

# STATISTICAL REFERENCE

## TABLE OF CONTENTS

<b>STATISTICS WITH UNWEIGHTED DATA.....</b>	<b>1</b>
Mean.....	1
Variance .....	1
Technical Comment: Variance of a Sample Proportion .....	2
Standard Deviation .....	2
Standard Error.....	3
Grouped Median .....	3
Skewness and Kurtosis .....	3
<b>STATISTICS WITH WEIGHTED DATA.....</b>	<b>5</b>
Weighted Mean .....	5
Weighted Variance.....	5
Weighted Standard Deviation.....	6
Weighted Sample Size.....	6
Effective Sample Size .....	6
Weighted Standard Error .....	7
<b>SIGNIFICANCE TESTING .....</b>	<b>8</b>
<b>T-TESTS – INDEPENDENT .....</b>	<b>13</b>
<b>UNWEIGHTED DATA.....</b>	<b>13</b>
Assume equal variances.....	13
Assume unequal variances .....	14
Technical Comment: A Note on Degrees of Freedom .....	14
Letting WinCross determine whether variances are equal or not .....	15
Part-Whole Comparisons .....	15
<b>SINGLY and MULTIPLY WEIGHTED DATA .....</b>	<b>16</b>
Assume equal variances.....	17
Assume unequal variances .....	17
Part-Whole Comparisons .....	18
<b>T-TESTS – DEPENDENT PAIRED/OVERLAP ( LOC+/VAR+).....</b>	<b>19</b>
<b>UNWEIGHTED DATA.....</b>	<b>20</b>
t-Test for Means with Partial Pairing .....	20
Technical Comment: On Calculating Covariances .....	21
Technical Comment: A Note on Perfect Pairing.....	22
Part-Whole Comparisons .....	22
<b>SINGLY WEIGHTED DATA .....</b>	<b>24</b>
t-Test for Means with Partial Pairing .....	25
Technical Comment: A Note on Perfect Pairing.....	26
Part-Whole Comparisons .....	27
<b>MULTIPLY WEIGHTED DATA .....</b>	<b>29</b>
t-Test for Means with Partial Pairing .....	30
<b>T-TESTS – DEPENDENT PAIRED/OVERLAP (MULTI) .....</b>	<b>32</b>
<b>UNWEIGHTED DATA .....</b>	<b>32</b>
Part-Whole Comparisons.....	34

SINGLY WEIGHTED DATA .....	36
Part-Whole Comparisons.....	38
MULTIPLY WEIGHTED DATA .....	39
Z-TESTS - INDEPENDENT .....	41
UNWEIGHTED DATA.....	41
Using unpooled proportions.....	41
Using pooled proportions .....	42
Technical Comment: Testing for Equality of Two Multinomial Proportions .....	42
Part-Whole Comparisons.....	43
SINGLY and MULTIPLY WEIGHTED DATA .....	44
Using unpooled proportions.....	45
Using pooled proportions .....	45
Part-Whole Comparisons .....	45
Z-TESTS – DEPENDENT PAIRED/OVERLAP ( LOC+/VAR+).....	47
UNWEIGHTED DATA.....	47
z-Test for Proportions with Partial Pairing.....	47
Technical Comment: A Note on Perfect Pairing.....	48
Part-Whole Comparisons.....	49
SINGLY WEIGHTED DATA .....	50
z-Test for Proportions with Partial Pairing.....	51
Technical Comment: A Note on Perfect Pairing.....	52
Part-Whole Comparisons.....	52
MULTIPLY WEIGHTED DATA .....	54
Z-TESTS – DEPENDENT PAIRED/OVERLAP (MULTI) .....	56
UNWEIGHTED DATA.....	56
Part-Whole Comparisons.....	57
SINGLY WEIGHTED DATA .....	58
Part-Whole Comparisons.....	59
MULTIPLY WEIGHTED DATA .....	60
COMPARING VOLUMETRIC PERCENTAGES .....	62
DEPENDENT PAIRED/OVERLAP ( LOC+/VAR+) .....	63
UNWEIGHTED DATA.....	63
SINGLY WEIGHTED DATA .....	66
MULTIPLY WEIGHTED DATA.....	67
DEPENDENT PAIRED/OVERLAP (MULTI).....	68
UNWEIGHTED DATA.....	68
SINGLY WEIGHTED DATA .....	70
MULTIPLY WEIGHTED DATA .....	72
COMPARISON WITH TOTAL .....	73
LOC+/VAR+: UNWEIGHTED & SINGLY WEIGHTED .....	74
MULTI: UNWEIGHTED & SINGLY WEIGHTED.....	77
ONE-WAY ANOVA .....	80
CHI-SQUARE .....	85
FACTOR ANALYSIS .....	86
SAMPLE BALANCING .....	90

<b>Appendix I .....</b>	<b>94</b>
<b>Critical Value for t-Distribution Table .....</b>	<b>95</b>

## 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  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ . 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  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ , and the frequencies of occurrence of these specific values in the data set will be denoted by  $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k$ . In that case,

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

When each of the observations is weighted we will denote the weights by  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ .

## 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

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

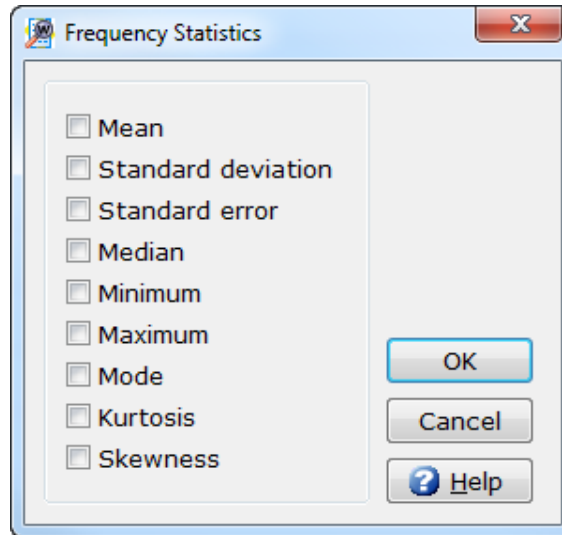
i.e., the cumulative count up to but not including row  $m$ .

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:

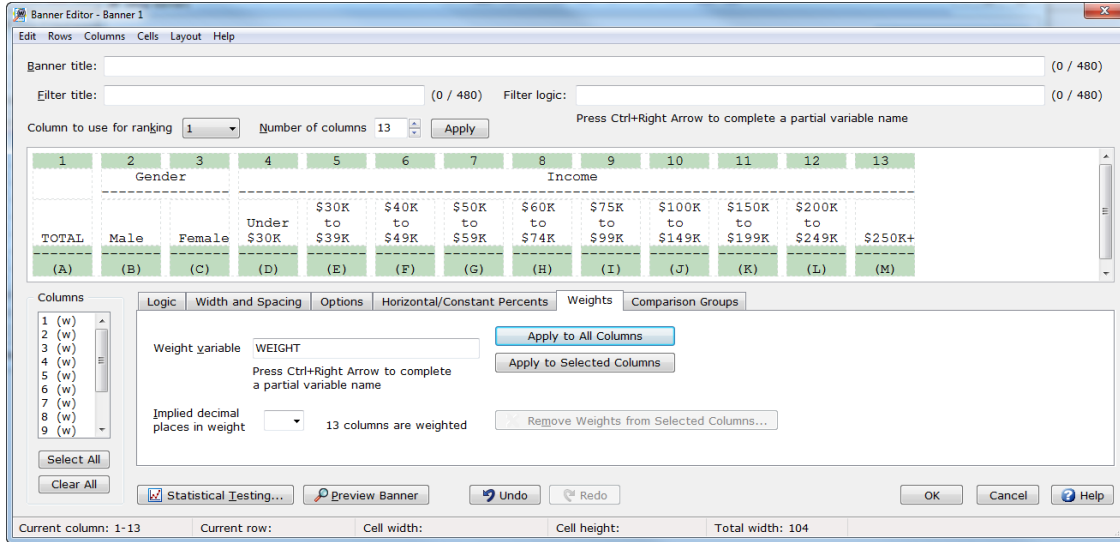
$$se_s = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

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

[http://www.analyticalgroup.com/support\\_wc\\_faqs.htm](http://www.analyticalgroup.com/support_wc_faqs.htm)

