An Analysis of WinCross, SPSS, and Mentor
Procedures for Estimating the
Variance of a Weighted Mean

Dr. Albert Madansky
Vice President, The Analytical Group, Inc.
and
H.G.B. Alexander Professor Emeritus of Business Administration
Graduate School of Business
University of Chicago

The creators of statistical processing software for the marketing research community have confronted them with a variety of approaches in dealing with significance testing relating to weighted sample means. Each of these approaches produces a different variance of the weighted sample mean, and thus a different test statistic. The purpose of this note is to explain their bases, compare their approaches, and make some recommendations.

1. <u>Terminology</u>

The formula for the weighted mean is

$$x^* = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

And so the variance of the weighted mean is

$$Var(x^*) = \frac{\sum_{i=1}^{n} w_i^2 Var(x_i)}{(\sum_{i=1}^{n} w_i)^2}.$$

If each of the x's has the same variance, σ^2, then this reduces to

$$Var(x^*) = \frac{\sigma^2 \sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2} = \frac{\sigma^2}{f},$$

where the "effective sample size" f is given by

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

2. Estimation of σ^2

a. WinCross

If each of the x's has the same expected value μ and variance σ^2, then the usual estimate of the variance, namely

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

where

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

is an unbiased estimate of σ^2. It is this estimate that is used by WinCross in computing the variance of the weighted mean x*, i.e., the WinCross estimate of the variance of the weighted mean is s^2/f .

b. SPSS

An alternative estimate of the variance σ^2, used by SPSS in its computations, is based on the weighted data, namely

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

It can be shown that this estimate is a biased estimate of σ^2, in that

$$E(s_w^2) = \sigma^2 \frac{(\sum_{i=1}^{n} w_i)^2 - \sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2 - \sum_{i=1}^{n} w_i} = \frac{\sigma^2}{g}$$

so that a proper unbiased estimate of σ^2 based on s_w^2 would be g s_w^2, where the unbiasing factor g is given by

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

SPSS does not perform this adjustment, but instead uses the biased estimate s_w^2. SPSS compounds the estimation problem by estimating the variance of the weighted mean by

$$\frac{s_w^2}{\sum_{i=1}^{n} w_i}$$

instead of by s_w^2/f. That is, instead of dividing s_w^2 by the "effective sample size" f it divides s_w^2 by the sum of the weights, the "weighted sample size,"

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

c. Mentor

First let us establish a glossary relating CfMC Mentor's notation to ours.

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

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

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

$$Y = \sum_{i=1}^{n} w_i^2$$

So E=F^2 /Y is the effective sample size, which we call f. The weighted mean is M=S/F, which we call x*. Mentor calculates an "adjusted sum of squares" A, via the following formula:

$$A = \frac{Z - S^2 / F}{Y / F}$$

$$= \frac{\sum_{i=1}^{n} w_i x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2 / \sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i^2 / \sum_{i=1}^{n} w_i}$$

This can be rewritten as

$$A = \frac{\sum_{i=1}^{n} w_i x_i^2 - (\sum_{i=1}^{n} w_i) x^{*2}}{\sum_{i=1}^{n} w_i^2 / \sum_{i=1}^{n} w_i}$$

$$= s_w^2 \frac{\sum_{i=1}^{n} w - 1}{\sum_{i=1}^{n} w_i^2 / \sum_{i=1}^{n} w_i}$$

so that the expected value of A is

$$
\begin{aligned}
E(A) &= E(s_w^2)\frac{\sum_{i=1}^{n} w - 1}{\sum_{i=1}^{n} w_i^2 \;/\; \sum_{i=1}^{n} w_i} \\
&= \frac{\sigma^2 \;/\; g}{\sum_{i=1}^{n} w_i^2 \;/\; \sum_{i=1}^{n} w_i} \\
&= \sigma^2 \left[\frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2} - 1\right] \\
&= \sigma^2 (f - 1)
\end{aligned}
$$

Mentor's estimate of σ^2 is given by V=A/(E-1), or

$$
s_c^2 = A\left[\frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2} - 1\right],
$$

and so we see that it is an unbiased estimate of σ^2.

