TYPE I ERROR RATES FOR THE UNEQUAL VARIANCE T TEST

J. Philip Miller, Reimut Wette and Robert P. Parks
Washington University

The Behrens-Fisher problem of comparing two sample means from normal populations when the variances cannot be assumed to be equal is a statistical problem with a long and controversial history. It is a problem in which Fisher's fiducial inference theory and Neyman-Pearson's confidence interval theory produce clear differences. (Kendall & Stuart, 1973, pp. 146-157). It is one of the classic problems of statistical inference, for on the surface it is a simple problem, yet no similar region exists.

The present report is not, however, concerned with the theoretical aspects but rather addresses the problem on a more applied level appropriate for those engaged in data analysis activities. SAS PROC TTEST, as practically all other statistical packages, routinely prints results for an "unequal variance t test". The current research was originally directed towards understanding the performance of that test when it was used with the F test on the equality of the variances as a pretest estimator. That procedure is explicitly prescribed by the SPSS manual (Nie, et al 1975, p. 270) and appears to be the one followed by many applied statisticians. It is characterized by performing an F test for equality of the two variances and applying the unequal variance t test if the F is significant at some level α', say; otherwise applying the regular t test which assumes the equality of the two variances. This procedure is usually justified by a belief that the unequal variance t test is conservative when applied to samples which actually have identical population variances.

As we began to perform some simulations of the procedure, we discovered that, although that contention was true for samples of equal size, it was not true for samples of unequal size. This called for an examination of the test performed by SAS and the search for a test with superior performance, if one existed.

For the unequal variance t test SAS computes the statistic

$$t^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} , \qquad (1)$$

where s_1^2 and s_2^2 are the unbiased estimates of the variances in groups 1 and 2, $\bar{X}_1$ and $\bar{X}_2$ the sample means, and n_1 and n_2 the sample sizes. t^* would be distributed normally with expectation 0 and variance 1 if the actual population variances were used, or as Student's t if their ratio were known.

If the variances are estimated from each sample separately, then the distribution of t^* is not known. SAS implements an approximate degree of freedom (ADF) solution in which the first and second moments of the distribution are asymptotically equated to a Student's t whose degrees of freedom are given by:

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

This approach is variously attributed to Satterthwaite (1941) or Smith (1936) and is also implemented in BMDP, IMSL, SPSS and it is found in such texts as Bennett and Franklin (1954). It is perhaps easier to visualize when (2) is written as

$$\frac{1}{df} = \frac{1}{n_1-1}\left(\frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)^2 + \frac{1}{n_2-1}\left(\frac{\frac{s_2^2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)^2 . \qquad (3)$$

Generally, the number of degrees of freedom is not integer valued. SAS (79.5) does a linear interpolation in P between the bracketing integer df values. SPSS (7.2) utilizes FLOOR(df) for its P-value.

Since the understanding of the test is virtually impossible from analytic considerations, we have approached the problem via Monte Carlo techniques. Initial explorations were accomplished utilizing the random number generation facilities in SAS, PROC PRINTTO to save the output, and SAS to summarize the results. In order to provide more flexible analyses, however, the simulations which are reported here were accomplished with Fortran programs run on a Harris 125 computer and numerical routines were drawn from IMSL (ed. 7) wherever possible. The results reported here represent some 10 million t tests on over a million different samples.

We examined the case with identical population variances for each sample. Since we were interested in the empirical Type I error rates, we began with the case of identical population means. Specifically, the values within each sample were generated with means 0 and variances 1. The samples were generated with the IMSL routine GGNML. An experiment consisted of from 1,000 to 5,000 replications of the sample for a given set of values for n_1 and

n_2. P values were calculated for each sample. For each experiment the empirical two-sided rejection rates at nominal rejection levels of .01, .05, .10 and .20 were recorded. For the case of equal sample sizes these empirical Type I error rates are shown in Figure 1.

