending with the mean of W (1.42). It finds that the mean of V (2.41) is not significantly different from that of W (142). So now WinCross puts a #2 under the means of this subset, i.e., under the means of W, U, S, R, Q, and V. At this point there is no need to compare the subset of means beginning with the mean of U, as it has been found to be not significantly different from all the means smaller than it.

The anova assumes that all the populations have the same variance, and estimates this variance as

$$s^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{\sum_{i=1}^{m} n_i - m}$$

For notational convenience, we define

$$f = \sum_{i=1}^{m} n_i - m$$

This value f is sometimes referred to as the "error degrees of freedom", and s^2 is sometimes referred to as the "error variance."

Let $S_{m,f} = (\bar{x}_m - \bar{x}_1)/s$, and $S_{k,f} = (\bar{x}_j - \bar{x}_i)/s$, where k=j-i+1 denotes the number of sample means being considered in a particular subset being tested. Statistics of this form are called "Studentized ranges," and there are special tables available with percentage points of the distribution of these statistics.
In what follows we assume that we are considering the subset $\bar{x}_i, \bar{x}_{i+1}, ..., \bar{x}_j$ and comparing the difference $\bar{x}_j - \bar{x}_i$ with a critical value c_k, where k=j-i+1. Following are the appropriate values of c_k associated with the one-way anova procedures provided in WinCross.

Least-significant difference

The difference $\bar{x}_j - \bar{x}_i$ is in this procedure should be compared with

$$s\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\sqrt{2F_{\alpha,1,f}}$$

where $F_{\alpha,1,f}$ is the upper 100α percent point of the F distribution with 1 and f degrees of freedom. (In SPSS this procedure is the LSD Post Hoc Multiple Comparisons test. In MINITAB this procedure is called the Fisher procedure.)

Student Newman Keuls

The difference $\bar{x}_j - \bar{x}_i$ is in this procedure should be compared with

$$s\sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}}S_{\alpha,m,f}$$

where $S_{\alpha,m,f}$ is the upper 100α percent point of the Studentized range distribution with m and f degrees of freedom. (In SPSS this procedure is the S-N-K Post Hoc Multiple Comparisons test.)

Kramer Tukey B

The difference $\bar{x}_j - \bar{x}_i$ is in this procedure should be compared with

$$s\sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}}\,\frac{S_{\alpha,m,f} + S_{\alpha,k,f}}{2}$$

where $S_{\alpha,m,f}$ is the upper 100α percent point of the Studentized range distribution with m and f degrees of freedom and $S_{\alpha,k,f}$ is the upper 100α percent point of the Studentized range distribution with k and f degrees of freedom, and where k=i-j+1. (In SPSS this procedure is the Tukey's-b Post Hoc Multiple Comparisons test.)

Kramer Tukey

The difference $\bar{x}_j - \bar{x}_i$ is in this procedure should be compared with

$$s\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}S_{\alpha,k,f}$$

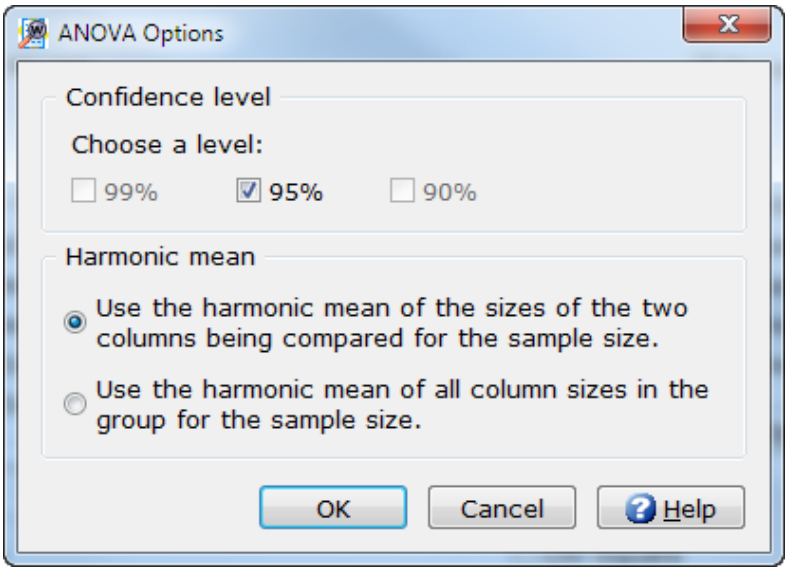
where $S_{\alpha,k,f}$ is the upper 100α percent point of the Studentized range distribution with k and f degrees of freedom, and where k=i-j+1. (In SPSS this procedure is the Tukey HSD Post Hoc Multiple Comparisons test. In MINITAB this procedure is called the Tukey procedure.)

Scheffe

The difference $\bar{x}_j - \bar{x}_i$ is in this procedure should be compared with

$$s\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\sqrt{2(m-1)F_{\alpha,m-1,f}}$$

where $F_{\alpha,m,f}$ is the upper 100α percent point of the F distribution with m and f degrees of freedom. (In SPSS this procedure is the Scheffe Post Hoc Multiple Comparisons test.)



WinCross allows a choice of one of three levels of α, namely 0.10, 0.05, and 0.01, corresponding to confidence levels of 90%, 95%, and 99% .

WinCross limits to 20 the number of columns being compared.

As one can see from the format of the various values of c_k given above, there are two types of multipliers, one being

$$\sqrt{\frac{1}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

and the other being

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}}$$

The first is the harmonic mean of the sizes of the two columns being compared and the second is the harmonic mean of all column sizes in the comparison group. We recommend the first of these as the basis for the significance test, as it is the one used in determining the tables of critical values for the significance tests. Since SPSS uses the second of these, for compatibility with SPSS WinCross provides the user with the option of using the second factor in performing the test.

CHI-SQUARE

General notation

The chi-square test computation is applied to a designated subset of a table with R contiguous rows and C contiguous columns. It tests whether there is association between the variable defining the rows and the variable defining the columns. We denote by n_{ij} the count in row i, column j of the table subset (i=1,…,R, j=1,…,C). We denote by r_i the total count in row i, by c_j the total count in row j, and by m the total count in the subset of the table. That is,

$$r_i = \sum_{j=1}^{C} n_{ij}$$

$$c_j = \sum_{i=1}^{R} n_{ij}$$

$$m = \sum_{j=1}^{C} r_i = \sum_{i=1}^{R} c_j = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}$$

Under the hypothesis of lack of association of rows and columns, the expected value of the count in cell (i,j) is given by

$$e_{ij} = \frac{r_i c_j}{m}$$

The chi-square test

The test statistic is calculated as

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

and it has a chi-square distribution with (R-1)(C-1) degrees of freedom.

FACTOR ANALYSIS

The premise of factor analysis is that there is a small set of variables, called "factors," underlying the responses to a set of questionnaire items. The factor analysis model posits, moreover, that each respondent's response to each questionnaire item can be represented by a fixed linear combination of respondent-specific values of each of the factors. The respondent-specific values of each of the factors are called the respondent's "factor scores," and the coefficients of the linear function that combines the factor scores are called the "factor loadings."

For example, suppose we were measuring mental acuity, gave each respondent a battery of 100 questions, and the factor analysis found that there were three underlying factors, "verbal ability," "numeric ability," and "memory, each of which contributes with differing weights in the respondent's determining the answers to the various questions. Each respondent would then receive three factor scores, numeric scores on each of the three underlying factors. One might, after the factor analysis is completed, sort the respondents into groups, with each group associated with the factor whose factor score is highest for that respondent. This is what WinCross does.

One cautionary note must be inserted here. A genius who has factor scores of 3.1, 3.2, and 3.3 on these three factors (i.e., is in the 99^{th} percentile on all three) should not just be pigeonholed into segment 3, the "memory" segment. Admittedly, this is his strongest suit, but not by much. Moreover, his scores on the two factors into which he is not assigned are probably higher than those of the individuals who were assigned to those segments.

Second of all, someone with extremely low scores on all three factors should, I believe, not be slotted into any segment. Again using this analogy, a respondent who has factor scores of -3.1, -3.2, and -3.3 on these three factors should not be pigeonholed into the "verbal ability" factor just because his score on that factor is the highest of his three scores.