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

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

## Weighted Standard Error

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

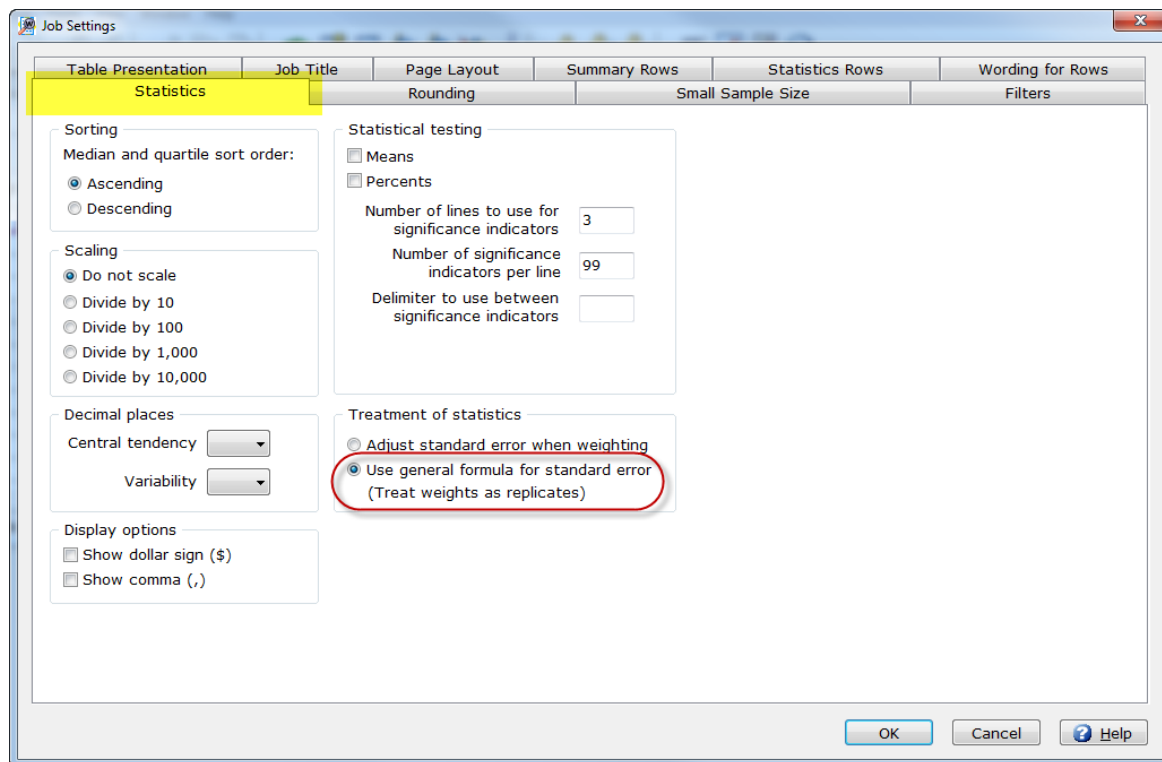
$$s_{\bar{x}_w} = s / \sqrt{e}$$

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

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

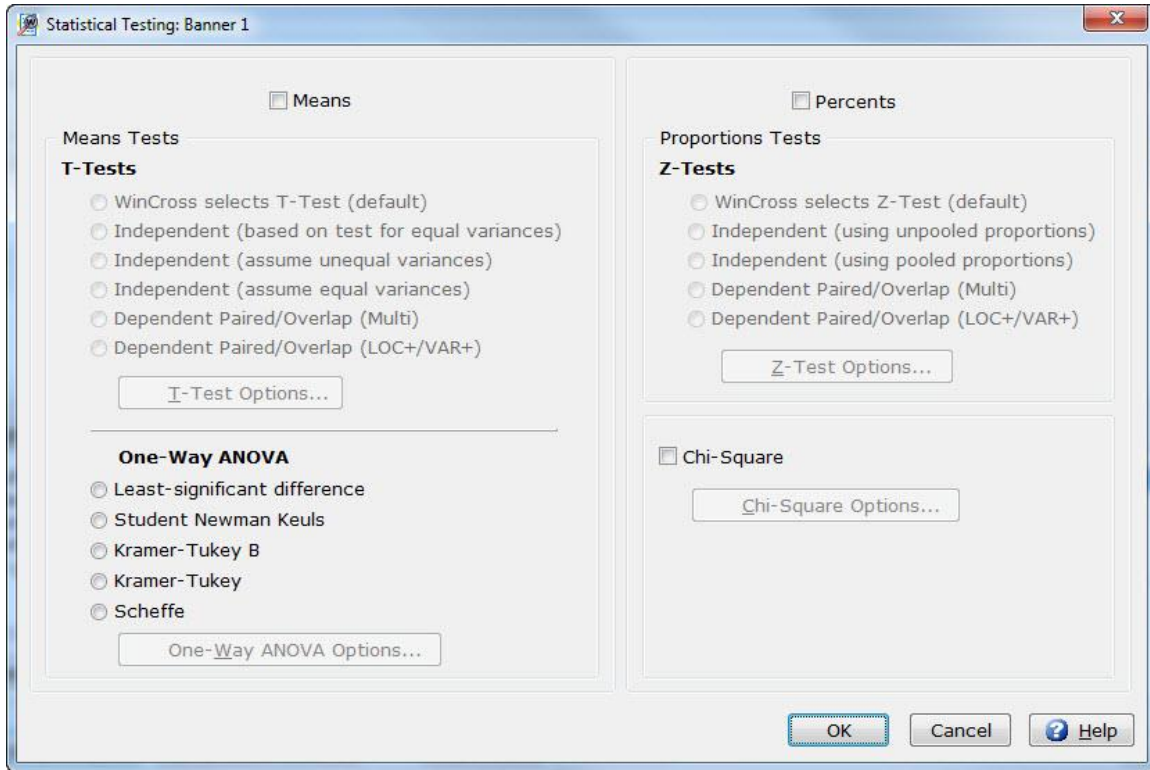
WinCross produces the weighted standard error  $s_{\bar{x}_w}^*$  given above as a descriptive statistic, but only as an option does it use it in calculating the t statistic for weighted data. To invoke this option, on the **Job Settings|Statistics** tab, select the **Use General formula for standard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

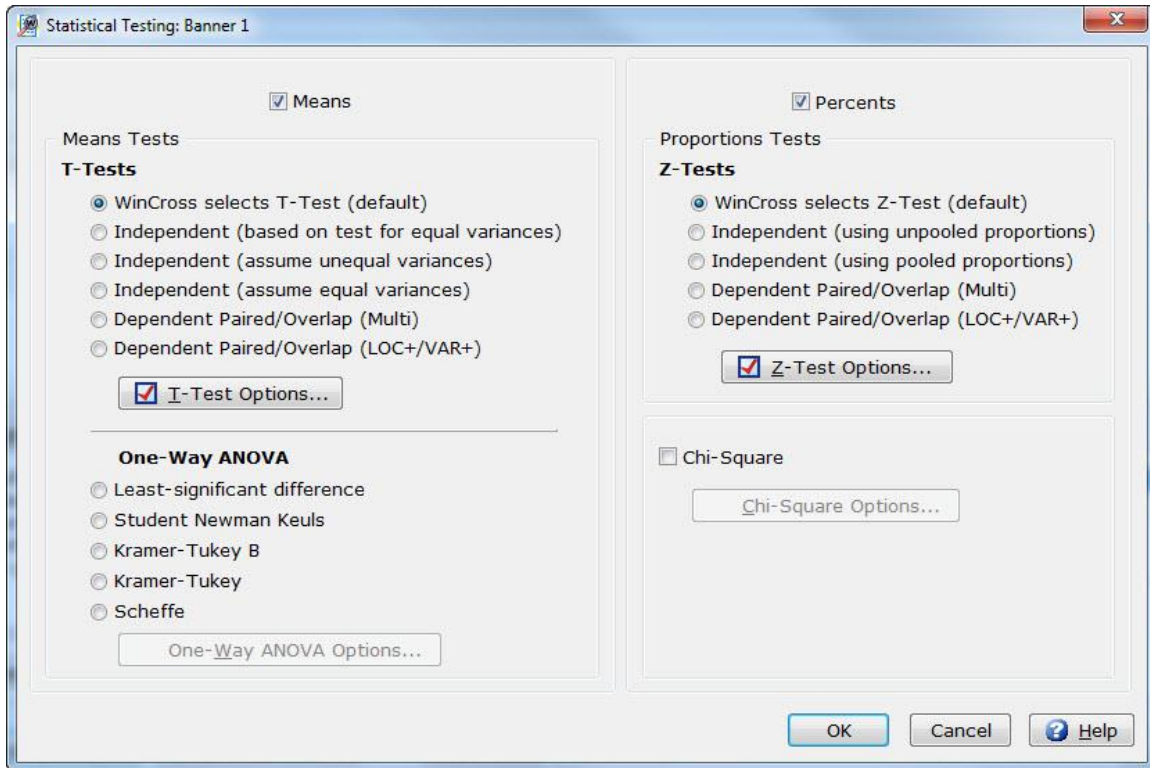
## SIGNIFICANCE TESTING

Following is the **Statistical Testing** window 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 80.) 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 85.)

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** window looks like this:

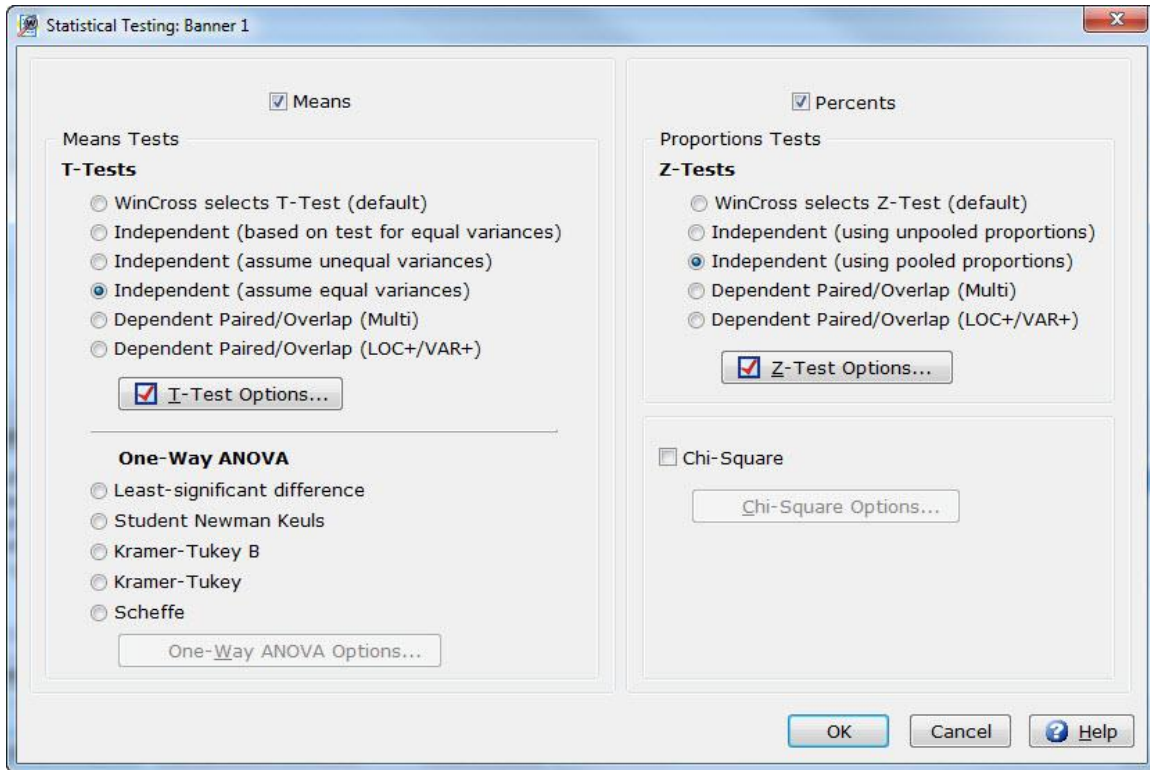


We note here that the first of the test options listed, “WinCross selects T-Test” and “WinCross selects Z-Test” are the “default” options, in that if the user does not check another option, WinCross will determine the appropriate test and perform it. WinCross knows from the structure of the table that 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.)