Following is an algebraically simplified expression for s_c^2, Mentor's estimate of σ^2.

$$
s_c^2 = \frac{(\sum_{i=1}^{n} w_i)\sum_{i=1}^{n} w_i (x_i^2 - x^{*2})}{(\sum_{i=1}^{n} w_i)^2 - \sum_{i=1}^{n} w_i^2}.
$$

Mentor then estimates the variance of x* by s_c^2/f.

3. <u>Comparison</u>

a. <u>Variance of WinCross estimate of variance of x*</u>

Since both the WinCross estimate and the Mentor estimate of the variance of x* are unbiased, the way one must compare the two estimates is by determining which one of these estimates has the smaller variance. Since $(n-1)s^2/\sigma^2$ has a chi-square distribution with n-1 degrees of freedom, we know that the variance of $(n-1)s\ s^2/\sigma^2$ is 2(n-1), so that variance of the WinCross estimate of s, namely s^2, is $2\sigma^4$ /(n-1), and the variance of the WinCross estimate of the variance of x* is $2\sigma^4/f^2$ (n-1).

b. Variance of Mentor estimate of variance of x*

Since both WinCross and Mentor estimate the variance of x* by dividing their estimates of σ^2 by f, one need only compare the variance of s^2, the WinCross estimate of σ^2, with the variance of s_c^2, Mentor's estimate of σ^2. We first establish some notation. Let X be the n-vector of observations, E be the n-vector of 1's, and I be the identity matrix. Then s^2 can be expressed as

$$s^2 = aX'AX$$

where a=1/(n-1) and A = I - (1/n)EE´.

We can express s_c^2 as

$$s_c^2 = bX'BX$$

where, as above, c is the "weighted sample size,"

$$b = \frac{c}{c^2 - \sum_{i=1}^{n} w_i^2},$$

$B = D_w - (1/c)WW'$, W is the n-vector of weights, and D_w is a diagonal matrix with the weights w_i on the diagonal.

The symmetric matrices A and B can each be expressed as a product of orthogonal and diagonal matrices, where the orthogonal matrices are the matrices of eigenvectors of A and B and the diagonal matrices are matrices containing the eigenvalues of A and B. Let the decompositions of A and B be expressed as $A=Q_A'D_A Q_A$ and $B=Q_B'D_B Q_B$. Then

$$s^2 = aX'Q_A'D_AQ_AX = aY_A'D_AY_A$$

and

$$s_c^2 = bX'Q_B'D_BQ_BX = bY_B'D_BY_B$$

Since the covariance matrix of X is $\sigma^2 I$, and both Q_A and Q_B are orthogonal matrices, the covariance matrix of Y_A is $\sigma^2 Q_A'Q_A = \sigma^2 I$ and the covariance matrix of Y_B is $\sigma^2 Q_B'Q_B = \sigma^2 I$. Therefore s^2 and s_c^2 are expressible as a weighted sum of squares of independent variables with common variance σ^2, and where the weights are the eigenvalues of aD_A and bD_B, respectively. That is

$$s^2 = a\sum_{i=1}^{n} \lambda_{Ai} y_{Ai}^2$$

$$s_c^2 = b\sum_{i=1}^{n} \lambda_{Bi} y_{Bi}^2$$

and so, since y_{Ai}^2/σ^2 and y_{Bi}^2/σ^2 have chi-square distributions with 1 degree of freedom, so that $\mathrm{Var}(y_{Ai}^2)=\mathrm{Var}(y_{Bi}^2)=2\sigma^4$, we see that the variances of the two estimates are expressible in terms of the sum of squares of the eigenvalues of aD_A and bD_B, namely

$$Var(s^2) = 2\sigma^4 a^2 \sum_{i=1}^{n} \lambda_{Ai}^2$$

$$Var(s_c^2) = 2\sigma^4 b^2 \sum_{i=1}^{n} \lambda_{Bi}^2$$

It remains to determine these eigenvalues.