General Notation

Let p be the number of questionnaire items and n the number of respondents. The factor analysis module begins with the p x p correlation matrix R of the questionnaire items. Let f be the number of factors underlying the responses. Let L be the p x f matrix of factor loadings. The aim of the factor analysis is to find a matrix L such that R is well approximated by the matrix product LL^T, where the T superscript denotes the transpose matrix.

Let G be an orthogonal matrix, and let L*=LG. Since $GG^T=I$, the identity matrix, $L^*L^{*T}=LGG^TL^T=LL^T$. Thus there is no unique representation of R as the product of a p x f matrix with its transpose. What factor analysts do, when given some matrix L such that R is approximately LL^T is seek an orthogonal matrix G such that the resulting matrix L*=LG is a more interpretable matrix of factor loadings. There are many mathematical techniques for finding the initial factor loading matrix L; WinCross uses the Jacobi method for finding L, and calls L the "factor matrix." There are also many mathematical

techniques for finding the G that produces the most interpretable L*; WinCross the varimax method for finding G, and calls L* the "rotated factor matrix."

Sometimes one can preset a required value of f. Most times, though, one determines the value of f by looking at the p eigenvalues of the matrix R and letting f be the number of eigenvalues that exceed 1.0. In either event the number f is referred to in WinCross as the "number of factor groups."

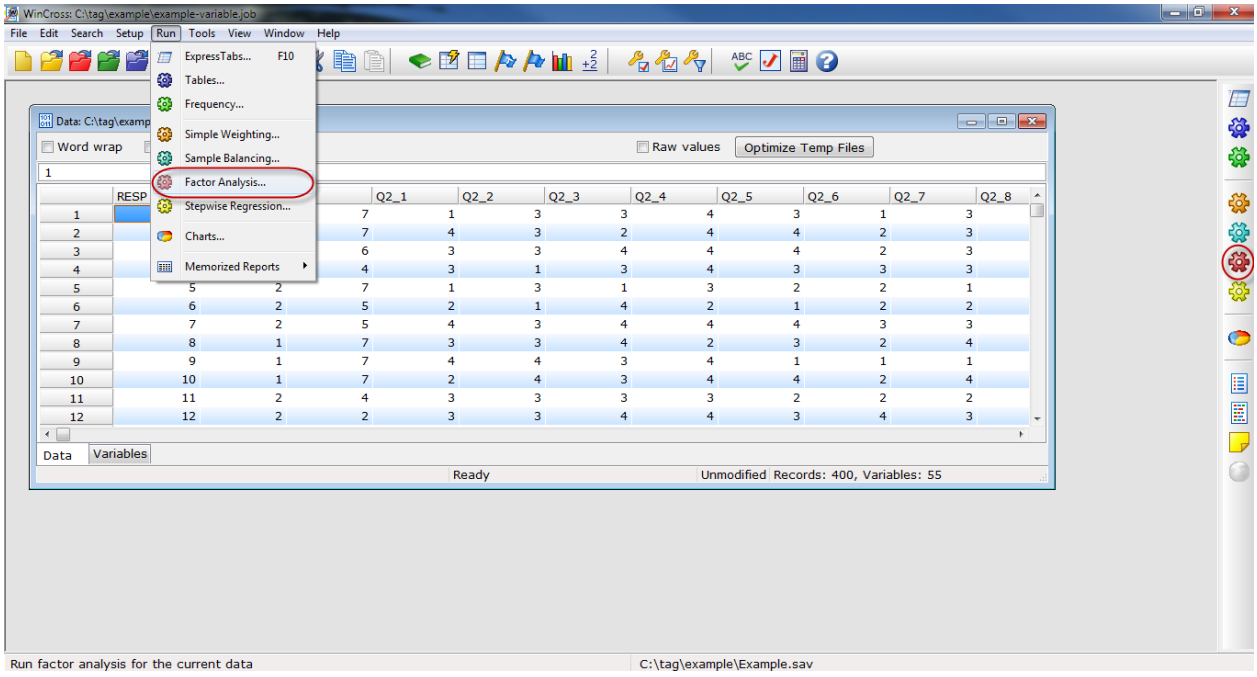
Let's now look at the responses. The i-th respondent's data can be arrayed as a p x 1 vector which we will call X_i. The i-th respondent's factor scores can be arrayed as an f x 1 vector which we will call F_i. The factor analysis model says that X_i can be approximated by the vector LF_i. Suppose we stacked all n respondents' data vectors into a p x n matrix $X = [X_1\ X_2 \cdots X_n]$ and all n respondents' factor score vectors into an f x n matrix $F = [F_1\ F_2 \cdots F_n]$. Then X = LF, and we can "solve" this equation for F as

$$F=(L^TL)^{-1}L^TX.$$

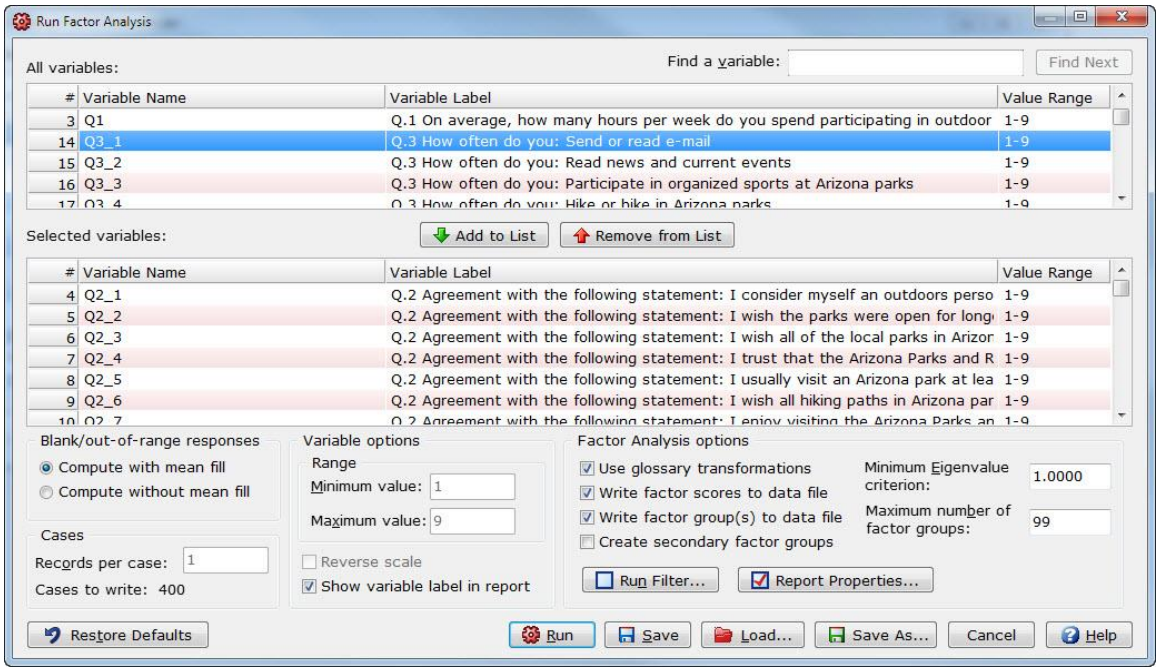
This solution is called the "regression method" for determining factor scores. WinCross applies this solution to the standardized data to produce standardized factor scores, i.e., factor scores with zero mean and unit standard deviation.

Usage

After a dataset is opened one can run the factor analysis module on the data by clicking on the **Run** command and then either on the **Factor Analysis** command, as illustrated in this screenshot, or on the red gear in the right margin of the screen.