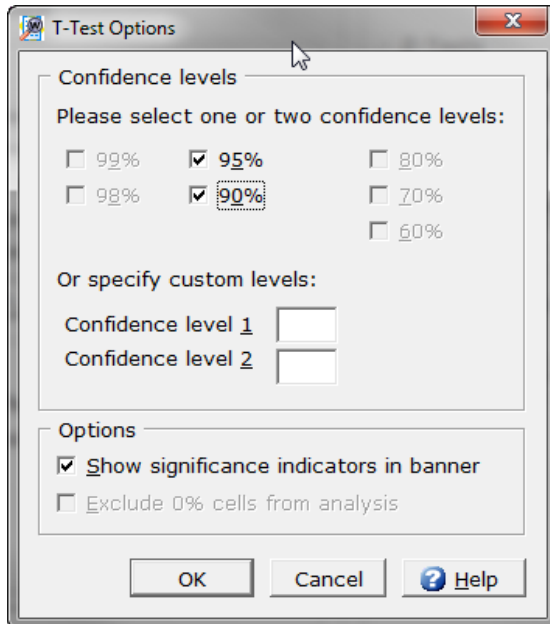
The independent t-test has two variants, one based on the assumption of equal variances and one based on the assumption of unequal variances. In many instances the user does not know whether or not to assume equal variances, and so WinCross has a built-in test which decides which assumption is more based 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.

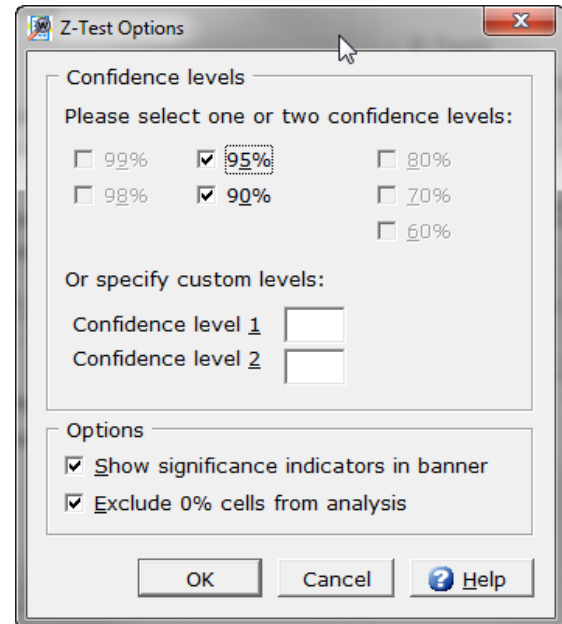
If one does not use the WinCross default, then one can click on the user-determined test variant, as illustrated below.



## T-TEST OPTIONS

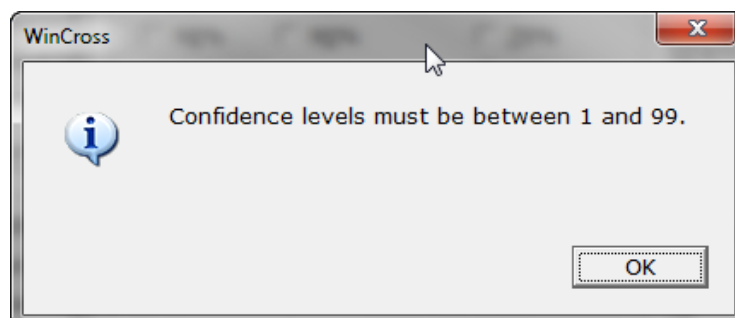


## Z-TEST OPTIONS



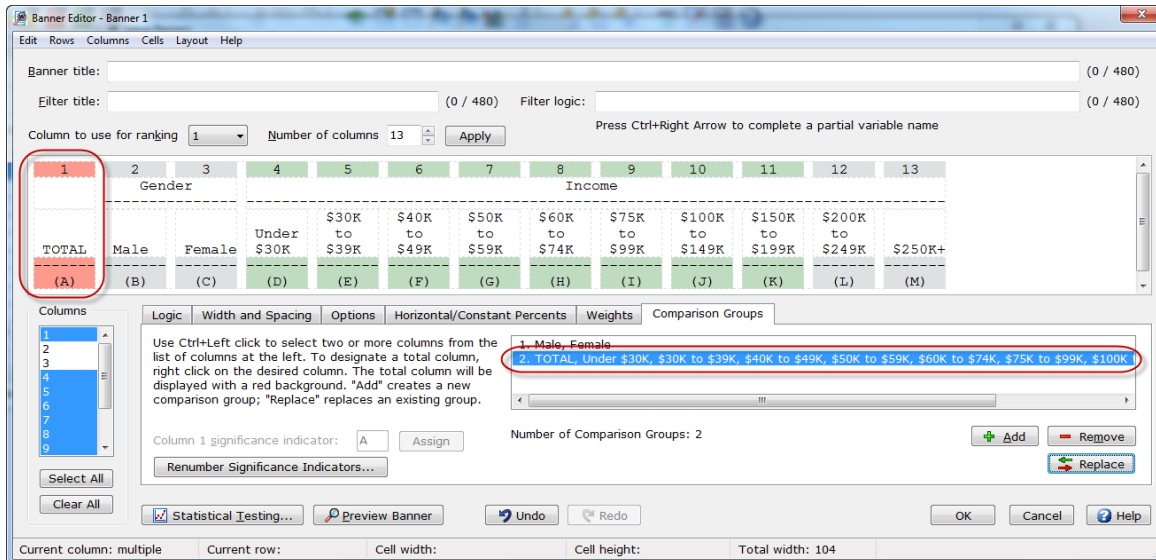
The significance tests are two-tailed tests. In performing either the T-tests or the Z-tests one can select up to two levels of significance, either corresponding to confidence levels of 60%, 70%, 80%, 90%, 95%, 98%, and 99% or any two confidence levels that you specify. If you choose to specify your 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  $\alpha$ , 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.



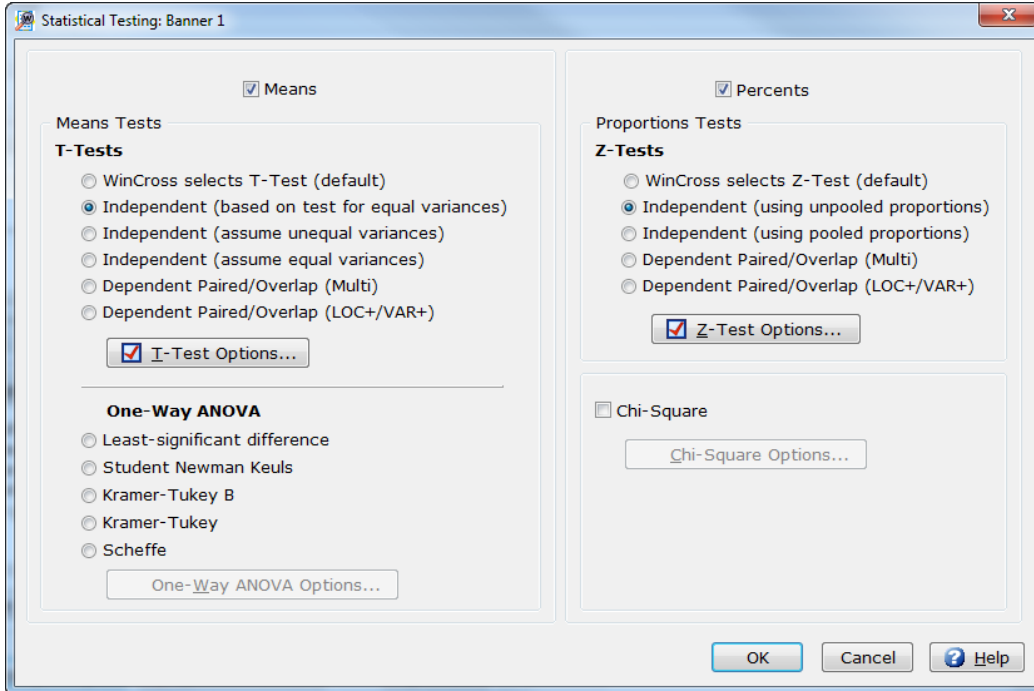
## T-TESTS - INDEPENDENT

### UNWEIGHTED DATA

#### General Notation

We consider here the situation in which we have data from two populations, where  $n_1$  is the number of observations in data set 1,  $n_2$  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  $\sigma^2$ , 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<sup>1</sup>, developed specifically for the two sample t test. The degrees of freedom of the Welch approximation is given by

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

---

<sup>1</sup> 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.

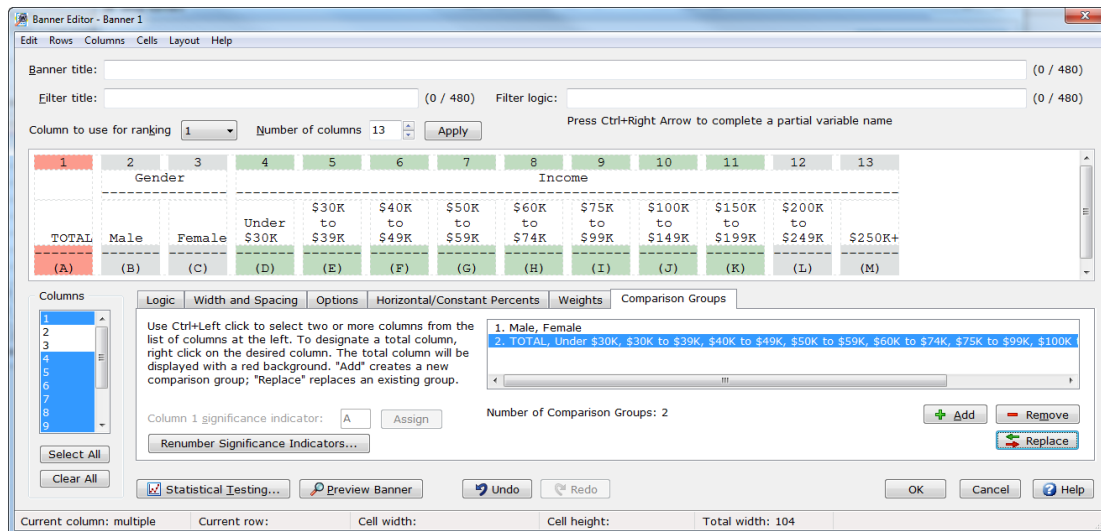
However, SPSS uses a different approximation, the Satterthwaite approximation<sup>2</sup>, which is a specialization to the two sample t test of a more general approximation useful in analysis of variance situations. The degrees of freedom of the Satterthwaite approximation is as given on the previous 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.