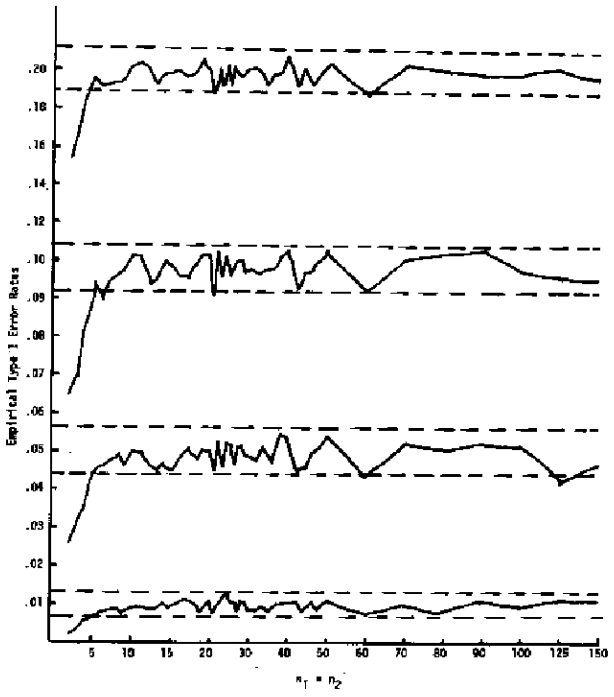

Figure 1

Empirical two-sided Type I error rates for nominal significance levels of .01, .05, .10 and .20. Rates are for the *SAS* method and are based on 5,000 replications. 95% confidence intervals (based on arc sine transformation) are shown with dashed lines.

These results demonstrated the basic conservatism assumed of the test. For all practical purposes the conservatism is negligible for $n_1=n_2$ over 10. Since as summary statistics these empirical Type I error rates are subject to large sampling variability, we also computed several other summary statistics which are somewhat more stable descriptors of the test's performance. For each sample generated, we computed the P associated with the t* observed using (2) for the df and SAS's interpolation algorithm. This P value was then compared with the correct P_e value which was that computed by the standard, equal variance t test by way of the log of the odds ratio, i.e.

$$LO = \ln \frac{\frac{P}{1-P}}{\frac{P_e}{1-P_e}} \quad . \qquad (4)$$

Positive values for LO indicate that the P for the ADF test is larger than that for the equal variance t test, that is, that the test is conservative. Negative values indicate P values which are too small, thus invalidating the nominal significance levels. The average LO value for an experiment was then computed as well as the standard deviation of the LO values. The latter provided an index of the consistency of the performance of the test relative to the equal variance t test.

Other ADF approaches have been proposed in texts and other computer packages. IMSL and BMDP both utilize the same formulas as SAS, only instead of interpolating for the fractional degrees of freedom, they compute the P-values by transforming t to a Beta-distributed variable, which is then defined for noninteger df. We shall refer to this approach as the Beta method. Dixon and Massey, in the first (1951) and second (1957) editions offer a different formula for the df, viz.,

$$df_{DM} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2+1}} - 2 \quad . \qquad (5)$$

The *DM* method is also offered in such texts as Anderson and Bancroft (1952) and Remington and Schork (1970). With no comment and no reference Dixon and Massey replaced formulas in their third edition (1969) to one given earlier by Bliss (1967), viz.,

$$df_{Bliss} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2}} \quad . \qquad (6)$$

Certainly all of these ADF methods are asymptotically equivalent as the degrees of freedom increase, since the exact value of the degrees of freedom becomes much less important in determining the significance level. These ADF methods, even when employing the Beta distribution, provide computationally rapid methods for the calculation of the appropriate P values.

Another approach, which is found in certain texts, is to utilize a critical t value which is a function of tabulated t values. One of the earliest of these is by Cochran (1964) and compares t_α^* with the observed t*, where t_α^* is given by

$$t_\alpha^* = C\, t_{n_1-1,\alpha} + (1-C)\, t_{n_2-1,\alpha} \quad , \qquad (7)$$

where

$$C = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad , \qquad (8)$$

and $t_{df,\alpha}$ is the critical value for a two-sided t distribution with *df* degrees of freedom at the α level of significance. While this method is simple to implement using a standard t table for a preselected α level, the computation of a P-value requires an iterative approximation to find the level for which the observed t* and C provide an equality. This weighting method, which is also found in Snedecor and Cochran (1967), Bliss (1967), and Sokal and Rohlf (1969), we shall refer to as the *C* method.

McCullough (1960) and Banerjee (1960) each independently proposed a similar weighting, but using a weighted mean square average, viz.,

$$t^*_\alpha = \sqrt{C\, t^2_{n_1-1,\alpha} + (1-C)\, t^2_{n_2-1,\alpha}} \quad . \qquad (9)$$

Pagurova (1968) has provided a solution which is a weighted combination of three t values, a t with n_1-1, a t with n_2-1, and a t with n_1+n_2-2 degrees of freedom. We shall refer to these two as the *MB* and the *P* methods.