WinCross Factor Analysis dialogs
Following is the first of the WinCross dialogs used in its Factor Analysis:



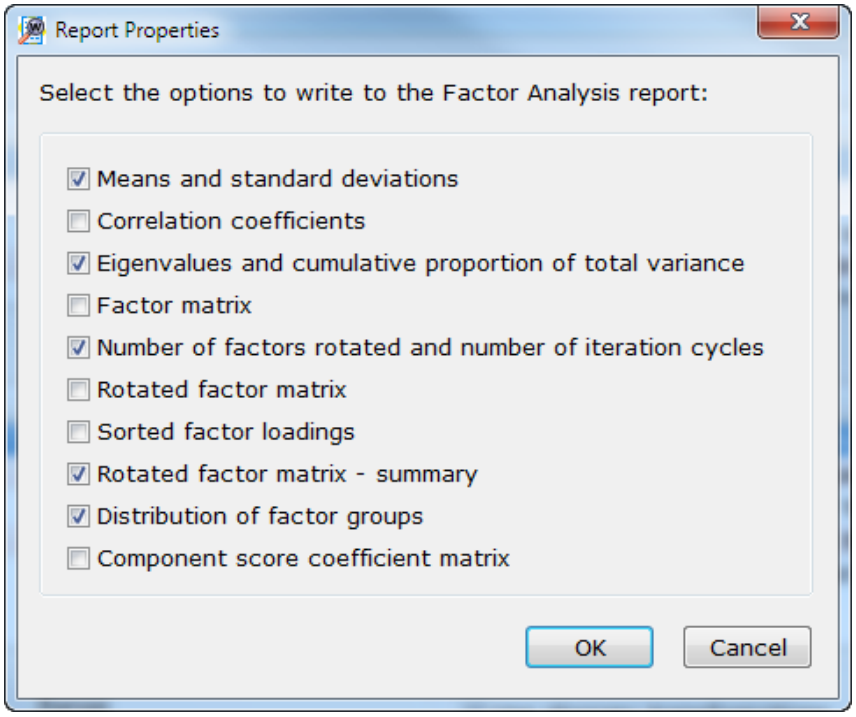
The **Blank/out-of-range responses** box gives the user two options to deal with such data. One option, **Compute with mean fill**, replaces each blank or out-of- range response with the mean of that variable. The other option, **Compute without mean fill**, determines that if any of the n variables for a respondent is blank or out of range then none of that respondent's data will be used in computing the correlation matrix (this procedure is sometimes called "listwise deletion").

The **Minimum eigenvalue criterion** enables the user to set the minimum value of the eigenvalue as the determinant of the number of factors in the factor analysis (usually set at 1.0), and the **Maximum number of factor groups** enables the user to preset the number of factors in the factor analysis. If the number of factors as determined by the **Minimum eigenvalue criterion** is smaller than the **Maximum number of factor groups**, WinCross will set the number of factors at the number determined by the **Minimum eigenvalue criterion**. If the number of factors as determined by the **Minimum eigenvalue criterion** is larger than the **Maximum number of factor groups**, WinCross will set the number of factors at the number determined by the **Maximum number of factor groups**.

The **Write factor scores to data file** option creates f factor scores, using the regression method described above, and adds them as f additional columns in the data file. The **Write factor group(s) to data file** option determines which of the f factor scores is the largest and writes the index of that factor score into a column in the data file. When writing the factor group(s) to the data file, the primary factor group is added as a new variable called GRP. The **Create secondary factor groups** option is designed to

designate respondents to a “secondary factor group” if their second-highest factor score is (a) greater than or equal to 2 and (b) is within 0.1 of their highest factor score. This grouping picks up respondents with very strong factor scores (greater than 2) that are close enough to their highest factor score (within 0.1 of it) that they should be considered as part of that factor group as well. If the user creates a secondary factor group, the group is in a new variable called GRP2. A distribution for the GPR2 is automatically generated.

Following is the second of the WinCross dialogs used in its Factor Analysis:



We describe each of these outputs in turn.

Means and standard deviations: The means and standard deviations of each of the p items are entered into the output file.

Correlation coefficients: The p x p correlation matrix R is entered into the output file.

Eigenvalues and cumulative proportion of total variance: All p eigenvalues of the correlation matrix are output, in descending order. Since the sum of the eigenvalues must equal p, the contribution of each factor to the explanation of the total variance of the data is equal to that factor’s associated eigenvalue divided by p. These ratios are accumulated and entered into the output file.

Factor matrix: This is the matrix L produced by the Jacobi method.

Number of factors rotated and number of iteration cycles: The number of factors f is output. Also, since the varimax search for L* takes multiple iterations on the computer, WinCross outputs the count of the number of iterations it took to find L*.

Rotated factor matrix: This is the matrix L* produced by the varimax procedure.

Sorted factor loadings: This is the matrix L* sorted so that (a) the coefficients of the first factor are in descending order, then (b) sorted in descending order only for those variables whose coefficients of the second factor exceed that of the first factor, then (c) sorted in descending order only for those variables whose coefficients of the third factor exceed that of the second factor, etc. This enables the user to see which variables are the most important in determining each factor.

Rotated factor matrix-summary: Same as sorted factor loadings, but with the largest coefficients highlighted in bold face type.

Distribution of factor groups: The counts of the number of respondents assigned to each of the f factors based on their factor scores is entered into the output file. This option must be checked to produce the distribution of GRP. If you are also creating secondary factor groups, then by checking this option, you will automatically produce the distribution of GRP2.

Component score coefficient matrix: The matrix $(L^{*T}L^{*})^{-1}L^{*T}$ which multiplies the standardized data to produce the factor scores.

SAMPLE BALANCING

The goal of the "Sample Balancing" module is to provide a weight for each respondent in the sample such that the weighted marginals on each of a set of characteristics matches preset values of those marginals. This process is sometimes called "raking" or "rim weighting." The most common procedure used to produce these weights is "iterative proportional fitting", a procedure devised by W. Edwards Deming and Frederick F. Stephan, first published in their December, 1940 paper, "On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known," in Volume 11 of *The Annals of Mathematical Statistics*, pages 427-444, and further explicated in Chapter 7 of Deming's book, Statistical Adjustment of Data (New York: John Wiley & Sons, 1943). Though "iterative proportional fitting" has the nice property of converging to a set of nonnegative weights, these weights do not have any optimal properties (such as the minimization of some measure of goodness of fit.)

WinCross's adaptation was developed by J. Stephens Stock, a colleague of Deming, in the 1960s with the express goal that the weights that it produces optimize a measure of goodness of fit. Unfortunately, Stock and his Market-Math, Inc. partner Jerry Green never published their algorithm, but made it available to the market research community. The Analytical Group, Inc. has utilized this algorithm since its incorporation in 1970. (In Public Opinion of Criminal Justice in California, a 1974 report for the Institute of Environmental Studies at the University of California Berkeley by the Field Research Corporation, we find a use of this algorithm, with the note (page 118) "…the weighting correction is based on a design concept originated by the late J. Stephens Stock and Market-Math, Inc. It is currently used by Field Research Organization and several other leading research organizations.")

Unfortunately, this algorithm (and any other algorithm that seeks to find weights which will optimize some criterion, such as the linear and GREG weighting procedures, see Deveille and Särndal "Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association* (1992) 87: 376-82 and Deville, Särndal, and Sautory, "Generalized Raking Procedures in Survey Sampling" *Journal of the American Statistical Association*, (1993) 88: 1013-20) may arrive at negative weights for some of the observations. This is because the data may be so inconsistent with the target marginal that the only way to reconcile the two is to create some negative weights. In page 57 of their book, Statistics for Real-Life Sample Surveys (Cambridge: Cambridge University Press, 2006), Dorofeev and Grant (2006, page 57) have presented an example of a weighting situation where the only possible set of weights which work include some negative weights. Their example is the following:

level	1	2	3	target
1	5	7	10	25
2	3	0	0	15
3	9	10	1	5
target	10	15	20	

In order to weight cell (2,2) so that the second row sum is 15, the weight must be 5. But 5 times 3 exceeds 10, so that the (1,1) and (3,1) cells must have negative weights in order for the first column sum to be 10.

Of course this is an unnatural example, in that there are 0s in columns 2 and 3 of row 2. But it illustrates the problem, which can occur even in perturbations of this example where the 0s are replaced by small nonzero frequencies. (Mathematically, for this example we must solve 6 linear equations—corresponding to the six targets—for 9 unknowns—the nine cell weights, with the constraint that the 9 unknowns must all be positive. There are lots of solutions to this mathematical problem, of which iterative proportional fitting may converge on one -- but the moment a "goodness-of-fit" criterion is superimposed on this problem, the imbalance of the data with the targets shows up in the form of negative weights.)