### Letting WinCross 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.

<sup>2</sup> 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

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

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

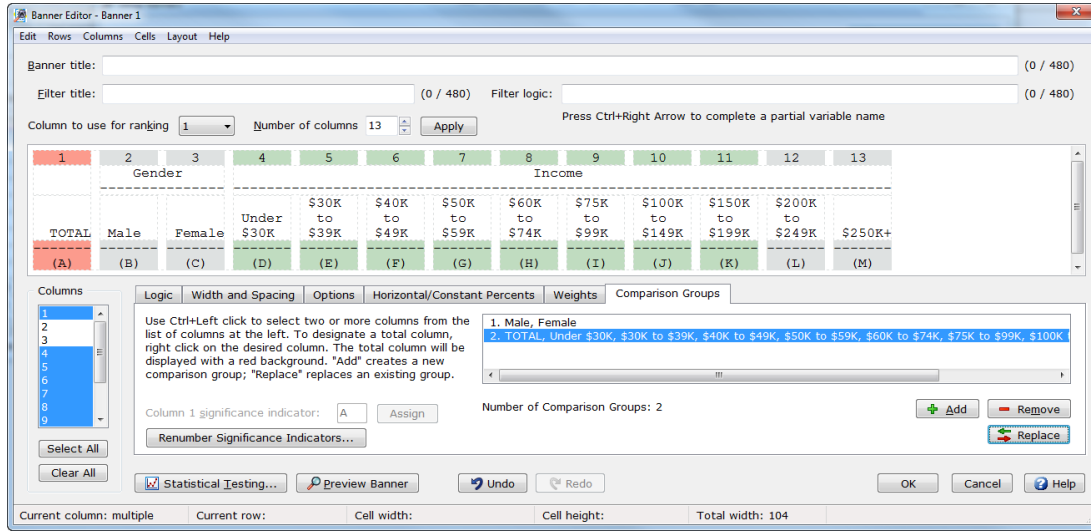
$$df_s = \frac{(n_1 - 1)(n_2 - 1)}{(n_1 - 1)(1 - c)^2 + (n_2 - 1)c^2}$$

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

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

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  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_2$  independent observations  $x_{21}, \dots, x_{2n_2}$  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  $(x_{11}, x_{21}), (x_{12}, x_{22}), \dots, (x_{1n_0}, x_{2n_0})$  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

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

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

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 unequal variances.**

#### Technical Comment:

##### 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

$$\text{Cov}(x_1, x_2) = E(x_1 - E x_1)(x_2 - E x_2),$$

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

$$\text{Cov}(x_1, x_2) = E x_1 x_2 - E x_1 E x_2$$

Thus, to estimate  $\text{Cov}(x_1, x_2)$  one might use the best estimates of  $E x_1 x_2$ ,  $E x_1$ , and  $E x_2$  in the computation. The best estimate of  $E x_1 x_2$  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  $E x_1$  is the mean of all the  $x_1$ ; the best estimate of  $E x_2$  is the mean of all the  $x_2$ . Putting all this together we obtain as an estimate of the sample covariance between the two sets of paired observations

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

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  $\text{Cov}(x_1, x_2) = E(x_1 - E x_1)(x_2 - E x_2)$  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:

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

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  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_j$  independent observations  $x_{j1}, \dots, x_{jn_j}$  from population  $j$ ,  $j=2, \dots, m$ . 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  $(x_{1i}, x_{ji})$   $i=1, \dots, n, j=2, \dots, m$  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  $(x_{1i}, x_{ji})$   $i=1, \dots, n, j=2, \dots, m$  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  $n_{0j}$  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

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

The 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

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

## SINGLY WEIGHTED DATA

### General Notation

Suppose we have  $n_1$  independent observations  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_2$  independent observations  $x_{21}, \dots, x_{2n_2}$  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  $w_{11}, \dots, w_{1n_1}$  the weights for the respondents from population 1,  $w_{21}, \dots, w_{2n_2}$  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.,  $w_{11} = w_{21} = w_1, \dots, w_{1n_0} = w_{2n_0} = w_{n_0}$ .

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

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

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.

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

$$t = \frac{\bar{x}_{1w} - \bar{x}_{2w}}{\sqrt{\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0c}{e_1e_2}}}$$

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 unequal variances.**

#### 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

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

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  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_j$  independent observations  $x_{j1}, \dots, x_{jn_j}$  from population  $j$ ,  $j=2, \dots, m$ . 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  $w_1, \dots, w_n$  and where the weights applied to each of the observations on the first  $n_0$  respondents are identical for all the observations, i.e.,  $w_{11} = w_{21} = w_1, \dots, w_{1n_0} = w_{2n_0} = w_{n_0}$ .

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

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

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

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

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

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

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

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

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

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

## MULTIPLY WEIGHTED DATA

### General Notation

Suppose we have  $n_1$  independent observations  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_2$  independent observations  $x_{21}, \dots, x_{2n_2}$  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  $w_{11}, \dots, w_{1n_1}$  the weights for the respondents from population 1,  $w_{21}, \dots, w_{2n_2}$  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.,

$$w_{11} \neq w_{21}, \dots, w_{1n_0} \neq w_{2n_0}.$$

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

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

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  $i$  ( $i = 1, 2, \dots, n$ ), by  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ), and by  $z_i$  the observed measurement for respondent  $i$  ( $i = n+m+1, n+m+2, \dots, n+m+p$ ). (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

$$\begin{aligned}\bar{X}_1 - \bar{X}_2 &= \frac{\sum_{i=1}^n x_i + \sum_{i=n+1}^{n+m} y_i}{n+m} - \frac{\sum_{i=1}^n x_i + \sum_{i=n+m+1}^{n+m+p} z_i}{n+p} \\ &= \left(\frac{1}{n+m} - \frac{1}{n+p}\right)n\bar{x} + \left(\frac{1}{n+m}\right)m\bar{y} - \left(\frac{1}{n+p}\right)p\bar{z}\end{aligned}$$

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

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

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

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

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  $m \leq 1$  then  $s_y^2$  is 0 and the second term in the expression for  $s_d^2$  is eliminated.

When  $p \leq 1$  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  $i$  ( $i = 1, 2, \dots, n$ ) and by  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ). (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

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

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

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

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

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

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

$$\begin{aligned} \bar{X}_{1w} - \bar{X}_{2w} &= \frac{\sum_{i=1}^n w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} - \frac{\sum_{i=1}^n w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{\sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i} \\ &= \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} - \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i} \right) \sum_{i=1}^n w_i x_i + \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \right) \sum_{i=n+1}^{n+m} w_i y_i - \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i} \right) \sum_{i=n+m+1}^{n+m+p} w_i z_i \\ &= \left( \frac{\sum_{i=n+m+1}^{n+m+p} w_i - \sum_{i=n+1}^{n+m} w_i}{\left[ \sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i \right] \left[ \sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i \right]} \right) \sum_{i=1}^n w_i x_i + \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \right) \sum_{i=n+1}^{n+m} w_i y_i - \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i} \right) \sum_{i=n+m+1}^{n+m+p} w_i z_i \end{aligned}$$

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  $m \leq 1$  then  $s_y^2$  is 0 and the second term in the expression for  $s_d^2$  is eliminated.

When  $p \leq 1$  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

$$\begin{aligned} \bar{X}_{1w} - \bar{X}_{Tw} &= \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} - \frac{\sum_{i=1}^n w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \\ &= \left( \frac{1}{\sum_{i=1}^n w_i} - \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \right) \sum_{i=1}^n w_i x_i - \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \right) \sum_{i=n+1}^{n+m} w_i y_i \\ &= \frac{\sum_{i=n+1}^{n+m} w_i}{\left[ \sum_{i=1}^n w_i \right] \left[ \sum_{i=1}^{n+m} w_i \right]} \sum_{i=1}^n w_i x_i - \frac{1}{\sum_{i=1}^{n+m} w_i} \sum_{i=n+1}^{n+m} w_i y_i \end{aligned}$$

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

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

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  $x_i$ s, 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

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

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

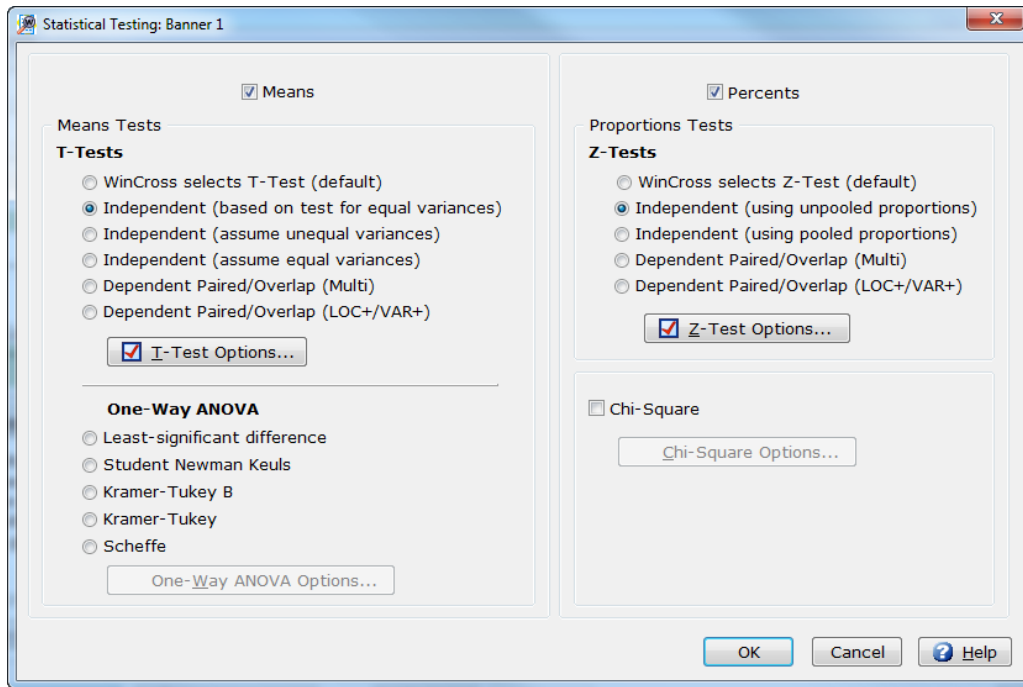
$$t = \frac{\bar{X}_{1w} - \bar{X}_{2w}}{s_d}$$

## Z-TESTS - INDEPENDENT

### General Notation

We consider here the situation in which we have proportions from two populations, where  $n_1$  is the number of observations in data set 1,  $n_2$  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  $p_1$  and  $p_2$ . 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

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

and finally calculate

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

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  $n_1, n_2, \dots, n_m$  in  $m$  categories (with  $n_1 + n_2 + \dots + n_m = n$ ), 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  $k$ ,  $k = 1, 2, \dots, m$  (where  $P_1 + P_2 + \dots + P_m = 1$ ). 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  $P$ s 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  $P_i = P_j$ .

