SPECIAL CONSIDERATIONS FOR VOLUMETRIC Z-TEST FOR PROPORTIONS

One's instinctive reaction to the question of whether two percentages are significantly different from each other is to treat them as if they were proportions in which the denominator is the sample size and the numerator has a binomial distribution, and then apply the standard statistical test for significant difference of proportions. But in fact that is not the case.

Consider the following example. A respondent is asked how many bottles of each of 5 soft drinks he consumes in a week. His data then gets summarized into 5 percentages, namely his consumption percentage of each of the soft drinks. The average of these percentages is calculated across a random sample of respondents, and the question of whether the average percentage of Coke is significantly different from that of Pepsi.

Now consider a second example. A random sample of respondents is asked four questions, (1) Have you ever eaten at the Ritz-Carleton restaurant? (2) Have you ever eaten at the Four Seasons restaurant? (3) What fast-food restaurants have you visited in the last month? (3) For each of these restaurants, how many bottles of Coke did you buy? For each of the respondents who ever ate at the Ritz-Carleton the total number of bottles of Coke purchased are calculated for each fast-food restaurant. Similarly, for each of the respondents who ever ate at the Four Seasons the total number of bottles of Coke purchased are calculated for each fast-food restaurant. (Of course, there are respondents who ate at both the Ritz-Carleton and the Four Seasons.) From these one calculates, for example, the percentage of Coke bottles purchased at McDonald's by respondents who ate at the Ritz-Carleton and who ate at the Four Seasons. The statistical question is whether the percentage of Coke purchased at McDonald's is significantly different for Ritz-Carleton and Four Seasons patrons.

Now consider a third example. A random sample of respondents is asked two questions, (1) What fast-food restaurants have you visited in the last month? (2) For each of these restaurants, how many bottles of each of a list of soft drinks did you buy? The total number of bottles of soft drinks, as well as the total number of bottles of each soft drink, are calculated for each fast-food restaurant. From these one calculates, for example, the percentage of Coke bottles purchased at McDonald's and the percentage of Coke bottles purchased at Burger King. The statistical question is whether the consumption percentage of Coke in McDonald's is significantly different from the consumption percentage of Coke in Burger King.

These three examples are qualitatively different, in that in the first example the data about brand A are not independent of the data about brand B, as they were based on the same respondent (and in addition, since the percentages have to sum to 1, the higher the brand A percentage the lower will be the brand B percentage). In the second example, a subset of the respondents will have been patrons of both the Ritz-Carleton and Four Seasons restaurants, and so their Coke consumption in McDonald's is identical! In the third example, for those respondents who frequented both McDonald's and Burger King the Coke consumptions in the two fast-food restaurants are correlated.

What makes them similar, though, is that what drives these percentages is the underlying volumetric data that forms the basis for the percentage calculation. It is the distribution of these data that determine the distribution of the percentages, and, in turn, the proper method of testing these percentages for significant differences.

EXAMPLE 1: AVERAGES OF MULTINOMIAL PERCENTAGES

Suppose a respondent is asked to determine his percentage allocation across p products (e.g., what fraction of his dollar expenditure in a given category does he spend on each product in the category?) A sample of n respondents is drawn and the average percentage is calculated for each product. One now wants to know if the percentage for product 1 is significantly different from the percentage for product 2. We know first of all that, for each respondent, the p percentages are correlated, because they are required to sum to 1. If p_{1i} is the percentage for product 1 and p_{2i} is the percentage for product 2 for the i-th respondent, then the variance of p_{1i}-p_{2i} is estimated by $v_i=p_{1i}(1-p_{1i}) + p_{2i}(1-p_{2i}) +2\ p_{1i}\ p_{2i} = p_{1i} + p_{2i} - (p_{1i} - p_{2i})^2$. The variance of the average of the difference of these proportions is therefore estimated by

$$v = \frac{\sum_{i=1}^{n} v_i}{n^2} = \frac{\sum_{i=1}^{n} [p_{1i} + p_{2i} - (p_{1i} - p_{2i})^2]}{n^2}$$