All but one of the eigenvalues of A are equal to 1, with the n-th eigenvalue equal to 0 (see S.N. Roy, B.G. Greenberg, and A.E. Sarhan “Evaluation of Determinants, Characteristic Equations and their Roots for a Class of Patterned Matrices” *Journal of the Royal Statistical Society*. *Series B (Methodological)*, Vol. 22, No. 2. (1960), pp. 348-359).. Thus the sum of the eigenvalues of A is n-1, and so, since $a^2 = 1/(n-1)^2$, we see that $Var(s^2) = 2\sigma^4/(n-1)$, as demonstrated earlier using a nonmatricial derivation.

We need not determine the eigenvalues of B to calculate their sum of squares, for $B^2 = Q_B' D_B Q_B Q_B' D_B Q_B = Q_B' D_B D_B Q_B$, and so the sum of squares of the eigenvalues of B is equal to the sum of eigenvalues of B^2 . But the sum of eigenvalues of a symmetric matrix is equal to the trace of that matrix, i.e., the sum of its diagonal terms. So we need only look at the diagonal terms of B^2 to obtain this required quantity.

Since $B = D_w - (1/c)WW'$, $B^2 = D_W^2 - (1/c)D_W WW' - (1/c) WW'D_W + (1/c)^2 WW'WW'$, and

$$trB^2 = \sum_{i=1}^{n} w_i^2 - 2\frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i} + \frac{(\sum_{i=1}^{n} w_i^2)^2}{(\sum_{i=1}^{n} w_i)^2}$$

and so

$$Var(s_c^2) = 2\sigma^4 b^2 [\sum_{i=1}^{n} w_i^2 - 2\frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i} + \frac{(\sum_{i=1}^{n} w_i^2)^2}{(\sum_{i=1}^{n} w_i)^2}]$$

$$= 2\sigma^4 \frac{[\sum_{i=1}^{n} w_i]^2}{\{(\sum_{i=1}^{n} w_i)^2 - \sum_{i=1}^{n} w_i^2]^2} \frac{\sum_{i=1}^{n} w_i^2 (\sum_{i=1}^{n} w_i)^2 - 2\sum_{i=1}^{n} w_i^3 \sum_{i=1}^{n} w_i + (\sum_{i=1}^{n} w_i^2)^2}{[\sum_{i=1}^{n} w_i]^2}$$

$$= 2\sigma^4 \frac{\sum_{i=1}^{n} w_i^2 (\sum_{i=1}^{n} w_i)^2 - 2\sum_{i=1}^{n} w_i^3 \sum_{i=1}^{n} w_i + (\sum_{i=1}^{n} w_i^2)^2}{(\sum_{i=1}^{n} w_i)^4 - 2\sum_{i=1}^{n} w_i^2 (\sum_{i=1}^{n} w_i)^2 + (\sum_{i=1}^{n} w_i^2)^2}$$

c. <u>Comparison</u>

Before proceeding with a proof that Var(s^2) $\le$ Var(s_c^2), I will illustrate these computations with an example. I selected as weights 100 random numbers from a uniform distribution between 0 and 1. These weights, along with their squares and cubes, are given in Appendix I. The variance of s^2, excluding the factor $2\sigma^4$, is 1/99= 0.010101. The various sums needed to compute the variance of s_c^2 are

$$\sum_{i=1}^{n} w_i = 45.040576$$

$$\sum_{i=1}^{n} w_i^2 = 29.631266$$

$$\sum_{i=1}^{n} w_i^3 = 22.913609$$

The variance of s_c^2, again excluding the factor $2\sigma^4$, is calculated as

$$\frac{29.631266\ x\ (45.0405756)^2 - 2\ x\ 22.913609\ x\ 45.040576 + (29.631266)^2}{(45.0405756)^4 - 2\ x\ 29.631266\ x\ (45.040576)^2 + (29.631266)^2} = 0.014756$$

Thus in this example use of s_c^2 will produce an estimate of the variance of x* with 1.46 times the variance compared with the use of s_c^2.