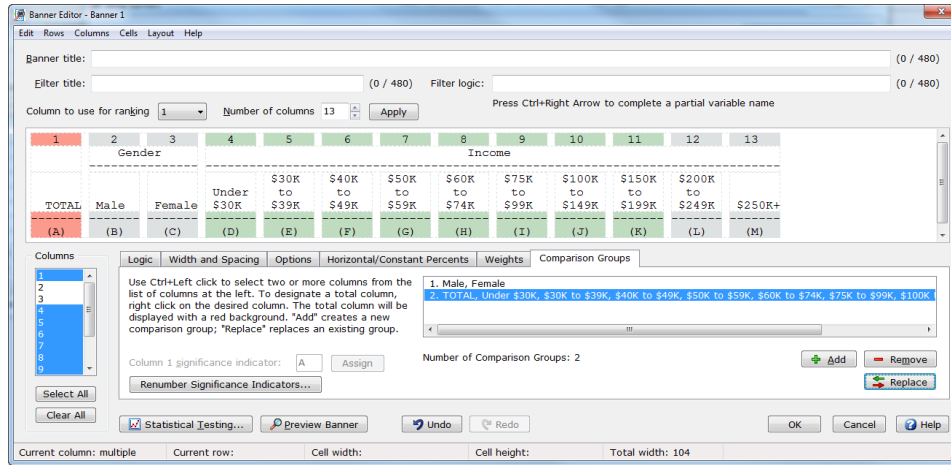
$$z = \frac{p_i - p_j}{\sqrt{\frac{p_i(1-p_i) + p_j(1-p_j) + 2p_i p_j}{n}}}$$

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](http://www.analyticalgroup.com) program found on our website: [www.analyticalgroup.com](http://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.

## 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,  $\mathbf{p}_1(1 - \mathbf{p}_1)$  and  $\mathbf{p}_2(1 - \mathbf{p}_2)$ .

### 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

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

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

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

where

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

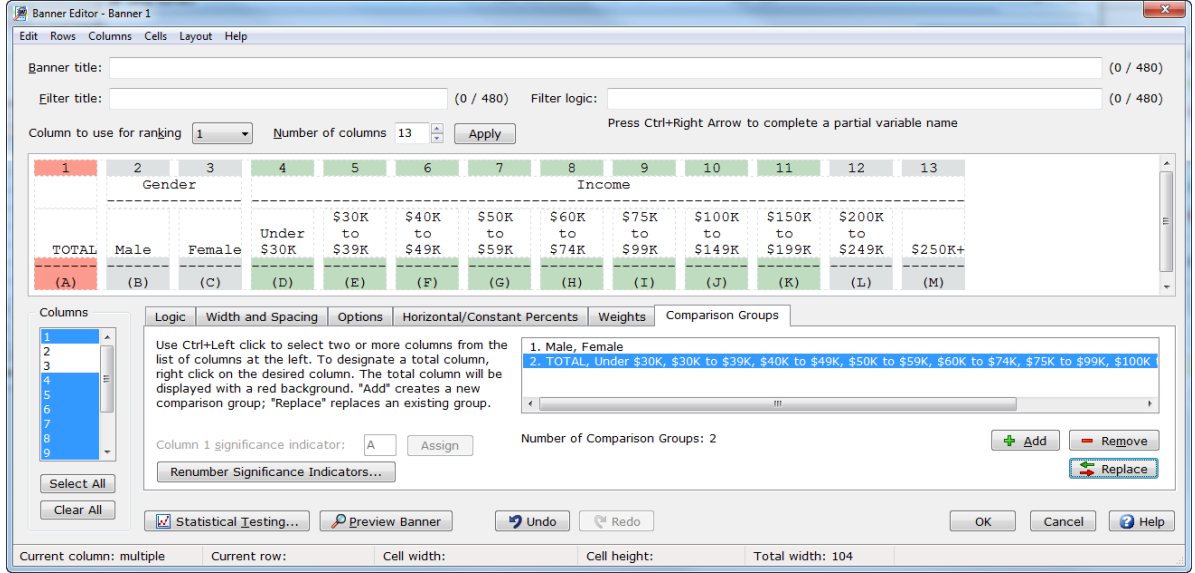
is the unweighted pooled proportion.

### 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

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

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  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_2$  independent observations  $x_{21}, \dots, x_{2n_2}$  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

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

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

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

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

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

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  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_j$  independent observations  $x_{j1}, \dots, x_{jn_j}$  from population  $j$ ,  $j=2, \dots, m$ , 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  $(x_{1i}, x_{ji})$   $i=1, \dots, n$ ,  $j=2, \dots, m$  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

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

For each of the  $m-1$  pairs of observations  $(x_{1i}, x_{ji})$   $i=1, \dots, n$ ,  $j=2, \dots, m$  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  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_2$  independent observations  $x_{21}, \dots, x_{2n_2}$  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  $w_{11}, \dots, w_{1n_1}$  the weights for the respondents from population 1,  $w_{21}, \dots, w_{2n_2}$  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.,  $w_{11} = w_{21} = w_1, \dots, w_{1n_0} = w_{2n_0} = w_{n_0}$ .

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

$$\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0c}{e_1e_2},$$

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

$$z = \frac{p_{1w} - p_{2w}}{\sqrt{\frac{s_1^2}{e_1} + \frac{s_2^2}{e_2} - \frac{2e_0c}{e_1e_2}}}$$

### 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  $x_{11}, \dots, x_{1n_1}$  from population 1 and  $n_j$  independent observations  $x_{j1}, \dots, x_{jn_j}$  from population  $j$ ,  $j=2, \dots, m$ , 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  $w_1, \dots, w_n$  and where the weights applied to each of the observations on the first  $n_0$  respondents are identical for all the observations, i.e.,  $w_{11} = w_{21} = w_1, \dots, w_{1n_0} = w_{2n_0} = w_{n_0}$ .

The two sample proportions are

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

The two weighted sample proportions are

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

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

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

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

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

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

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

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

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

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  $i$  ( $i = 1, 2, \dots, n$ ), by  $y_i$  the observed attribute value for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ), and by  $z_i$  the observed attribute value for respondent  $i$  ( $i = n+m+1, n+m+2, \dots, n+m+p$ ). (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

$$\begin{aligned} q_1 - q_2 &= \frac{\sum_{i=1}^n x_i + \sum_{i=n+1}^{n+m} y_i}{n+m} - \frac{\sum_{i=1}^n x_i + \sum_{i=n+m+1}^{n+m+p} z_i}{n+p} \\ &= \left(\frac{1}{n+m} - \frac{1}{n+p}\right)nq_x + \left(\frac{1}{n+m}\right)mq_y - \left(\frac{1}{n+p}\right)pq_z \end{aligned}$$

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

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

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  $i$  ( $i = 1, 2, \dots, n$ ) and by  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ). (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

$$\begin{aligned}
p_1 - p_T &= \frac{\sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n x_i + \sum_{i=n+1}^{n+m} y_i}{n+m} \\
&= \left(\frac{1}{n} - \frac{1}{n+m}\right)n\bar{x} - \left(\frac{1}{n+m}\right)m\bar{y}
\end{aligned}$$

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

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

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

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

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

$$\begin{aligned}
q_{1w} - q_{2w} &= \frac{\sum_{i=1}^n w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} - \frac{\sum_{i=1}^n w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{\sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i} \\
&= \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} - \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i} \right) \sum_{i=1}^n w_i x_i + \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \right) \sum_{i=n+1}^{n+m} w_i y_i - \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i} \right) \sum_{i=n+m+1}^n w_i z_i \\
&= \left( \frac{\sum_{i=n+m+1}^{n+m+p} w_i - \sum_{i=n+1}^{n+m} w_i}{\left[ \sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i \right] \left[ \sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i \right]} \right) \sum_{i=1}^n w_i x_i + \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \right) \sum_{i=n+1}^{n+m} w_i y_i - \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+m+1}^{n+m+p} w_i} \right) \sum_{i=n+m+1}^n w_i z_i
\end{aligned}$$

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

$$\begin{aligned} p_{1w} - p_{Tw} &= \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} - \frac{\sum_{i=1}^n w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \\ &= \left( \frac{1}{\sum_{i=1}^n w_i} - \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \right) \sum_{i=1}^n w_i x_i - \left( \frac{1}{\sum_{i=1}^n w_i + \sum_{i=n+1}^{n+m} w_i} \right) \sum_{i=n+1}^{n+m} w_i y_i \\ &= \frac{\sum_{i=n+1}^{n+m} w_i}{\left[ \sum_{i=1}^n w_i \right] \left[ \sum_{i=1}^{n+m} w_i \right]} \sum_{i=1}^n w_i x_i - \frac{1}{\sum_{i=1}^{n+m} w_i} \sum_{i=n+1}^{n+m} w_i y_i \end{aligned}$$

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.

