A NOTE ON SPURIOUS SIGNIFICANCE

A. COMPARISONS WITH OVERLAPPING SAMPLES

1. Comparison of means

Following is the variance of the difference in means, where there are n observations on both products 1 and 2, with m additional observations on product 1 only and p additional observations on product 2 only.

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

Let's for simplicity assume that the variances are all equal. Then the variance reduces to

$$\sigma^2[(\frac{1}{n+m}-\frac{1}{n+p})^2 n.+(\frac{1}{n+m})^2 m+(\frac{1}{n+p})^2 p]$$

Let a=m/n and b=p/n, the respective ratios of the unique group sample sizes to the common group sample size. Then this variance becomes

$$\sigma^2[(\frac{1}{n(1+a)}-\frac{1}{n(1+b)})^2 n+(\frac{1}{n(1+a)})^2 na+(\frac{1}{n(1+b)})^2 nb]$$

$$=\frac{\sigma^2}{n}[\frac{(b-a)^2}{(1+a)(1+b)}+\frac{a}{(1+a)^2}+\frac{b}{(1+b)^2}]$$

$$=\frac{\sigma^2}{n}[\frac{a+b}{(1+a)(1+b)}]$$

As a and b both approach 0 (i.e., as there are hardly any unique observations, relative to the number of overlap observations) this variance goes to 0. Thus any t test with the square root of this variance in the denominator will get very large and so will register that the difference between the means is significant. This is just as an artifact of the implications of the small values of m and p rather than on the significance of the difference between the two means. We therefore caution users of this test to only use it if the multiplier of σ^2/n is greater than 0.05.

2. Comparison of proportions

Following is the variance of the difference of proportions, as given in the Statistical Reference manual and rewritten using the notation from the above section for comparison of means:

$$s_d^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - \frac{2n_0c}{n_1n_2} = \frac{s_1^2}{n+m} + \frac{s_2^2}{n+p} - \frac{2nc}{(n+m)(n+p)}$$

Let's assume that there is no significant difference in the population proportions (so they have a common value P) and that the samples mimic the populations, i.e. $p_1=p_{10}=p_2=p_{20}=P$. Then

$$s_d^2 = \frac{P(1-P)}{n+m} + \frac{P(1-P)}{n+p} - \frac{2n(p_{110}-P^2)}{(n+m)(n+p)}$$

Finally, let's assume that $p_{110}=P$, and that, as above, let a=m/n and b=p/n. Then

$$s_d^2 = \frac{P(1-P)}{n}[\frac{1}{1+a} + \frac{1}{1+b} - \frac{2}{(1+a)(1+b)}]$$
$$= \frac{P(1-P)}{n}[\frac{a+b}{(1+a)(1+b)}]$$

Once again as a and b approach 0 this variance goes to 0. Thus any z test with the square root of this variance in the denominator will get very large and so will register that the difference between the means is significant. This is just as an artifact of the implications of the small values of m and p rather than on the significance of the difference between the two proportions. We therefore caution users of this test to only use it if the multiplier of P(1-P)/n is greater than 0.05.

Conclusion

Given the possibility of spuriously finding "significant" differences due only because of the degree of overlap of the two samples, WinCross has adopted the safeguard of declaring all such differences not significant if the (a+b)/[(1+a)(1+b)] as defined above is less than 5%.

B. PART-WHOLE COMPARISONS

1. Comparison of means

The variance of the difference between the part and whole mean is

$$V[\bar{x}_1 - \bar{x}_T] = (1-\frac{m}{n})^2[\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n-m}]$$

where σ_1^2 is the variance of population 1 and σ_2^2 is the variance of population 2.

That is, if p=m/n is the fraction of the n observations that come from population 1, then the variance of $\bar{x}_1 - \bar{x}_T$ is given by $(1-p)^2$ times the sum of squares of the standard errors of population 1 and population 2. When the variances of the two populations are identical, the common variance may be estimated by s_T^2 and so the t statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_T}{s_T(1-p)\sqrt{\frac{1}{m} + \frac{1}{n-m}}}$$

The denominator of t can be rewritten as

$$s_T(1-p)\sqrt{\frac{1}{np} + \frac{1}{n(1-p)}} = \frac{s_T}{\sqrt{n}}\sqrt{\frac{1-p}{p}}$$

As p gets close to 1 the denominator of t approaches 0, so that t approaches infinity. In this case the test will find significant differences whenever the "part" is almost equal to the "whole." As a precaution, I recommend not performing this test when p > 0.95. As p gets close to 0 the denominator of t approaches infinity, so that t approaches 0. In this case the test will find no significant differences whenever the "part" is an infinitesimal part of the "whole." As a precaution, I recommend not performing this test when p < 0.05.

2. Comparison of proportions

The variance of the difference between the part and whole proportions is

$$\frac{(n-m)^2 p_1(1-p_1)}{mn^2} + \frac{(n-m)p_{\sim 1}(1-p_{\sim 1})}{n^2}$$

where p_1 is the proportion in the sample from population 1 and $p_{\sim 1}$is the proportion in the sample from population 2, the complementary part of the sample. Suppose these two proportions are equal to Q. Let p=m/n be the fraction of the n observations that come from population 1. Then this variance can be rewritten as

$$Q(1-Q)[\frac{(n-m)^2}{mn^2} + \frac{(n-m)}{n^2}$$

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

$$= \frac{Q(1-Q)}{n}\frac{(1-p)}{p}$$

As p gets close to 1 the denominator of z approaches 0, so that z approaches infinity. In this case the test will find significant differences whenever the "part" is almost equal to the "whole." As a precaution, I recommend not performing this test when p > 0.95. As p gets close to 0 the denominator of z approaches infinity, so that z approaches 0. In this case the test will find no significant differences whenever the "part" is an infinitesimal part of the "whole." As a precaution, I recommend not performing this test when p < 0.05.

Conclusion

Given the possibility of spuriously finding “significant” differences due only because of the degree of overlap of the part to the whole, WinCross has adopted the safeguard of declaring all such differences not significant if the fraction of the part to the whole is less than 5% or greater than 95%.