Now suppose that we have n_1 observations only on product 1, n_2 observations only on product 2, and n observations on the pair of products. In this case the average for product 1 is

$$\frac{\sum_{i=1}^{n} p_{1i} + \sum_{i=n+1}^{n+n_1} p_{1i}}{n+n_1}$$

and the average for product 2 is

$$\frac{\sum_{i=1}^{n} p_{2i} + \sum_{i=n+1}^{n+n_{21}} p_{2i}}{n+n_2}$$

The variance of the difference of these two averages is

$$\frac{\sum_{i=1}^{n+n_1} p_{1i}(1-p_{1i})}{(n+n_1)^2} + \frac{\sum_{i=1}^{n+n_2} p_{2i}(1-p_{2i})}{(n+n_2)^2} + \frac{2\sum_{i=1}^{n} p_{1i}\,p_{2i}}{(n+n_1)(n+n_2)}$$

EXAMPLE 2: DEPENDENT PAIRED/OVERLAP (MULTI)

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 in a week by people who have ever eaten at the Ritz-Carleton, column

2 records the number of bottles of Coke consumed in a week by people who have ever eaten at the Four Seasons, the total row contains the total consumption for Coke in the respective columns, and row 1 contains the consumption for Coke of those respondents that had a specific attribute (e.g., Cokes purchased by respondents whose age was between 18 and 35). The percentages in question here are the percentage that Coke purchases make up of the total volume of soft drinks purchased among those 18-35 year old respondents who have ever eaten at the Ritz-Carleton restaurant and that for those who have ever eaten at the Four Seasons restaurant. The possible paired/overlap situation is that there are respondents who have eaten at both restaurants. Here is what such a table would look like:

	Volume of soft drinks purchased by Respondents aged 18 to 35 who have ever eaten at		
	Ritz-Carleton	4 Seasons	Hyatt
Total	51616	53184	56281
	100.0%	100.0%	100.0%
Coke	5427	5642	6032
	10.5%	10.6%	10.7%
Pepsi	5459	5058	5292
	10.6%	9.5%	9.4%
Seven Up	5113	5664	5566
	9.9%	10.6%	9.9%
Sprite	5307	5555	6126
	10.3%	10.4%	10.9%
Fanta	4535	4638	4897
	8.8%	8.7%	8.7%
Dr. Pepper	5143	5368	5745
	10.0%	10.1%	10.2%
Diet Pepsi	5063	5213	5521
	9.8%	9.8%	9.8%
Diet Coke	5472	5721	6066
	10.6%	10.8%	10.8%
Dr. Brown Cherry	4783	4820	5129
	9.3%	9.1%	9.1%
Dr. Brown Cel Ray	5314	5505	5907
	10.3%	10.4%	10.5%

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 (e.g., are between 18 and 35 and have eaten at both the Ritz-Carleton and Four Seasons restaurants), the next m respondents provide data only for column 1 (e.g., are between 18 and 35

and have only eaten at the Ritz-Carleton), and the last p respondents provide data only for column 2 (e.g., are between 18 and 35 and have only eaten at the Four Seasons). (There may be still other respondents that provided data on some, if not all, of the other banner items, but not on items 1 or 2. These will be disregarded in this analysis.)

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

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

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

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

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

Let X_1 be the 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_2 be the 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

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

The difference of the two percentages is given by

$$d = p_1 - p_2 = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} y_i}{X_1} - \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+m+1}^{n+m+p} z_i}{X_2}$$

$$= (\frac{1}{X_1} - \frac{1}{X_2})n\bar{x} + (\frac{1}{X_1})m\bar{y} - (\frac{1}{X_2})p\bar{z}$$

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

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

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

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

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

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

EXAMPLE 3: DEPENDENT PAIRED/OVERLAP (LOC+)

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 each of a number of soft drinks consumed in a week by people who ate at McDonald's, column 2 records the number of bottles of each of a number of soft drinks consumed in a week by people who ate at Burger King, the total row contains the total consumption of soft drinks in the respective columns, and row 1 contains the consumption for Coke in each of the two restaurants,. The percentages in question here are the percentage of the total McDonald's soft drink consumption that is attributable of Coke and the percentage of the total Burger King soft drink consumption that is attributable of Coke. The possible paired/overlap situation is that there are respondents who purchased soft drinks (not necessarily Coke) at both restaurants.