General Notation

Let v be the number of variables to be considered in the balancing. Let c_i denote the number of levels (sometimes referred to as "breaks") of the i-th variable, i=1,…,v. Let $p_{j_1 \ldots j_v}$ denote the proportion of respondents in the sample in level j_1 on variable 1, j_2 on variable 2, …, j_v on variable v, where $j_i = 1, \ldots, c_i$. Let $f^i_{j_i}$ denote the marginal proportion in the sample of level j_i of variable i (j_i=1,…,c_i, i=1,…,v).

To make things concrete, let v=3, with the three variables being income (i=1), age (i=2), and region (i=3). Suppose there are 5 income breaks (c_1=5), 10 age breaks (c_2=10), and 9 region breaks (c_3=9). Then, in our notation, if for example j_1=2, j_2=1, and j_3=4, then $p_{j_1 j_2 j_3} = p_{214}$ is the proportion of the sample that are of income level 2, age level 1, and region level 4. And, as another example of the interpretation of this notation, if i=3 then $f^i_{j_i} = f^3_{j_3} = f^3_2$ is the proportion of the sample that are in region level 2 (the superscript "3" indicates that we are looking at variable 3, region, and the subscript "2" indicates that we are looking at level 2 of that variable).

$f^i_{j_i}$ can be determined by adding up all the $p_{j_1 \ldots j_v}$ across all the values of each of the v-1 j_k for which $k \neq i$. For example, to obtain f^3_2 one adds up all the proportions $p_{j_1 j_2 2}$ across j_1=1,2 and j_2=1,2,3. We express this relation symbolically as

$$f^i_{j_i} = \sum_{j_k, k \neq i} p_{j_1 \ldots j_v}$$

These $f^i_{j_i}$ are called sample rim percents.

Suppose that the preset distributions on the v variables are given by the set of target proportions $g^i_{j_i}$. The object of the sample balancing module is to find a set of weights $w_{j_1 \ldots j_v}$ such that if, when looking at the j_i-th break, instead of adding up the $p_{j_1 \ldots j_v}$ across

all but the i-th category, we add up the $w_{j_1 \ldots j_v} p_{j_1 \ldots j_v}$ across all but the i-th category, we will obtain the $g^i_{j_i}$. That is,

$$g^i_{j_i} = \sum_{j_k, k \neq i} w_{j_1 \ldots j_v} p_{j_1 \ldots j_v}$$

These $g^i_{j_i}$ are called target rim percents. If this were a simple one-dimensional sample balancing situation (i.e., v=1), then the ratios of the target rim percents to the sample rim percents would be the appropriate weights for the various levels.

Goodness-of-fit minimization technique

The procedure for determining the weights is iterative. Each iterative "round" consists of v "passes," one "pass" through each of the v variables. We begin at "round 0" by setting all weights $w_{j_1 \ldots j_v}(0,i)$ equal to 1, i.e., we begin with the unweighted data.

Suppose we are on the i-th "pass" in "round t+1." Let $w_{j_1 \ldots j_v}(t,i)$ denote the weights at this point in the iterative process. Let $g^i_{j_i}(t)$ denote the results of the computation

$$g^i_{j_i}(t) = \sum_{j_k, k \neq i} w_{j_1 \ldots j_v}(t,i) p_{j_1 \ldots j_v}$$

These $g^i_{j_i}(t)$ are called estimated target rim percents.

At the first pass (i=1) of the t-th round of the iterative procedure the module calculates a set of increments $d_{j_1 \ldots j_v}(t,1)$ to add to the $w_{j_1 \ldots j_v}(t-1,v)$, producing $w_{j_1 \ldots j_v}(t,1) = w_{j_1 \ldots j_v}(t-1,v) + d_{j_1 \ldots j_v}(t,1)$. At the i-th pass (i>1) of the t-th round of the iterative procedure the module calculates a set of increments $d_{j_1 \ldots j_v}(t,i)$ to add to the $w_{j_1 \ldots j_v}(t,i-1)$, producing $w_{j_1 \ldots j_v}(t,i) = w_{j_1 \ldots j_v}(t,i-1) + d_{j_1 \ldots j_v}(t,i)$.

These increments are given by the formula

$$d_{j_1 j_2 \ldots j_v} = [g^i_{j_i}(t) - g^i_{j_i}] / f^i_{j_i}$$

That is, we compare the ratio of the estimated target rim percent to the sample rim percent to the ratio of the target rim percent to the sample rim percent, and increment or decrement by the difference between these two ratios.

The WinCross goodness-of-fit minimization sample balancing module now applies these new weights to the respondents and begins round 2, once again in pass 1 looking at the income marginals. The principle in each step is the same: adjust the weights so that the ratio of estimated target rim percents data to sample rim percents equals the ratio of target rim percents to sample rim percents.

The module continues iterating until a criterion of goodness of fit has been met. WinCross uses the measure

$$\sqrt{\sum_{i=1}^{v}\sum_{j_i=1}^{c_i}[(g^i_{j_i}(t)-g^i_{j_i})/\ f^i_{j_i}]^2\ /\ m}$$

where

$$m=\sum_{i=1}^{v}c_i$$

is the total number of levels in the balancing process.

The square of this measure is the average across levels and variables of the sum of squares of deviations between the ratio of estimated target rim percents to sample rim percents and the ratio of actual target rim percents to sample rim percents. The module iterates until this measure is less than some preset value (with default set at 0.00005).

Iterative proportional fitting technique

The procedure for determining the weights is iterative. Each iterative "round" consists of v "passes," one "pass" through each of the v variables. We begin at "round 0" by setting the estimated target rim percents $p^{(0)}{}_{j_1\cdots j_v}$ as equal to the sample rim percents $p_{j_1\cdots j_v}$.

Suppose we are in "round t+1." At the i-th pass of the t+1-st round of the iterative procedure the module calculates

$$p^{(t+1)}{}_{j_1\cdots j_v}=p^{(t)}{}_{j_1\cdots j_v}\frac{\sum_{j_k,k\neq i}p_{j_1\cdots j_v}}{\sum_{j_k,k\neq i}p^{(t)}{}_{j_1\cdots j_v}}$$

That is, we calculate the ratio of the sample rim percent to the estimated target rim percent at round t. and multiply the estimated target rim percent at round t by this ratio.

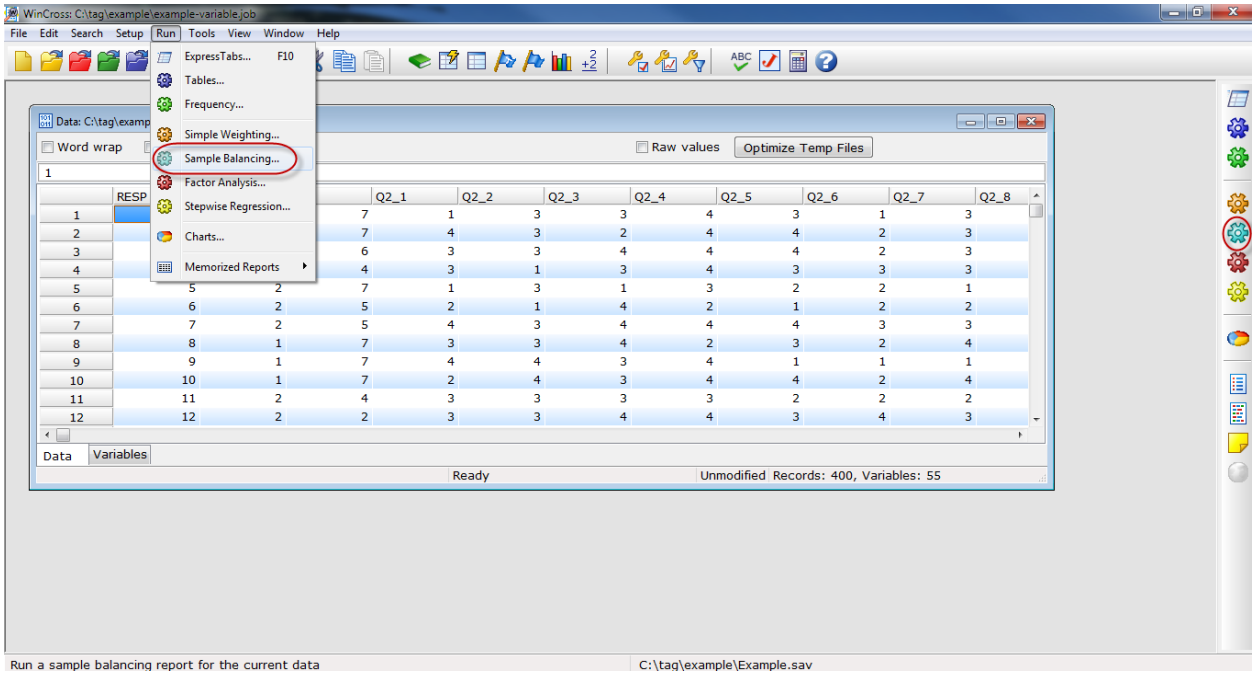
Iterative proportional fitting has no overall criterion for goodness of the adjusted weight. It merely iterates until each $p^{(t+1)}{}_{j_1\cdots j_v}$ is within some preset distance from $p^{(t)}{}_{j_1\cdots j_v}$, that is, until each estimated target rim percent is within some preset distance from the its predecessor estimated target rim percent. It has been proven that, except for extreme data situations in which some of the cells are 0, the iterative procedures terminates. And, as mentioned earlier, the results are nonnegative weights. But there are no known optimal properties of this procedure.