Mickey and Brown (1966) have shown that the correct solution is bounded by two Student's t distributions, one with the df equal to the minimum of n_1-1 and n_2-1 and the other with $df=n_1+n_2-2$. We shall refer to these two as the *Low* and *High* methods.

The classical tabulated solutions to the problem are provided by Sukhatme (1942) in the Fisher and Yates (1957) tables and by Aspin and Welch (1949) in the Biometrika Tables Pearson and Hartley, 1966). The Aspin-Welch tables are provided by an expansion to order -4 in df. Aspin (1948) reports that, unfortunately, it took her over 100 pages to analytically go from order -3 to -4, so that additional terms in the expansion are not likely to be quickly forthcoming. The tabulations provided in Fisher and Yates are integrations of weighted convolutions of t distributions such that the weighted sum of t values is greater than the observed t*. Perhaps it is unfair to compare Fisher's fiducially motivated test with empirical Type I error rates. Presumably, such a comparison would not have been accepted as valid by Fisher. For each sample we computed the observed P value by each of the above eleven methods. These were then summarized over each experiment. For the equal sample size case we see in Figure 2 the empirical Type I error rates corresponding to a nominal alpha of .05.

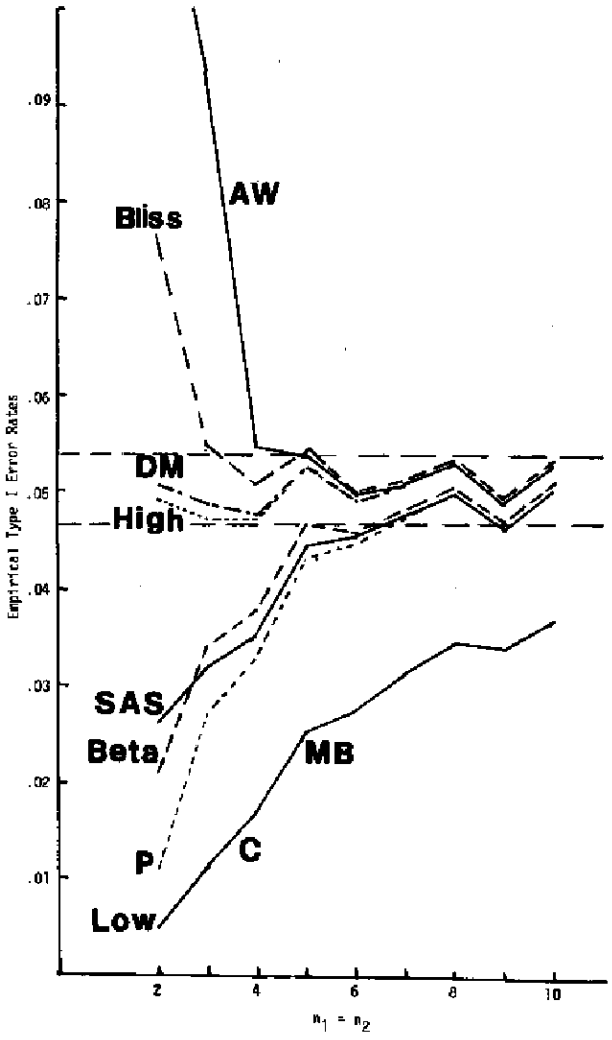

Figure 2

Empirical two-sided Type I error rates for a nominal significance level of .05 for various methods. The experiments ranged from $n_1=n_2=2$ to $n_1=n_2=10$. The dashed line represents 95% confidence intervals (based on arc sine transformation) for 15,000 replications.

For the equal sample size case the *C* and *MB* methods are equivalent to the lower bound, which in fact does provide the lower bound for all methods. The *SAS* and *Beta* methods are close and except for the $n_1=n_2=2$ case the *Beta* method is always closer to the correct .05 level. In fairness to the Aspin-Welch method it should be pointed out that the Biometrika Tables do not go below $n_1=n_2=6$, an area in which its equal sample size performance appears to be quite satisfactory. The upper bound is, for the equal sample size case, equivalent to the ordinary equal variance t test, which would be the correct test for the cases studied. Figure 3 demonstrates the average LO value for $n_1=n_2$ from

2 to 150 and demonstrates the rapid convergence to an average LO of 0.0.