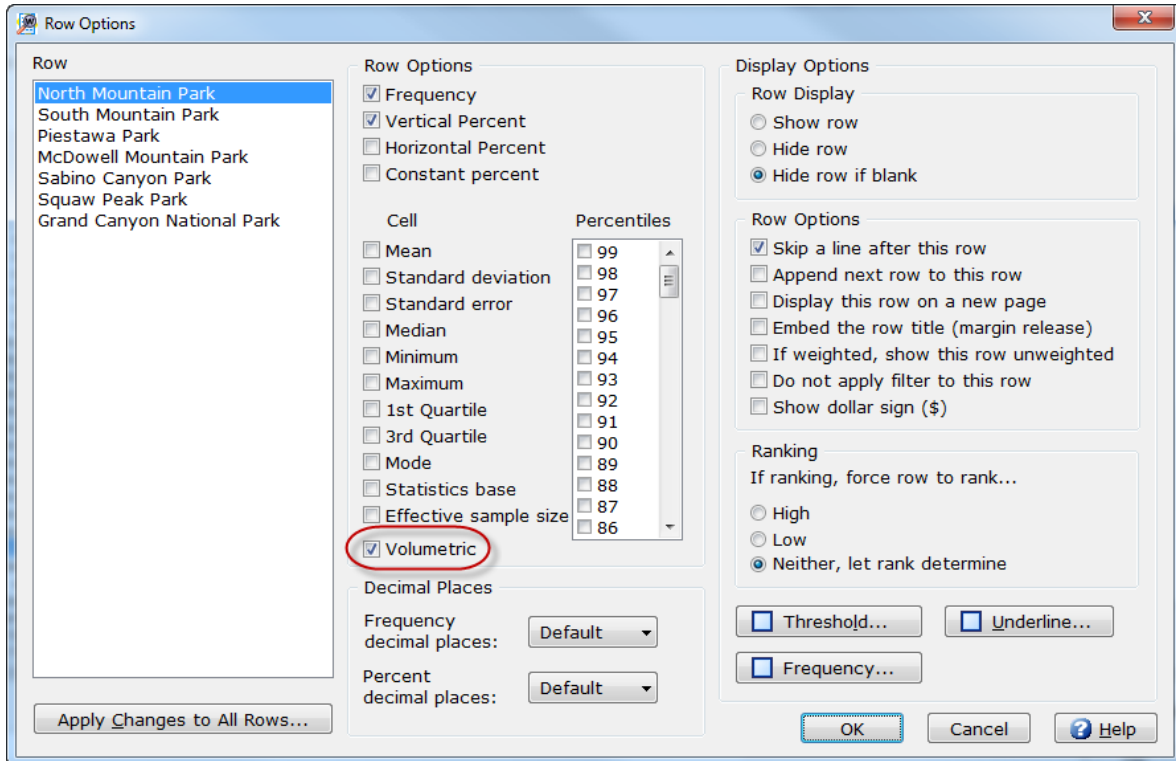
The z-statistic for testing the difference of proportions is given by

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

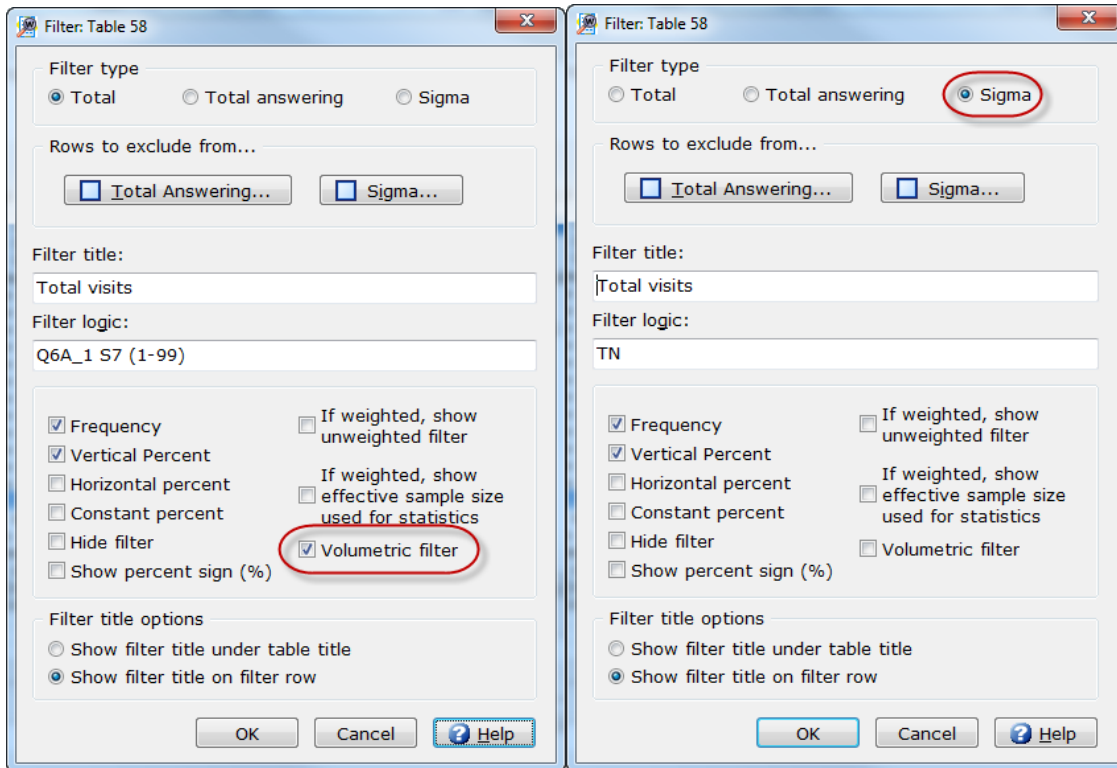


## COMPARING VOLUMETRIC PERCENTAGES

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  $i$  ( $i = 1, 2, \dots, n$ ), by  $x_{2i}$  the observed measurement for column 2 for respondent  $i$  ( $i = 1, 2, \dots, n$ ), by  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ), and by  $z_i$  the observed measurement for respondent  $i$  ( $i = n+m+1, n+m+2, \dots, n+m+p$ ). (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

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

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

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

where  $\sigma_{x1}^2$  is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2,  $\sigma_{x2}^2$  is the variance of the measurements in column 2 of those respondents who qualified for both columns 1 and 2,  $\rho$  is the correlation between the measurements in column 1 and column 2 of those respondents who qualified for both columns 1 and 2,  $\sigma_y^2$  is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and  $\sigma_z^2$  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

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

## SINGLY WEIGHTED DATA

Let us denote by  $x_{1i}$  the observed measurement for column 1 for respondent  $i$  ( $i = 1, 2, \dots, n$ ), by  $x_{2i}$  the observed measurement for column 2 for respondent  $i$  ( $i = 1, 2, \dots, n$ ), by  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ), and by  $z_i$  the observed measurement for respondent  $i$  ( $i = n+m+1, n+m+2, \dots, n+m+p$ ). (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

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

The difference of the two percentages is given by

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

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

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

where  $\sigma_{x1}^2$  is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2,  $\sigma_{x2}^2$  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,  $\sigma_y^2$  is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and  $\sigma_z^2$  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

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

### MULTIPLY WEIGHTED DATA

Let us denote by  $x_{1i}$  the observed measurement for column 1 for respondent  $i$  ( $i = 1, 2, \dots, n$ ), by  $x_{2i}$  the observed measurement for column 2 for respondent  $i$  ( $i = 1, 2, \dots, n$ ), by  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ), and by  $z_i$  the observed measurement for respondent  $i$  ( $i = n+m+1, n+m+2, \dots, n+m+p$ ). (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

$$d = p_{1w} - p_2 = \frac{\sum_{i=1}^n w_{1i} x_{1i} + \sum_{i=n+1}^{n+m} w_{1i} y_i}{X_{1w}} - \frac{\sum_{i=1}^n w_{2i} x_{2i} + \sum_{i=n+m+1}^{n+m+p} w_{2i} z_i}{X_{2w}}$$

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

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

where  $\sigma_{x1}^2$  is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2,  $\sigma_{x2}^2$  is the variance of the measurements in column 2 of those respondents who qualified for both columns 1 and 2,  $\rho$  is the correlation between the measurements in column 1 and column 2 of those respondents who qualified for both columns 1 and 2,  $\sigma_y^2$  is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and  $\sigma_z^2$  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  $i$  ( $i = 1, 2, \dots, n$ ),  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ), and by  $z_i$  the observed measurement for respondent  $i$  ( $i = n+m+1, n+m+2, \dots, n+m+p$ ). (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

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

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

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

where  $\sigma_x^2$  is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2,  $\sigma_y^2$  is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and  $\sigma_z^2$  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

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

### SINGLY WEIGHTED DATA

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

(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

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

where  $\sigma_x^2$  is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2,  $\sigma_y^2$  is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and  $\sigma_z^2$  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

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

## MULTIPLY WEIGHTED DATA

Let us denote by  $x_i$  the observed measurement for both columns 1 and 2 for respondent  $i$  ( $i = 1, 2, \dots, n$ ),  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ), and by  $z_i$  the observed measurement for respondent  $i$  ( $i = n+m+1, n+m+2, \dots, n+m+p$ ). (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  $\sigma_x^2$  is the variance of the measurements in column 1 of those respondents who qualified for both columns 1 and 2,  $\sigma_y^2$  is the variance of the measurements in column 1 of those respondents who only qualified for column 1, and  $\sigma_z^2$  is the variance of the measurements in column 2 of those respondents who only qualified for column 2,

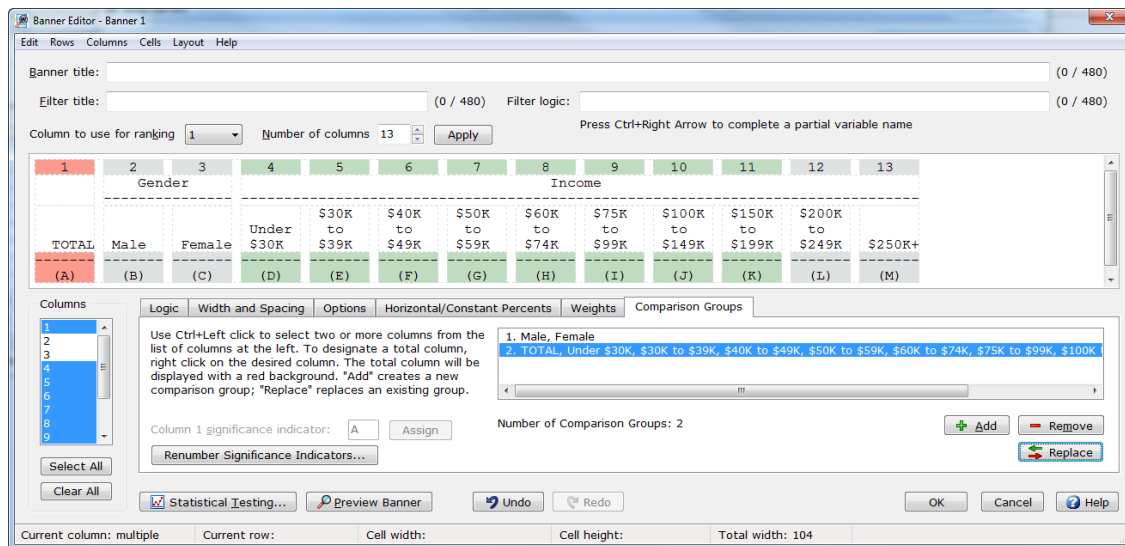
The estimate of the variance of the difference of the two percentages is given by