You can find an example of the use of the WinCross goodness-of-fit minimization module and a contrast with that of iterative proportional fitting on our web site:

For the goodness-of-fit minimization module the value of the goodness of fit is 0.000431586; for iterative proportional fitting this value is 0.006785959, over 15 times as large.

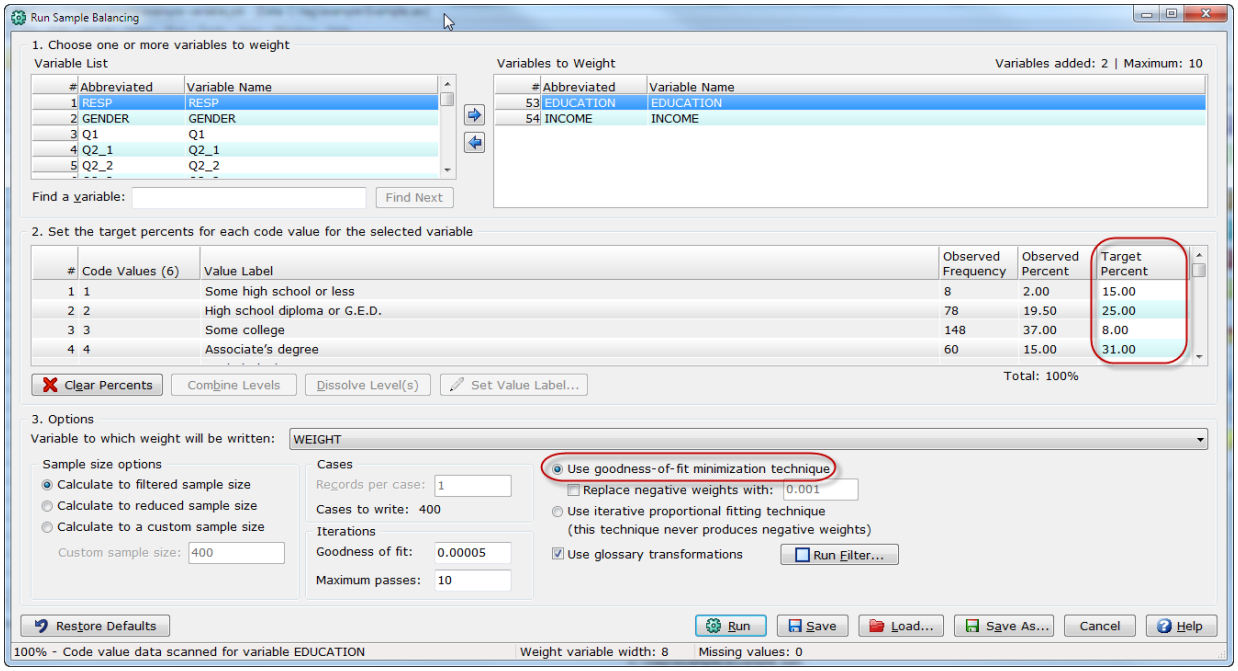
Usage

After a dataset is opened one can run the sample balancing module on the data by clicking on the **Run** command and then either on the **Sample Balancing** command, as illustrated in this screenshot, or on the turqoise gear in the right margin of the screen.



WinCross Sample Balancing dialog

Following is the WinCross dialog used in its Sample Balancing:



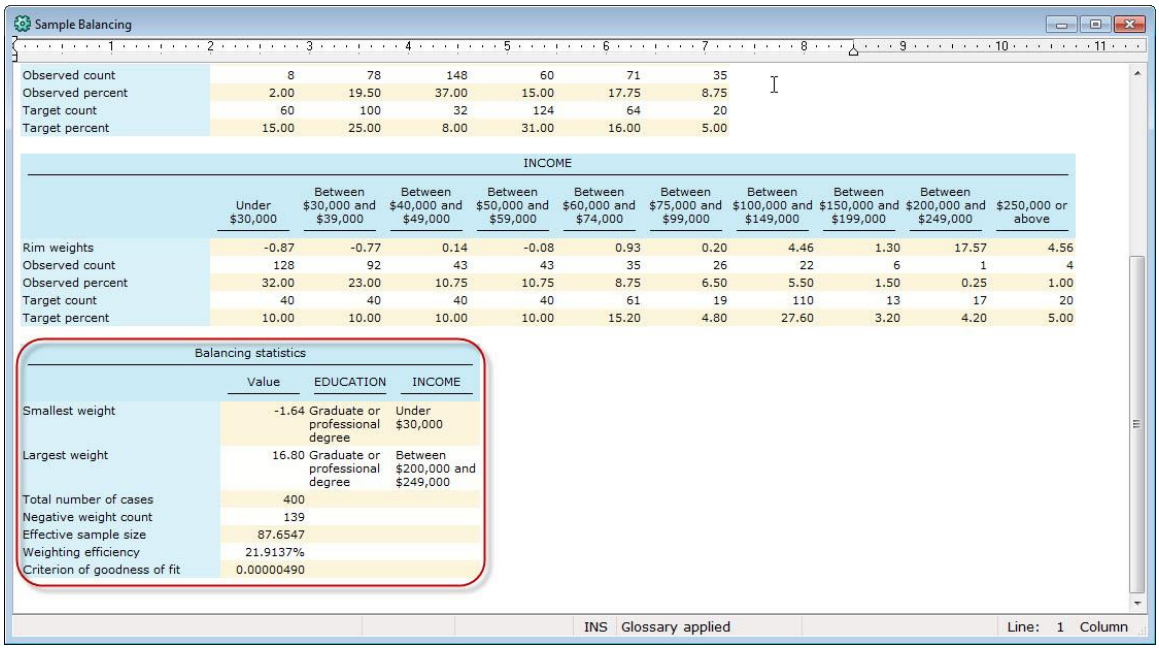
Note, that for each variable in the **Code Values** list, you are to enter the associated target rim percent in the **Target Percent** column. Note also, that you are able to enter a default replacement value for any negative weights generated by the sample balancing module. Also, if any respondent receives a weight of 0 then you have the option of deleting him from the sample and recalculating the weight with that respondent not included in the sample rim percents.