	Volume of soft drinks purchased at McDonald's	Burger King	Al's
Total	70090	49366	60373
	100.0%	100.0%	100.0%
Coke	7595	4824	6238
	10.8%	9.8%	10.3%
Pepsi	6528	5071	5743
	9.3%	10.3%	9.5%
Seven Up	6874	4992	6028
	9.8%	10.1%	10.0%

Sprite	7330	5224	7088
	10.5%	10.6%	11.7%
Fanta	6091	4846	6912
	8.7%	9.8%	11.4%
Dr. Pepper	7113	4829	6409
	10.1%	9.8%	10.6%
Diet Pepsi	7169	4691	6158
	10.2%	9.5%	10.2%
Diet Coke	7707	4770	5529
	11.0%	9.7%	9.2%
Dr. Brown Cherry	6404	5052	5617
	9.1%	10.2%	9.3%
Dr. Brown CelRay	7279	5067	4651
	10.4%	10.3%	7.7%

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

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

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

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

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

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

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

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

The difference of the two percentages is given by

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

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

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

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

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

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

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

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

EXAMPLE 2: WEIGHTED DEPENDENT PAIRED/OVERLAP (MULTI)

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

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

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

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

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

Let X_1 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_2 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

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

The difference of the two percentages is given by

$$d = p_1 - p_2 = \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{n+m} w_i y_i}{X_1} - \frac{\sum_{i=1}^{n} w_i x_i + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{X_2}$$

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

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

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

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

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

EXAMPLE 3: WEIGHTED DEPENDENT PAIRED/OVERLAP (LOC+)

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

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

$$X_1^+=\sum_{i=1}^{n}w_ix_{1i}+\sum_{i=n+1}^{n+m}w_iy_i.$$

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

$$X_2^+=\sum_{i=1}^{n}w_ix_{2i}+\sum_{i=n+m+1}^{n+m+p}w_iz_i$$

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

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

The difference of the two percentages is given by

$$d = p_1 - p_2 = \frac{\sum_{i=1}^{n} w_i x_{1i} + \sum_{i=n+1}^{n+m} w_i y_i}{X_1} - \frac{\sum_{i=1}^{n} w_i x_{2i} + \sum_{i=n+m+1}^{n+m+p} w_i z_i}{X_2}$$

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

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

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

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

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

$$+ \frac{\sum_{i=n+m+1}^{n+m+p}(z_i-\bar{z})^2}{(p-1)X_2^2}\sum_{i=n+m+1}^{n+m+p} w_i^2$$

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

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.

EXAMPLE 2: COMPARISON WITH TOTAL (MULTI) UNWEIGHTED & WEIGHTED

Let us denote by x_i the observed measurement for column 1 for respondent i (i = 1, 2, …, n), y_i the observed measurement for respondent i (i = n+1, n+2, …, n+m).

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

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

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

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

Let X_1 be the total of the measurements for those responding to column 1 across all attributes (e.g., the total Coke consumption respondents of all ages who ever ate at the Ritz-Carleton) and X_T be the weighted total of the measurements for all respondents across all attributes (e.g., the total Coke consumption respondents of all ages). Then the percentages under consideration are

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

The difference of the two percentages is given by

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

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

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

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

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

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

If the differences are weighted, then

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

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

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

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

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

EXAMPLE 3: COMPARISON WITH TOTAL (LOC+) UNWEIGHTED & WEIGHTED

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. Define δ_{ji} as 1 if respondent i answered item j and as 0 if respondent i did not answer item j, for i = 1, 2, …, n and j=1, 2, …, c. Let us denote by $x_{ji}\delta_{ji}$ the observed measurement for column j for respondent i. (As you can see, the δ_{ji} are used to keep track of the "no answers" in the data.) The total of the measurements for that attribute for those responding to column 1 is given by

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

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

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

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

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

The difference of the two percentages is given by

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

$$= \left(\frac{1}{X_1} - \frac{1}{X_T}\right)\sum_{i=1}^{n_1} x_{1i}\delta_{1i} - \left(\frac{1}{X_T}\right)\sum_{j=2}^{c}\sum_{i=1}^{n} x_{ji}\delta_{ji}$$

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

$$\left(\frac{1}{X_1} - \frac{1}{X_T}\right)^2 \sigma_1^2 \sum_{i=1}^{n}\delta_{1i} + \left(\frac{1}{X_T}\right)^2 \sum_{j=2}^{c}\sigma_j^2 \sum_{i=1}^{n}\delta_{ji} - 2\left(\frac{1}{X_T}\right)\left(\frac{1}{X_1} - \frac{1}{X_T}\right)\sigma_1 \sum_{j=2}^{c}\rho_{1j}\sigma_j \sum_{i=1}^{n}\delta_{1i}\delta_{ji}$$

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

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

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

where

$$n_j=\sum_{i=1}^{n}\delta_{ji}$$

and

$$n_{1j}=\sum_{i=1}^{n}\delta_{1i}\delta_{ji}$$

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

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

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

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

The difference of the two percentages is given by

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

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

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

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

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

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