Z-TESTS - DEPENDENT PAIRED/OVERLAP (MULTI)
A PRECAUTIONARY NOTE

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. We also deal in this note only with unweighted data.

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

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

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

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

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

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

The difference of the two proportions is given by

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

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

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

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

p=m

This is the situation in which the number who drank only Coke (m) is equal to the number who drank only Pepsi (p), and where n is the number who drank both Coke and Pepsi. Note that when p=m the first term of the difference of the two proportions is 0 and

$$q_1 - q_2 = (\frac{1}{n+m})mq_y - (\frac{1}{n+p})pq_z$$

so that the difference in the proportions depends only on the proportions in those who were positive only on item 1 and those who were positive only on item 2.

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

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

which is very small if n, the number who were positive on both items 1 and 2, is large. One would thus tend to find significant differences in the two proportions even when q_y and q_z are not very different, just because n is very large.

m=0

This is the situation in which the number everyone who drank only Coke also drank Pepsi (so m=0), and where p is the number who drank only Pepsi and n is the number who drank both Coke and Pepsi. Note that when m=0

$$q_1 - q_2 = (\frac{p}{n+p})q_x - (\frac{1}{n+p})pq_z = (\frac{p}{n+p})(q_x - q_z)$$

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

$$s_d^2 = (\frac{p}{(n+p)})^2[\frac{q_x(1-q_x)}{n} + \frac{q_z(1-q_z)}{p}]$$

When p=1, q_z will be either 0 or 1, and so q_1-q_2 will either be

$$q_1 - q_2 = (\frac{1}{n+1})q_x$$

or

$$q_1 - q_2 = (\frac{1}{n+1})(q_x - 1)$$

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

$$s_d^2 = \frac{1}{(n+1)^2} \frac{q_x(1-q_x)}{n}$$

Note that this is the estimated variance of q_x multiplied by $1/(n+1)^2$, which will make the estimated variance so much smaller as to render almost any difference between q_1 and q_2 significant.