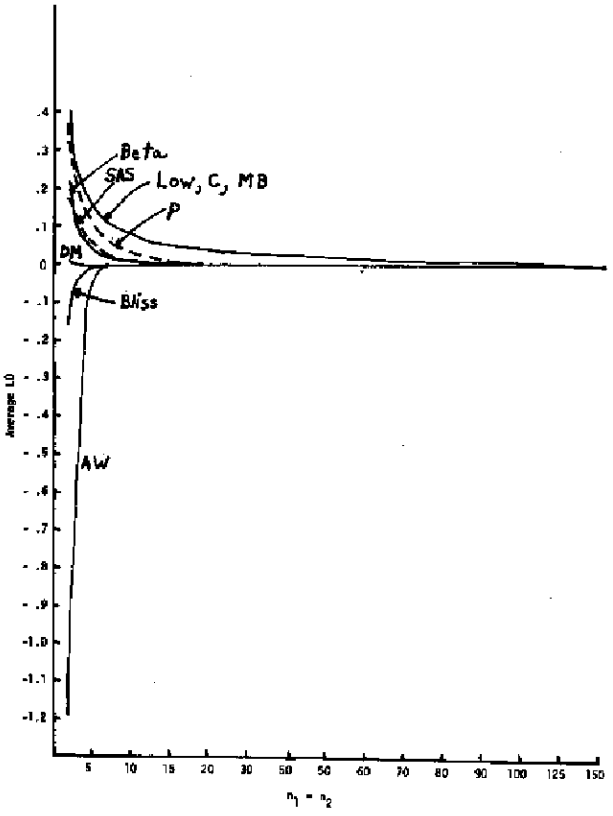

Figure 3

Average LO for each method when $n_1=n_2$ for various methods. Each point represents at least 5,000 replications.

When the sample sizes are unequal, however, the performance is not near so favorable, as shown in Figure 4. The sample size in group 1 was held constant and the size of group 2 varied from 2 to 750. The *SAS* method begins with the conservative performance we saw earlier for samples of $n_1=n_2=2$, but by $n_1=2$, $n_2=5$ it produced a negative average LO, indicating an empirical Type I error rate greater than the nominal level. For low sample sizes there were no differences between the *SAS* and *Beta* methods, but at larger n_2 values the average LO value for *SAS* was more negative. The *DM* procedure which appears to have desirable properties for the small equal sample size case (Figure 2), consistently performs worse than *SAS* or *Beta*. The *Bliss* method, which is the one that Dixon and Massey converted to, performs even worse. The *AW* method performs worse than the boundary condition until $n_1=2$ and $n_2=8$ and inferior to the *SAS* and *Beta* methods until about the $n_1=2$, $n_2=20$ case; it never seriously appears to overstate the P value. The mean square weighting (*MB*) method appears to not have any important advantage over the *C* method. Utilizing 3 Student's t distributions as in the *P* method also seems to be inferior with decidedly negative LO values at larger n.

At $n_1=n_2=2$ and $n_1=2$, $n_2=3$ the *BF* method is more conservative than the lower bound and does not appear to offer advantages over the computationally much more efficient weighted t methods.