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

## 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.

	Total	Volume of soft drinks consumed		
		Coke	Pepsi	Sprite
	-----	-----	-----	-----
Total	14618 100.0%	5539 100.0%	2842 100.0%	3002 100.0%
breakfast	2283 15.6%	850 15.3%	438 15.4%	491 16.4%
lunch	3776 25.8%	1424 25.7%	714 25.1%	785 26.1%
dinner	5381 36.8%	2094 37.8%	998 35.1%	1084 36.1%
other	3178 21.7%	1171 21.1%	692 24.3%	642 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  $\delta_{ji}$  as 1 if respondent  $i$  answered item  $j$  and as 0 if respondent  $i$  did not answer item  $j$ , for  $i = 1, 2, \dots, n$  and  $j = 1, 2, \dots, c$ . Let us denote by  $x_{ji}\delta_{ji}$  the observed measurement for column  $j$  for respondent  $i$ . (As you can see, the  $\delta_{ji}$  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  $\delta_{ji}$  have the same value for all the columns, so we will designate that common value as the  $\delta_{Ti}$ .

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

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

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

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

where  $n_T$  is the number of respondents contributing to the total column,  $\sigma_j^2$  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

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

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

$$\begin{aligned} d_w = p_{1w} - p_{Tw} &= \frac{\sum_{i=1}^n x_{1i} w_i \delta_{Ti}}{X_{1w}} - \frac{\sum_{j=1}^c \sum_{i=1}^n x_{ji} w_i \delta_{Ti}}{X_{Tw}} \\ &= \left( \frac{1}{X_{1w}} - \frac{1}{X_{Tw}} \right) \sum_{i=1}^{n_1} x_{1i} w_i \delta_{Ti} - \left( \frac{1}{X_{Tw}} \right) \sum_{j=2}^c \sum_{i=1}^n x_{ji} w_i \delta_{Ti} \end{aligned}$$

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

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

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

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

## 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 percentages that oil changes at the dealer make up of the total number of oil changes made by Ford owners and by all car owners.

	Number of oil changes per year by respondents who have ever owned a			
	Total	Ford	Chevy	VW
	-----	-----	-----	-----
Total	3893 100.0%	1715 100.0%	2169 100.0%	1115 100.0%
At dealer	1905 48.9%	822 47.9%	1071 49.4%	540 48.4%
At garage	1348 34.6%	609 35.5%	756 34.9%	392 35.2%
Elsewhere	640 16.4%	284 16.6%	342 15.8%	183 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  $i$  ( $i = 1, 2, \dots, n$ ),  $y_i$  the observed measurement for respondent  $i$  ( $i = n+1, n+2, \dots, n+m$ ). 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

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

where  $\sigma_x^2$  is the variance of the measurements in column 1 of those respondents who qualified for column 1 and  $\sigma_y^2$  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

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

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

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

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

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

## 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  $\alpha$ , 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  $n_1, n_2, \dots, n_m$  denote the sample sizes from the  $m$  populations, and let  $x_{ij}$  ( $i=1, 2, \dots, m, j=1, 2, \dots, n_i$ ) 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  $\bar{x}_1 \leq \bar{x}_2 \leq \dots \leq \bar{x}_m$ . 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,  $\bar{x}_2, \bar{x}_3, \dots, \bar{x}_m$  and  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{m-1}$  to see if they are homogeneous. To check the first subset  $\bar{x}_2, \bar{x}_3, \dots, \bar{x}_m$  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,  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{m-2}$  and  $\bar{x}_2, \bar{x}_3, \dots, \bar{x}_{m-1}$  to see if they are homogeneous.

Similarly, to check the second subset  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{m-1}$  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  $\bar{x}_2, \bar{x}_3, \dots, \bar{x}_m$  and have labeled each of them with a #1 and no significant difference in the second subset,  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{m-1}$ , 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  $\bar{x}_2, \bar{x}_3, \dots, \bar{x}_{m-1}$  will be labeled both #1 and #2.

If we did find a significant difference in the first subset  $\bar{x}_2, \bar{x}_3, \dots, \bar{x}_m$  and no significant difference in the second subset,  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{m-1}$  then the members of the second subset are each labeled with a #1. And now we have to drill down further within  $\bar{x}_2, \bar{x}_3, \dots, \bar{x}_m$ . 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,  $\bar{x}_3, \bar{x}_2, \dots, \bar{x}_m$  and  $\bar{x}_2, \bar{x}_3, \dots, \bar{x}_{m-1}$  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 (2.57) 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 (1.42). 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  $\bar{x}_i, \bar{x}_{i+1}, \dots, \bar{x}_j$  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  $\bar{x}_j - \bar{x}_i$  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  $100\alpha$  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  $\bar{x}_j - \bar{x}_i$  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  $100\alpha$  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  $100\alpha$  percent point of the Studentized range distribution with  $m$  and  $f$  degrees of freedom and  $S_{\alpha,k,f}$  is the upper  $100\alpha$  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

$$s \sqrt{\frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right) S_{\alpha,k,f}}$$

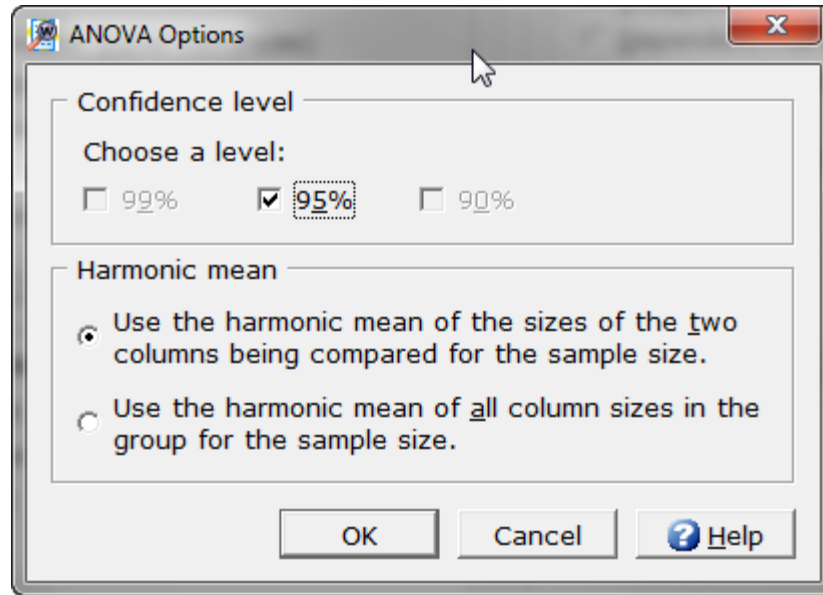
where  $S_{\alpha,k,f}$  is the upper  $100\alpha$  percent point of the Studentized range distribution with  $k$  and  $f$  degrees of freedom, and where  $k=i-j+1$ . (In SPSS this procedure is the Tukey HSD Post Hoc Multiple Comparisons test. In MINITAB this procedure is called the Tukey procedure.)

### **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  $100\alpha$  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  $\alpha$ , 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 compatability 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 ( $i=1,\dots,R, j=1,\dots,C$ ). 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<sup>th</sup> 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 program begins with the  $p \times p$  correlation matrix  $R$  of the questionnaire items. Let  $f$  be the number of factors underlying the responses. Let  $L$  be the  $p \times f$  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  $p \times f$  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  $p \times 1$  vector which we will call  $X_i$ . The  $i$ -th respondent’s factor scores can be arrayed as an  $f \times 1$  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  $p \times n$  matrix  $X = [X_1 X_2 \cdots X_n]$  and all  $n$  respondents’ factor score vectors into an  $f \times n$  matrix  $F = [F_1 F_2 \cdots F_n]$ . Then  $X = LF$ , and we can “solve” this equation for  $F$  as

$$F = (L^T L)^{-1} L^T X.$$

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.

### WinCross Factor Analysis dialogs

Following is the first of the WinCross dialogs used in its Factor Analysis:

**Run Factor Analysis**

Find a variable:  Find Next

#	Variable Name	Variable Label	Value Range
2	GENDER	Gender	1-9
3	Q1	Q.1 On average, how many hours per week do you spend participating in...	1-9
14	Q3_1	Q.3 How often do you: Send or read e-mail	1-9
15	Q3_2	Q.3 How often do you: Read news and current events	1-9

Selected variables:

#	Variable Name	Variable Label	Value Range
4	Q2_1	Q.2 Agreement with the following statement: I consider myself an outd...	1-4
5	Q2_2	Q.2 Agreement with the following statement: I wish the parks were ope...	1-4
6	Q2_3	Q.2 Agreement with the following statement: I wish all of the local p...	1-4
7	Q2_4	Q.2 Agreement with the following statement: I trust that the Arizona ...	1-4
8	Q2_5	Q.2 Agreement with the following statement: I usually visit an Arizon...	1-4

Blank/out-of-range responses  
☒ Compute with mean fill  
☐ Compute without mean fill

Cases  
 Recgdrs per case:   
 Cases to write: 400

Range  
 Minimum value:   
 Maximum value:

Reversal  
☐ Reverse scale

Factor Analysis options  
☒ Use glossary transformations  
☒ Write factor scores to data file  
☒ Write factor groups to data file  
☒ Show variable labels in report  
 ☒ Report Properties...