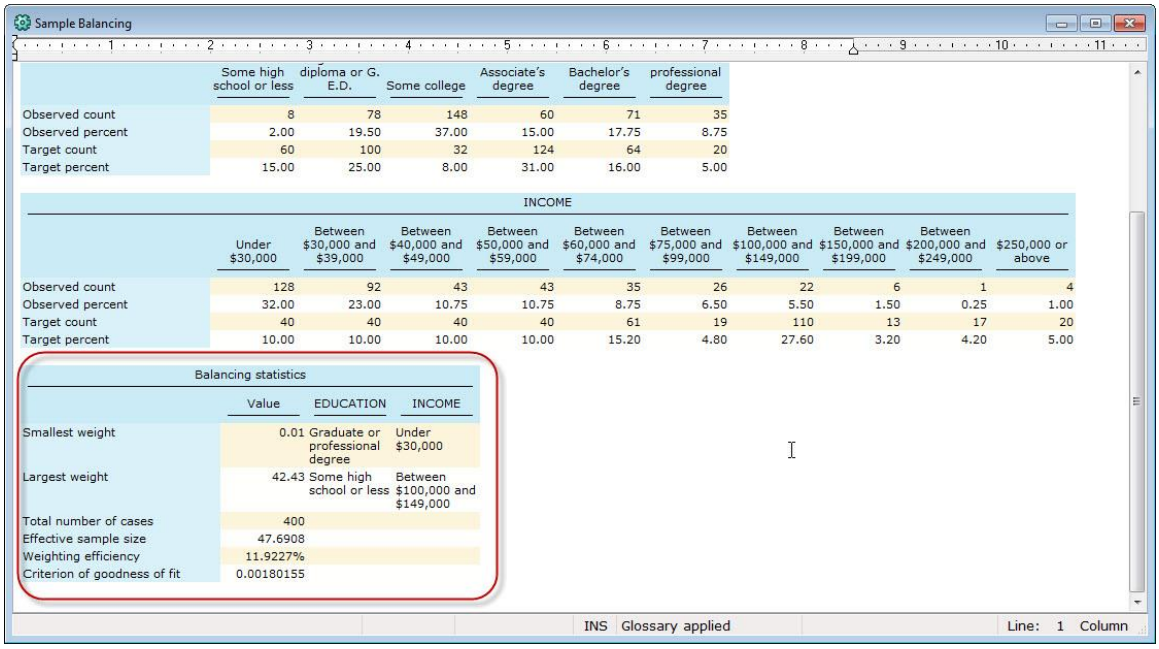
The default procedure is the goodness-of-fit-minimization technique described above. One situation that might occur in using this technique is that it will produce "negative" weights. One way to avoid this is to replace those weights with a small number, such as 0.001. Checking the **Use goodness-of-fit minimization technique** box invokes this weighting procedure and checking the **Replace negative weights with:** box enables the user to force the procedure to produce positive weights, with those items that were to receive negative weights having their weights replaced by a small positive number of your choice. Checking the **Use iterative proportional fitting technique** box invokes that procedure, and produces non-negative weights, but without any optimality property.

Though not used in iterative proportional fitting, the goodness-of-fit metric used in the goodness-of-fit minimization technique is calculated for the results of the iterative proportional fitting technique as well. Should that metric be smaller than 0.00005 the user should be advised that, if one wanted to improve on the iterative proportional fitting set of weights, one should reset the default value of the goodness-of-fit minimization technique cutoff value to something smaller than that of the iterative proportional fitting goodness-of-fit metric.

Also calculated for both procedures are the effective sample size based on the final weights and a measure of statistical efficiency, namely the ratio of the effective sample size to the sample size of the input data.

Here are examples of the output for each of the procedures:





Note in the lower left that the output includes three statistics: **Effective sample size, Weighting efficiency,** and **Criterion of goodness of fit.** The **Criterion of goodness of fit** is that given on page 98 above. The **Effective sample size** is the computation

described on page 6 above, sometimes called the “design effect,” which is the denominator of the variance of the mean to be used in statistical estimation and tests. The **Weighting efficiency** is the ratio of the effective sample size to the true sample size, multiplied by 100. So in the above examples the effective sample size based on the goodness-of-fit weights (87.65) is 21.9% of that of the full sample of 400, and the effective sample size based on iterative proportional fitting (47.69) is 11.9% of that of the full sample of 400.

REGRESSION

Background

The context of multiple regression is that there is a variable y (called the "dependent variable") and a set of m other variables, x_1, x_2, …, x_m (called the "independent variables") and one postulates that there is a linear relationship between the dependent and the independent variables, of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m$$

The goal of multiple regression analysis is to find best estimates a, b_1, b_2, …, b_m of α, β_1, β_2, …, β_m based on n observations of the set of variables (y, x_1, x_2, …,x_m).

There are two metrics for assessing the linear relationship, called R^2 and adjusted-R^2. R^2 is a measure of the fraction of the variation of the y's and is accounted for by the regression on the independent variables, i.e.,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - ... - b_m x_{mi})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

One feature of R^2 is that as more and more independent variables are added to the regression, R^2 is ever increasing. The other metric, adjusted R^2, is a measure of the fraction of the variance of the y's and is accounted for by the regression on the independent variables, i.e.,

$$adjusted - R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - ... - b_m x_{mi})^2 / (n - m - 1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n - 1)}$$

Adjusted-R^2 takes into account the number of independent variables used in the regression, and so the addition of one more independent variable may make the adjusted-R^2 smaller than its predecessor (and may even become negative!).

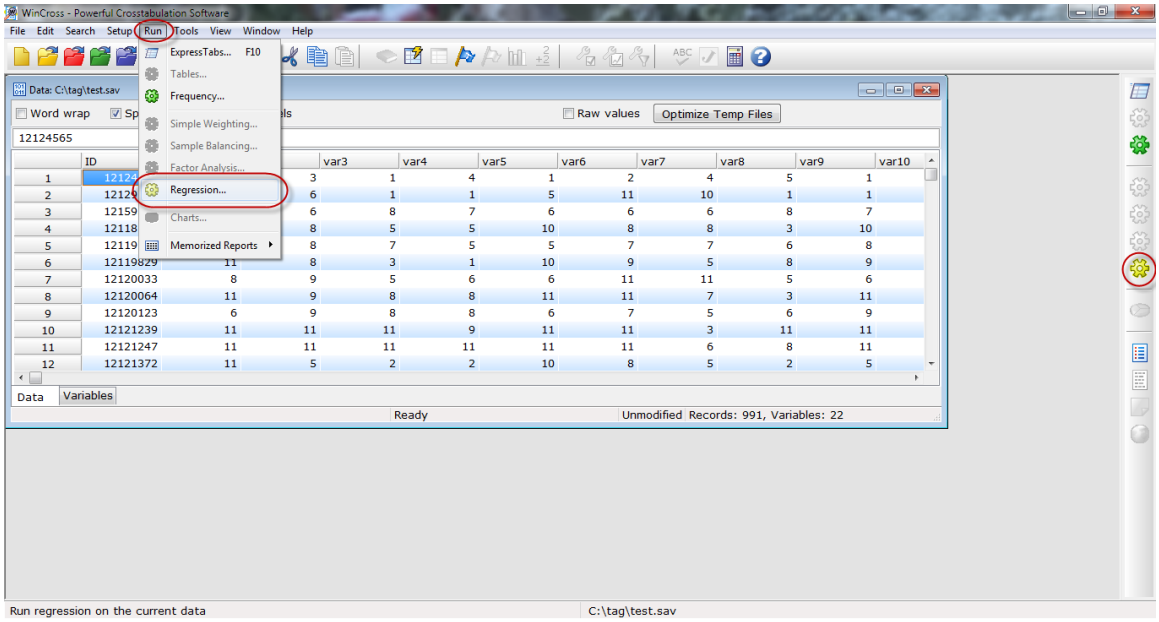
Regression starts by selecting the independent variable that is most correlated with y as its initial indpendent dependent variable. It then uses the adjusted-R^2 as its metric and, at each step of the process, selects the independent variable that, when added to those already selected, produces the largest adjusted-R^2. It halts when an independent variable is selected whose coefficent is not significantly different from 0 using the appropriate t statistic.

Sometimes these regressions are called "driver analysis," in that all the independent variables are positively correlated with the dependent variables and the analyst wants to know which of the independent variables "drives" the dependent variable. Multiple regression may produce negative coefficients for some of these variables, even though they are positively correlated with the dependent variable. This happens because the use of some of the independent variables will produce an overestimate of the dependent variable, which can be reduced by including an additional independent variable with a

negative coefficient in the regression. WinCross has added the facility to allow the regression to terminate when an independent variable is introduced and its coefficient is negative.