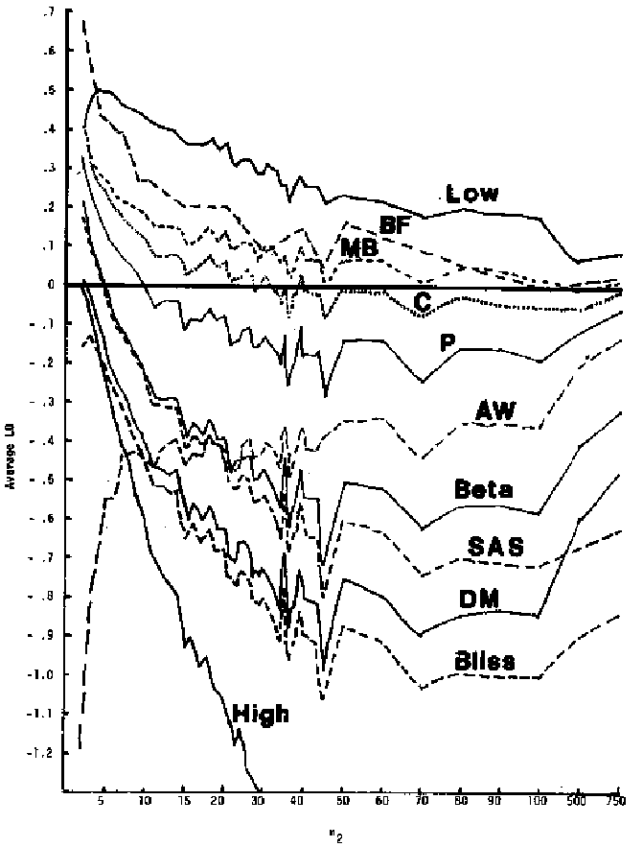

Figure 4

Average LO for each method when $n_1=2$ and n_2 ranges from 2 to 750. For each value of n_2 at least 5,000 replications were used.

In order to provide a better appreciation of the extent of these overstated alpha values, Table I shows the actual observed Type I error rates for nominal alpha levels of .01, .05, .10 and .20.

TABLE I
Empirical Type I Error Rates (Beta Method)
($n_1=2$, $n_2=50$)

	.01	.05	.10	.20	$\bar{P}$	$\overline{LO}$
ADF						
SAS	.075	.116	.150	.249	.471	- .60
Beta	.076	.117	.155	.230	.491	- .50
DM	.098	.151	.193	.270	.471	- .77
Bliss	.096	.157	.211	.296	.449	- .87
Weighted t						
C	.017	.053	.097	.185	.512	- .00
MB	.007	.037	.084	.176	.516	.07
P	.045	.076	.103	.169	.520	- .17
Other						
Low	.000	.001	.026	.132	.535	.24
High	.163	.242	.299	.375	.402	-1.58
AW	.056	.086	.115	.194	.500	- .36
BF	.005	.035	.074	.159	.530	.16

It is only their general conservatism that allow the *Low*, *MB* and *BF* methods not to overstate the P value. Similar results were observed for cases in which n_1 was allowed to be greater than 2 and n_2 ranged to several hundred. While the unbalanced sample sizes may not be the situations which the developers of these tests had in mind, this type of imbalance is quite frequent in studies in clinical medicine where, for example, the cases who die are compared to the survivors on a wide ensemble of measures. The unequal variance t test is then often utilized as a more parsimonious solution than transformations of the variables or non-parametric tests.

We now return to the original question about the use of the F test for the quality of the variances as a pretest estimator. Figure 5 shows the empirical Type I error rates corresponding to nominal test of .05.

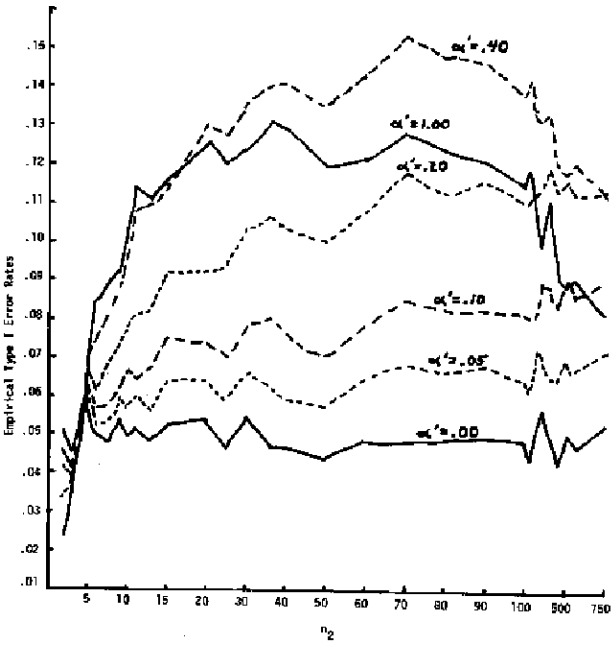

Figure 5

Empirical two-sided Type I error rates for a nominal significance level of .05 for various levels of significance for a F test of equality of the variance as a pretest estimator. Computations are for the *Beta* method with n_1=2 and various n_2 values. Each value of n_2 represents the results of 5,000 replications.