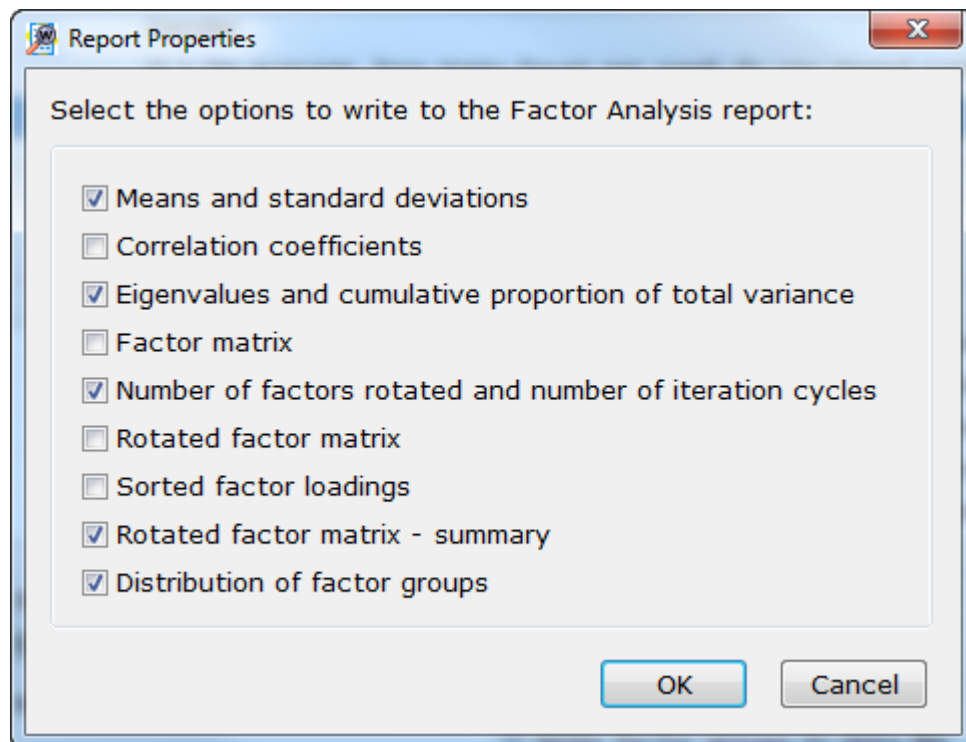
Minimum Eigenvalue criterion: 1.0000  
 Maximum number of factor groups: 99

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 “Max 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” creates f factor scores, using the regression method described above, and adds them as f additional columns in the data file. The “Write factor groups to data file” 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.

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  $p \times p$  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.

## SAMPLE BALANCING

The goal of the "Sample Balancing" program 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). WinCross's adaptation was developed by J. Stephens Stock, a colleague of Deming, in the 1960s.

### 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,  $i=1, \dots, v$ . Let  $p_{j_1 \dots j_v}$  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  $j_i = 1, \dots, c_i$ . Let  $f_{j_i}^i$  denote the marginal proportion in the sample of level  $j_i$  of variable  $i$  ( $j_i=1, \dots, c_i$ ,  $i=1, \dots, v$ ).

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  $f_{j_i}^i = f_{j_3}^3 = f_2^3$  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).

$f_{j_i}^i$  can be determined by adding up all the  $p_{j_1 \dots j_v}$  across all the values of each of the  $v-1$   $j_k$  for which  $k \neq i$ . For example, to obtain  $f_2^3$  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

$$f_{j_i}^i = \sum_{j_k, k \neq i} p_{j_1 \dots j_v}$$

These  $f_{j_i}^i$  are called sample rim percents.

Suppose that the preset distributions on the  $v$  variables are given by the set of target proportions  $g_{j_i}^i$ . The object of the sample balancing program is to find a set of weights  $w_{j_1 \dots j_v}$  such that if, when looking at the  $j_i$ -th break, instead of adding up the  $p_{j_1 \dots j_v}$  across

all but the i-th category, we add up the  $w_{j_1 \dots j_v} p_{j_1 \dots j_v}$  across all but the i-th category, we will obtain the  $g_{j_i}^i$ . That is,

$$g_{j_i}^i = \sum_{j_k, k \neq i} w_{j_1 \dots j_v} p_{j_1 \dots j_v}$$

These  $g_{j_i}^i$  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.

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  $w_{j_1 \dots j_v}(0, i)$  equal to 1, i.e., we begin with the unweighted data.

Suppose we are on the i-th "pass" in "round  $t+1$ ." Let  $w_{j_1 \dots j_v}(t, i)$  denote the weights at this point in the iterative process. Let  $g_{j_i}^i(t)$  denote the results of the computation

$$g_{j_i}^i(t) = \sum_{j_k, k \neq i} w_{j_1 \dots j_v}(t, i) p_{j_1 \dots j_v}$$

These  $g_{j_i}^i(t)$  are called estimated target rim percents.

At the first pass ( $i=1$ ) of the  $t$ -th round of the iterative procedure the program calculates a set of increments  $d_{j_1 \dots j_v}(t, 1)$  to add to the  $w_{j_1 \dots j_v}(t-1, v)$ , producing  $w_{j_1 \dots j_v}(t, 1) = w_{j_1 \dots j_v}(t-1, v) + d_{j_1 \dots j_v}(t, 1)$ . At the  $i$ -th pass ( $i > 1$ ) of the  $t$ -th round of the iterative procedure the program calculates a set of increments  $d_{j_1 \dots j_v}(t, i)$  to add to the  $w_{j_1 \dots j_v}(t, i-1)$ , producing  $w_{j_1 \dots j_v}(t, i) = w_{j_1 \dots j_v}(t, i-1) + d_{j_1 \dots j_v}(t, i)$ .

These increments are given by the formula

$$d_{j_1 j_2 \dots j_v} = [g_{j_i}^i(t) - g_{j_i}^i] / f_{j_i}^i$$

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 sample balancing program 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 program continues iterating until a criterion of goodness of fit has been met. WinCross uses the measure

$$\sqrt{\sum_{i=1}^v \sum_{j_i=1}^{c_i} [(g_{j_i}^i(t) - g_{j_i}^i) / f_{j_i}^i]^2}$$

This measure is the sum of squares of deviations across levels and variables 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 program iterates until this measure is less than some preset value (with default set at 0.0005). This is in contrast to iterative proportional fitting, which has no overall criterion and iterates until each  $g_{j_i}^i(t)$  is within some preset distance from  $g_{j_i}^i$ , that is, until each estimated target rim percent is within some preset distance from the actual target rim percent.

You can find an example of the use of the WinCross sample balancing program and a contrast with that of iterative proportional fitting on our web site:

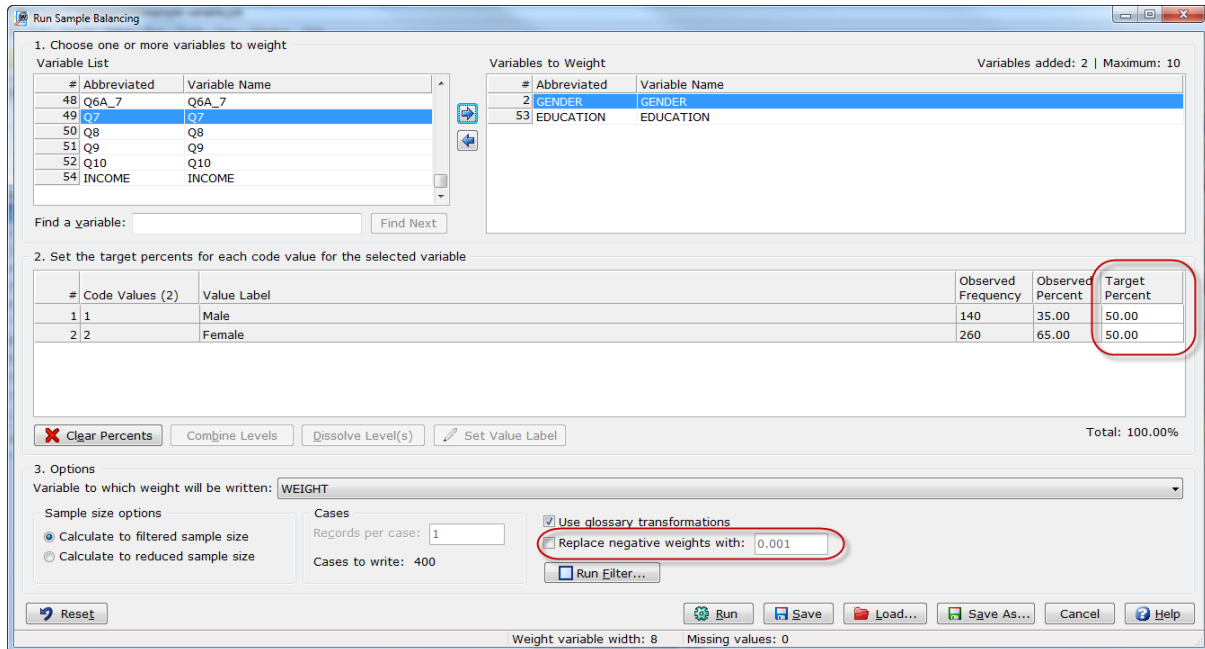
[WinCross's Sample Balancing Program](#)

For WinCross the value of the goodness of fit is 0.000431586; for iterative proportional fitting this value is 0.006785959, over 15 times as large.

One situation that might occur in using the WinCross sample balancing program 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.

## WinCross Sample Balancing dialog

Following is the WinCross dialog used in its Sample Balancing:



The WinCross Sample Balancing dialog box is shown with the following sections:

**1. Choose one or more variables to weight**

**Variable List**

#	Abbreviated	Variable Name
48	Q6A_7	Q6A_7
49	Q7	Q7
50	Q8	Q8
51	Q9	Q9
52	Q10	Q10
54	INCOME	INCOME

**Variables to Weight**

#	Abbreviated	Variable Name
2	GENDER	GENDER
53	EDUCATION	EDUCATION

**2. Set the target percents for each code value for the selected variable**

#	Code Values (2)	Value Label	Observed Frequency	Observed Percent	Target Percent
1	1	Male	140	35.00	50.00
2	2	Female	260	65.00	50.00

**3. Options**

Variable to which weight will be written: **WEIGHT**

**Sample size options**

☒ Calculate to filtered sample size  
☐ Calculate to reduced sample size

**Cases**

Records per case: 1  
Cases to write: 400

☒ Use glossary transformations  
☐ Replace negative weights with: 0.001  
☐ Run Filter...

**Buttons:** Clear Percents, Combine Levels, Dissolve Level(s), Set Value Label, Run, Save, Load..., Save As..., Cancel, Help

**Status:** Weight variable width: 8, Missing values: 0

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 program. 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.



## APPENDIX I

Our web site, [www.AnalyticalGroup.com](http://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

	<u>Confidence Level</u>						
	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