Now let us compare Var(s^2) with Var(s_c^2). One can simplify the expression for Var(s_c^2) by assuming that the weights sum to 1. This merely rescales the weights and will have no impact on the computation of Var(s_c^2). Then Var(s_c^2) reduces to

$$Var(s_c^2) = 2\sigma^4 \frac{\sum_{i=1}^{n} w_i^2 - 2\sum_{i=1}^{n} w_i^3 + (\sum_{i=1}^{n} w_i^2)^2}{1 - 2\sum_{i=1}^{n} w_i^2 + (\sum_{i=1}^{n} w_i^2)^2}$$

Note that when the w's are all equal to 1/n, then

$$Var(s_c^2) = 2\sigma^4 \frac{n/n^2 - 2n/n^3 + (n/n^2)^2}{1 - 2n/n^2 + (n/n^2)^2} = \frac{2\sigma^4}{n-1}$$

which is the same as Var(s^2) in that case.

Let us now determine what are the values of the w's that minimize Var(s_c^2) subject to the constraint that the sum of the w's is equal to 1. To do this we form the Lagrangean

$$L = \log(2\sigma^4) + \log[\sum_{i=1}^{n} w_i^2 - 2\sum_{i=1}^{n} w_i^3 + (\sum_{i=1}^{n} w_i^2)^2] - \log[1 - 2\sum_{i=1}^{n} w_i^2 + (\sum_{i=1}^{n} w_i^2)^2] - \xi(\sum_{i=1}^{n} w_i - 1),$$

set the derivative of L with respect to each of the w_i equal to 0, and solve for the minimizing values of the w_i and ξ, the Lagrange multiplier. The result of this is the set of n equations

$$\frac{\partial L}{\partial w_i} = \frac{2w_i(1+2\sum_{j=1}^{n} w_j^2) - 6w_i^2}{\sum_{j=1}^{n} w_j^2 - 2\sum_{j=1}^{n} w_j^3 + (\sum_{j=1}^{n} w_j^2)^2} + \frac{4w_i(1-\sum_{j=1}^{n} w_j^2)}{1-2\sum_{j=1}^{n} w_j^2 + (\sum_{j=1}^{n} w_j^2)^2} = \xi$$

The only way for this equation to hold for each of the w_i is when all of the w_i are equal, i.e., when $s_c^2 = s^2$. Otherwise Var(s_c^2) will be greater than Var(s^2).

4. Conclusion

Given both the bias in the SPSS estimate of σ^2 and its incorrect denominator in determining the standard error of x*, the probabilities calculated based on the t-statistic will be incorrect. The probabilities based on both the WinCross and Mentor statistics will be correct, but, because Mentor uses an estimate of the variance of x* with a larger variance than that of the estimate used by WinCross, it is more likely that one will find fewer significant differences using the Mentor procedure than using the WinCross procedure.