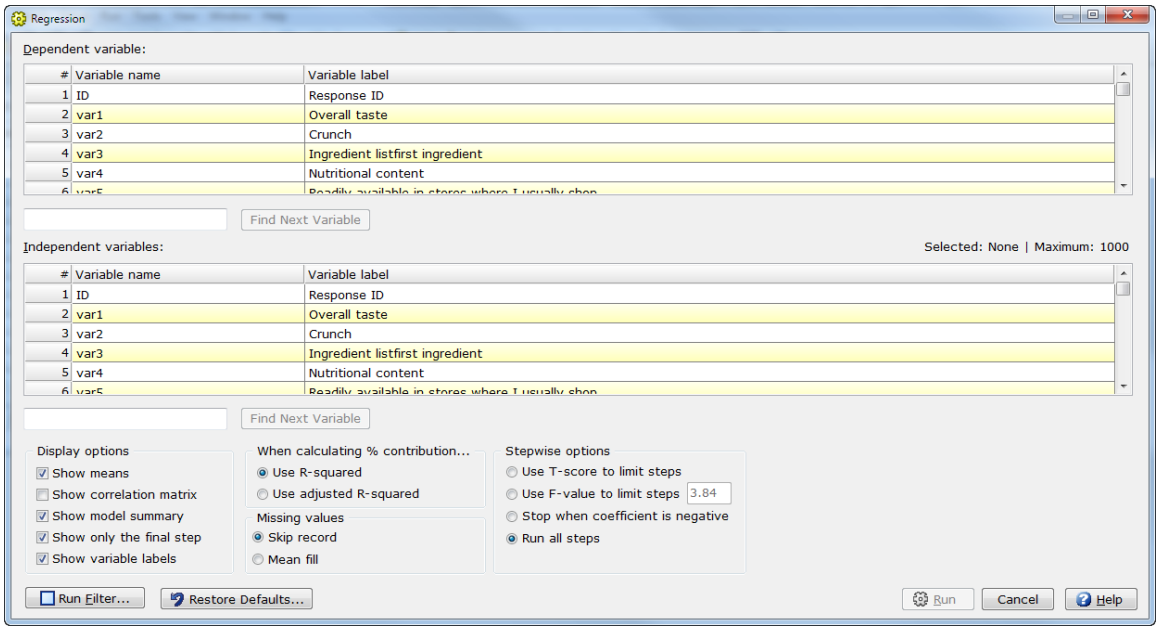
Usage

After a dataset is opened, one can run a regression on the data by clicking on the **Run** command and then either on the **Regression** command, as illustrated in this screenshot, or on the yellow gear in the right margin of the screen.



The variables in the data set are arrayed and one can select the dependent variable and the set of independent variables to be used in the regression using the following dialog.

To make your selections highlight your choice of dependent variable and your choices of independent variables.



The **Stepwise options** provide the user with four different stopping rules. The **Use T-score to limit steps** looks at the t-statistic at each step and terminates when the t-statistic shows that the coefficient is not significantly different from 0. The **Use F-value to limit steps** calculates the square of the current step's value of t and terminates when this is below a value determined by the user and entered into the box to the right of that option. The default is 3.84, which is the 95% point of the F distribution with 1 and ∞ as degrees of freedom. (The use of an F value as the stopping rule was recommended by the inventor of regression, and implemented in many regression programs, sometimes with values of F other than 3.84. This rule was studied by Wilkinson and Dallal in their 1981 paper, "Tests of significance in forward selection regression with an F-to-enter stopping rule," *Technometrics*, 23: 377–380. They showed that a final regression obtained by this selection rule based on the F value at 1%, was in fact only significant at 5%. We therefore do not recommend using this criterion, and have made the **Use T-score to limit steps** as the WinCross default.) The **Stop when coefficient is negative** option is used when the regression is a driver analysis and the user does not want to include negative coefficients in the model. The **Run all steps** option is used when one wants to use WinCross to perform a complete multiple regression using all of the independent variables.

An example of the use of this module using a data set with 21 independent variables, where the correlation of the dependent variable with each of the independent variables is given in the following table.

variable 1	0.4266
variable 2	0.3596
variable 3	0.4462
variable 4	0.3003
variable 5	0.1796
variable 6	0.3432
variable 7	0.5009
variable 8	0.4136
variable 9	0.4328
variable 10	0.3647
variable 11	0.2472
variable 12	0.2629
variable 13	0.3283
variable 14	0.3767
variable 15	0.3907
variable 16	0.4148
variable 17	0.4533
variable 18	0.4487
variable 19	0.4795
variable 20	0.2885
variable 21	0.3169

Note that all the correlations are positive. This, then, is an example of a driver analysis.

The results of running the regression using all the independent variables are given next. There are a few things to be noticed.

1. The coefficient of variable 13, introduced in step 3, is negative. Therefore if one is performing a driver analysis and has checked the **Stop when coefficient is negative** option, the regression would stop at step 2.
2. Note that when variable 12 gets introduced (at step 16) the adjusted-R^2 is lower than that of step 15.
3. Note also that R^2 continually increases until it reaches the final level of 0.4028.

step	variable	coefficient	std error	t value	p	R^2	adj R^2
1	variable 4	0.16640	0.02936	5.66752	0.0000	0.2360	0.2352
2	variable 5	0.20295	0.02764	7.34319	0.0000	0.3045	0.3031
3	variable 13	-0.08400	0.02806	-2.99395	0.0028	0.3424	0.3404
4	variable 18	0.10308	0.02563	4.02156	0.0001	0.3644	0.3618
5	variable 10	0.07316	0.02439	2.99985	0.0028	0.3749	0.3717
6	variable 15	0.07778	0.02256	3.44743	0.0006	0.3814	0.3776
7	variable 11	-0.04722	0.02472	-1.90983	0.0564	0.3850	0.3807
8	variable 2	0.07555	0.02553	2.95870	0.0032	0.3901	0.3852
9	variable 21	-0.02902	0.02302	-1.26028	0.2079	0.3930	0.3874
10	variable 1	0.03370	0.02286	1.47439	0.1407	0.3953	0.3891
11	variable 9	-0.03720	0.02232	-1.66658	0.0959	0.3971	0.3903
12	variable 6	-0.02942	0.01974	-1.49047	0.1364	0.3986	0.3912
13	variable 16	-0.02888	0.02493	-1.15848	0.2470	0.3998	0.3918
14	variable 20	0.02687	0.02470	1.08754	0.2771	0.4009	0.3923
15	variable 19	0.02591	0.02132	1.21560	0.2244	0.4017	0.3925
16	variable 12	-0.02418	0.02909	-0.83125	0.4060	0.4022	0.3924
17	variable 7	-0.01253	0.02470	-0.50740	0.6120	0.4024	0.3920
18	variable 8	0.01442	0.02676	0.53869	0.5902	0.4026	0.3915
19	variable 17	-0.01035	0.02236	-0.46300	0.6435	0.4027	0.3911
20	variable 3	-0.00921	0.02566	-0.35909	0.7196	0.4028	0.3905
21	variable 14	-0.00183	0.02541	-0.07218	0.9425	0.4028	0.3899
22	variable 22	0.00129	0.02103	0.06136	0.9511	0.4028	0.3893
	Constant:	4.78356					

Following is the result of the regression using the option **Use T-score to limit steps** or **Use F-value to limit steps**. The coefficient of the variable selected at step 10 (variable 1) has an associated t value that is below that at the 5% level of significance.

step	variable	coefficient	std error	t value	p	R^2	adj R^2
1	variable 4	0.18116	0.02730	6.63500	0.0000	0.2360	0.2352
2	variable 5	0.19642	0.02515	7.81000	0.0000	0.3045	0.3031
3	variable 13	-0.10645	0.02253	-4.72400	0.0000	0.3424	0.3404
4	variable 18	0.10591	0.02298	4.60900	0.0000	0.3644	0.3618
5	variable 10	0.07747	0.02291	3.38100	0.0010	0.3749	0.3717
6	variable 15	0.07607	0.02181	3.48800	0.0010	0.3814	0.3776
7	variable 11	-0.05880	0.02286	-2.57200	0.0100	0.3850	0.3807
8	variable 2	0.06345	0.01929	3.28900	0.0010	0.3901	0.3852
9	variable 21	-0.04405	0.02056	-2.14200	0.0320	0.3930	0.3874
	Constant:	4.76612					