The cases studied here are again those where n_1=2 and the sizes of the samples in group 2 range from 2 to 750. The different lines represent different α' levels elected for the F test. α'=.00 stands for never using the unequal variance t test unless the F is significant at the .05 level. The t test significance levels are computed by the *Beta* method but similar results are noted for the other methods. Probably the most notable features of these results are that the performance of the test is worse when the F as a pretest estimator is set at an α' of .40, which is a recommended level popular in much of the econometric literature. There is even cross-over for the .20 and .10 levels on this graph. The user is thus faced with a dilemma: the larger the α' for the pretest, the worse the performance of the combined test when the population variances are actually equal, and the lower the level, the less power there is to detect true inequalities in the variances so that the correct test can be applied.

In summary, our experience with this simple case where the population variances are indeed equal and only the Type I error rates are considered, allow several conclusions: 1) SAS's computation of the P value should be based on the Beta distribution function rather than on the interpolation in P. 2) One cannot be confident that this test is always conservative. It may be a safe assumption for the equal sample size case, but not for unequal sample sizes cases, especially not for gross discrepancies in sample sizes. 3) The use of the F test for the equality of the variances as a pre-test estimator does not provide an acceptable way out of the problem.

References

Anderson, RL and TA Bankroft. Statistical Theory in Research, McGraw-Hill, New York, 1952.

Aspin, AA. An examination and further development of a formula arising in the problem of comparing two mean values, Biometrika, 35:88-96, 1948.

Aspin, AA. Tables for use in comparisons whose accuracy involves two variances, separately estimated, (Appendix by BL Welch), Biometrika, 36:290-6, 1949.

Banerjee, SK. Approximate confidence interval for linear functions of means of k-populations when the population variances are not equal, Sankhya, 22:357-8, 1960.

Bennett, CA and NL Franklin. Statistical Analysis in Chemistry and the Chemical Industry, Wiley and Sons, New York, 1954.

Bliss, CI. Statistics in Biology, Vol. I, McGraw-Hill, New York, 1967.

Cochran, WG. Approximate significance levels of the Behrens-Fisher test, Biometrics 20:191-5, 1964.

Dixon WF and FJ Massey. Introduction to Statistical Analysis, McGraw-Hill, New York, 1951.

Dixon, WF and FJ Massey. Introduction to Statistical Analysis, (2nd Edition), Mc-Graw-Hill, New York, 1957.

Dixon, WF and FJ Massey. Introduction to Statistical Analysis, (3rd Edition), McGraw-Hill, New York, 1969.

Fisher, RF and F Yates. Statistical Tables for Biological, Agricultural and Medical Research, (5th Edition), Oliver & Boyd, London, 1957.

Kendall, MG and A Stuart. The Advanced Theory of Statistics, Vol. II, Inference and Relationship, (3rd Edition), Hafner, New York, 1973.

McCullough, RS, J Gurland and L Rosenberg. Small sample behaviour of certain tests of the hypothesis of equal means under variance heterogeneity, Biometrika, 47:345-53, 1960.

Mickey, MR and MB Brown. Bounds on the distribution functions of the Behrens-Fisher statistic, Ann. Math. Stat. 37:639, 1966.

Nie, NH, CH Hull, JG Jenkins, K Steinbrenner and DH Bent. SPSS: Statistical Package for the Social Sciences, (2nd Edition), McGraw-Hill, St. Louis, 1975.

Pagurova, VI. On a comparison of means of two normal samples, Theory of Probability and Its Applications, 13:527-34, 1968.

Pearson, ES and HO Hartley, (eds.). Biometrika Tables for Statisticians, (3rd Edition), Cambridge University Press, 1966.

Remington, RO and MA Schork. Statistics with Applications to the Biological and Health Sciences, Prentice-Hall, Englewood Cliffs, NJ, 1970.

Satterthwaite, FE. Synthesis of variance, Psychometrika, 6:309-16, 1941.

Smith, HF. The problem of comparing the results of two experiments with unequal errors, J. of Commercial Science and Industrial Research, Australia, 9:211-2, 1936.

Snedecor, GW and WG Cochran, Statistical Methods, (6th Edition), Iowa State Press, Ames, Iowa, 1967.

Sokal, RR and FJ Rohlf, Biometry: The Principles and Practice of Statistics in Biological Research, WH Freeman, San Francisco, 1969.

Sukhatme, PV. On Fisher and Behrens' test of significance for the difference in means of two normal samples, Sankhya, 4:39-48, 1942.