APPENDIX I

	w	w^2	w^3
1	0.995127	0.990278	0.98545212
2	0.991954	0.983973	0.97605569
3	0.989075	0.978269	0.96758176
4	0.982972	0.966234	0.94978092
5	0.971904	0.944597	0.91805798
6	0.968704	0.938387	0.90901967
7	0.965210	0.931630	0.89921892
8	0.954774	0.911593	0.87036567
9	0.952251	0.906782	0.86348403
10	0.941401	0.886236	0.83430331
11	0.938380	0.880557	0.82629710
12	0.919475	0.845434	0.77735568
13	0.917015	0.840917	0.77113305
14	0.888908	0.790157	0.70237726
15	0.882978	0.779650	0.68841393
16	0.853234	0.728008	0.62116140
17	0.837742	0.701812	0.58793710
18	0.823839	0.678711	0.55914834
19	0.817090	0.667636	0.54551875
20	0.810228	0.656469	0.53188990
21	0.805057	0.648117	0.52177095
22	0.793969	0.630387	0.50050756
23	0.782669	0.612571	0.47944015
24	0.781462	0.610683	0.47722545
25	0.736144	0.541908	0.39892231
26	0.722549	0.522077	0.37722626
27	0.718437	0.516152	0.37082250
28	0.693553	0.481016	0.33360993
29	0.663519	0.440257	0.29211919
30	0.648944	0.421128	0.27328869
31	0.610076	0.372193	0.22706585
32	0.578844	0.335060	0.19394769
33	0.575269	0.330934	0.19037631
34	0.571105	0.326161	0.18627213
35	0.538395	0.289869	0.15606412
36	0.537269	0.288658	0.15508698
37	0.523968	0.274542	0.14385147
38	0.521198	0.271647	0.14158206
39	0.507969	0.258033	0.13107251

40	0.471876	0.222667	0.10507119
41	0.462365	0.213781	0.09884503
42	0.456192	0.208111	0.09493864
43	0.445603	0.198562	0.08847984
44	0.441056	0.194530	0.08579880
45	0.437094	0.191051	0.08350732
46	0.422376	0.178401	0.07535251
47	0.421953	0.178044	0.07512634
48	0.417159	0.174022	0.07259469
49	0.405299	0.164267	0.06657736
50	0.392635	0.154162	0.06052949
51	0.387894	0.150462	0.05836321
52	0.383761	0.147273	0.05651744
53	0.377796	0.142730	0.05392275
54	0.371654	0.138127	0.05133534
55	0.357678	0.127934	0.04575902
56	0.341958	0.116935	0.03998695
57	0.306573	0.093987	0.02881388
58	0.305468	0.093311	0.02850343
59	0.296491	0.087907	0.02606361
60	0.289246	0.083663	0.02419926
61	0.283096	0.080143	0.02268826
62	0.280116	0.078465	0.02197929
63	0.269943	0.072869	0.01967054
64	0.266302	0.070917	0.01888527
65	0.265191	0.070326	0.01864989
66	0.257537	0.066325	0.01708122
67	0.249131	0.062066	0.01546263
68	0.233802	0.054663	0.01278041
69	0.231034	0.053377	0.01233183
70	0.227916	0.051946	0.01183926
71	0.207306	0.042976	0.00890914
72	0.206597	0.042682	0.00881804
73	0.192060	0.036887	0.00708453
74	0.190022	0.036108	0.00686138
75	0.188724	0.035617	0.00672174
76	0.184651	0.034096	0.00629586
77	0.180900	0.032725	0.00591992
78	0.179789	0.032324	0.00581151
79	0.169282	0.028656	0.00485101
80	0.155006	0.024027	0.00372431

81	0.151594	0.022981	0.00348374
82	0.149657	0.022397	0.00335190
83	0.146832	0.021560	0.00316564
84	0.123566	0.015269	0.00188667
85	0.121520	0.014767	0.00179450
86	0.117982	0.013920	0.00164228
87	0.111100	0.012343	0.00137133
88	0.109864	0.012070	0.00132607
89	0.101032	0.010207	0.00103128
90	0.094285	0.008890	0.00083816
91	0.091617	0.008394	0.00076900
92	0.088234	0.007785	0.00068692
93	0.067813	0.004599	0.00031185
94	0.063359	0.004014	0.00025435
95	0.050390	0.002539	0.00012795
96	0.034176	0.001168	0.00003992
97	0.029486	0.000869	0.00002564
98	0.029208	0.000853	0.00002492
99	0.026666	0.000711	0.00001896
100	0.009006	0.000081	0.00000073