Finally, here is the result of using the **Stop when coefficient is negative** option.

step	variable	coefficient	std error	t value	p	R^2	adj R^2
1	variable 4	0.27998	0.02520	11.11200	0.000	0.2360	0.2352
2	variable 5	0.25455	0.02581	9.86400	0.000	0.3045	0.3031
	Constant:	5.01366					

The **When calculating % contribution** provides the user with two choices, **Use R-squared** and **Use adjusted R-squared**. Following are the results of the three regressions reported above when using R^2 as the basis for calculating each variables contribution. The base in each case is the R^2 associated with the last step of the regression. In the case of the first variable in the regression, the percent contribution is the ratio of its R^2 to that of the last step in the regression. In the case of each of the other variables, the percent contribution is the ratio of the change in R^2 from that of the previous step to the R^2 from the last step in the regression. Though the R^2 in each step is always at least as large as that of the previous step, the changes in R^2 from step to step are not monotonically decreasing (note, for example, the % contribution at steps 6, 7, 8, and 9 of this regression). But, the percent contributions at each step are all non-negative.

	all steps		t test criterion		positive coeff crit	
step	R^2	% contrib	R^2	% contrib	R^2	% contrib
1	0.2360	58.6%	0.2360	60.0%	0.2360	77.5%
2	0.3045	17.0%	0.3045	17.4%	0.3045	22.5%
3	0.3424	9.4%	0.3424	9.6%		
4	0.3644	5.5%	0.3644	5.6%		
5	0.3749	2.6%	0.3749	2.7%		
6	0.3814	1.6%	0.3814	1.6%		
7	0.3850	0.9%	0.3850	0.9%		
8	0.3901	1.3%	0.3901	1.3%		
9	0.3930	0.7%	0.3930	0.7%		
10	0.3953	0.6%				
11	0.3971	0.4%				
12	0.3986	0.4%				
13	0.3998	0.3%				
14	0.4009	0.3%				
15	0.4017	0.2%				
16	0.4022	0.1%				
17	0.4024	0.1%				
18	0.4026	0.0%				
19	0.4027	0.0%				
20	0.4028	0.0%				
21	0.4028	0.0%				
22	0.4028	0.0%				

Following are the results of the three regressions reported above when using adjusted-R^2 as the basis for calculating each variables contribution. The base in each case is the adjusted-R^2 associated with the last step of the regression. In the case of the first variable in the regression, the percent contribution is the ratio of its adjusted-R^2 to that of the last step in the regression. In the case of each of the other variables, the percent contribution is the ratio of the change in adjusted-R^2 from that of the previous step to the adjusted-R^2 from the last step in the regression. Note that the adjusted-R^2 in each step is not always at least as large as that of the previous step (note, for example, steps 16 through 22). Also, the changes in adjusted-R^2 from step to step are not monotonically decreasing (note, for example, the % contribution at steps 6, 7, 8, and 9 of this regression). Finally, note that some of the percentage contributions using this metric are negative (see steps 17 through 22). As the percentage contributions are negative, we recommend the use of adjusted-R^2 as the criterion to look at in assessing the steps in a regression. For reporting purposes one may want to use R^2 as the basis for reporting percent contribution.

	all steps		t test criterion		positive coeff crit	
step	adj-R^2	% contrib	adj-R^2	% contrib	adj-R^2	% contrib
1	0.2352	60.4%	0.2352	60.7%	0.2352	77.6%
2	0.3031	17.4%	0.3031	17.5%	0.3031	22.4%
3	0.3404	9.6%	0.3404	9.6%		
4	0.3618	5.5%	0.3618	5.5%		
5	0.3717	2.5%	0.3717	2.6%		
6	0.3776	1.5%	0.3776	1.5%		
7	0.3807	0.8%	0.3807	0.8%		
8	0.3852	1.2%	0.3852	1.2%		
9	0.3874	0.6%	0.3874	0.6%		
10	0.3891	0.4%				
11	0.3903	0.3%				
12	0.3912	0.2%				
13	0.3918	0.1%				
14	0.3923	0.1%				
15	0.3925	0.0%				
16	0.3924	0.0%				
17	0.3920	-0.1%				
18	0.3915	-0.1%				
19	0.3911	-0.1%				
20	0.3905	-0.1%				
21	0.3899	-0.2%				
22	0.3893	-0.2%				

APPENDIX I

Our web site, www.AnalyticalGroup.com, contains four papers of varying technical levels:

1. "Weighted Standard Error and its Impact on Significance Testing (WinCross vs. Quantum & SPSS)"
This provides a basic derivation of the significance test used by WinCross along with a comparison with the computations provided by other software systems.

2. "A Simulation Comparison of WinCross, SPSS, and Mentor Procedures for Estimating the Variance of a Weighted Mean"
This shows by a simulation example that WinCross's procedure is the most precise.

3. "An Analysis of WinCross, SPSS, and Mentor Procedures for Estimating the Variance of a Weighted Mean"
This presents the mathematical proof that WinCross's procedure is the most precise.

4. "Alternative Approaches to Significance Testing with Weighted Means"
This presents a nonmathematical summary of these other papers.

Critical Value for t-Distribution Table

	Confidence Level						
	99%	98%	95%	90%	80%	70%	60%
Degrees of Freedom							
1	63.660	31.820	12.710	6.314	3.078	1.963	1.376
2	9.925	6.965	4.303	2.920	1.886	1.386	1.061
3	5.841	4.541	3.182	2.353	1.638	1.250	0.978
4	4.604	3.747	2.776	2.132	1.533	1.190	0.941
5	4.032	3.365	2.571	2.015	1.476	1.156	0.920
6	3.707	3.143	2.447	1.943	1.440	1.134	0.906
7	3.499	2.998	2.365	1.895	1.415	1.119	0.896
8	3.355	2.896	2.306	1.860	1.397	1.108	0.889
9	3.250	2.821	2.262	1.833	1.383	1.100	0.883
10	3.169	2.764	2.228	1.812	1.372	1.093	0.879
11	3.106	2.718	2.201	1.796	1.363	1.088	0.876
12	3.055	2.681	2.179	1.782	1.356	1.083	0.873
13	3.012	2.650	2.160	1.771	1.350	1.079	0.870
14	2.977	2.624	2.145	1.761	1.345	1.076	0.868
15	2.947	2.602	2.131	1.753	1.341	1.074	0.866
16	2.921	2.583	2.210	1.746	1.337	1.071	0.865
17	2.898	2.567	2.110	1.740	1.333	1.069	0.863
18	2.878	2.552	2.101	1.734	1.330	1.067	0.862
19	2.861	2.539	2.093	1.729	1.328	1.066	0.861
20	2.845	2.528	2.086	1.725	1.325	1.064	0.860
21	2.831	2.518	2.080	1.721	1.323	1.063	0.859
22	2.819	2.508	2.074	1.717	1.321	1.061	0.858
23	2.807	2.500	2.069	1.714	1.319	1.060	0.858
24	2.797	2.492	2.064	1.711	1.318	1.059	0.857
25	2.787	2.485	2.060	1.708	1.316	1.058	0.856
26	2.779	2.479	2.056	1.706	1.315	1.058	0.856
27	2.771	2.473	2.052	1.703	1.314	1.057	0.855
28	2.763	2.467	2.048	1.701	1.313	1.056	0.855
29	2.756	2.462	2.045	1.699	1.311	1.055	0.854
30	2.750	2.457	2.042	1.697	1.310	1.055	0.854
>30	2.704	2.423	2.021	1.684	1.303	1.050	0.851
>40	2.660	2.390	2.000	1.671	1.296	1.045	0.848
>60	2.617	2.358	1.980	1.658	1.289	1.041	0.845
>120	2.576	2.326	1.960	1.645	1.282	1.036	0.842