ON THE COMPARISON OF TWO MEANS FROM NORMAL POPULATIONS WITH UNKNOWN VARIANCES

By M. B. Golhar

[*Institute of Agricultural Research Statistics* (*I.C.A.R.*), *New Delhi*]*

1. Introduction

The problem of "Testing the equality of two means from normal populations with unknown (possibly unequal) variances", which is known as the Behrens-Fisher problem, was first considered by W. V. Behrens (1929) and then by R. A. Fisher (1935). Behrens suggested that the distribution of the difference between two means could be expressed in terms of observations in the samples from the normal populations, argument being independent of variances. Fisher obtained the same result in terms of fiducial probability. P. V. Sukhatme (1938) gave the tables for Fisher's method for testing the equality of two means. This method based on fiducial probability was criticised by Bartlett (1936, 1939), Neyman (1941) and others, criticism being based on the fact that the probability of rejecting the hypothesis, when true, is generally less than the nominal value of 0·05. G. S. James (1959) calculated this probability for the parameters $f_1 = f_2 = 1$ and has shown it lying in between 5% and 0·39%.

Bartlett tried to solve this problem by using the theory of confidence interval but the work remained unpublished. It was briefly mentioned by Welch (1938) and Neyman (1941) in their papers. The full solution was given by H. Scheffe (1943). He tried to minimise the confidence interval for the linear function of the sample values from two normal populations and obtained an appropriate statistic. This statistic is distributed in student's t-distribution with $n_1 - 1$ degrees of freedom ($n_1 \leqslant n_2$), where n_1 and n_2 are the sample sizes from two normal populations respectively. This statistic is also proved to be the best amongst all based on t-distribution and with maximum number of degrees of freedom.

* Now with Psychometric Unit, National Council of Educational Research and Training, Delhi 9.

Welch has given a new method (1947) in which he considers the statistic v, where

$$v = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}; \bar{x}_1, \bar{x}_2 \text{ and } s_1^2, s^2_2$$

being the sample means and the sample variances from two respective normal populations and finds out the critical value in series form. The same method was elaborately explained by Trickett and Welch (1954). The main difference between this method and Fisher's method based on fiducial probability is that here the statement is made by averaging over the joint probability distribution of s_i^2, *i.e.*, the fact that different values of s_i^2 can arise by chance in sampling from the populations with fixed variances is taken into account. Tables for this test have been given by A. A. Aspin (1949) and by Trickett, Welch and James (1956). The level of significance is also shown, by Welch, to be approximately 0·05, error being in the fourth decimal place (1949, 1954).

The two methods given by Scheffe and by Welch can be compared if we find the power function or the second kind of error of each test for the same significance level. Tables for the second kind of error for t-statistic has been given by Neyman (1935) and the second kind of error for Welch's test will be calculated here for some values of the parameters for the purpose of comparison. Bennett* and Hsu (1961) have also computed the power functions for the tests of Behrens-Fisher and Welch by Monte Carlo methods. The numerical values of the second kind of error calculated in this paper (Table I) can be compared with the values given by Bennett and Hsu.

2.1. *Welch's Critical Region*

Given x_{1i} $(i = 1 \ldots n_1)$, and x_{2j} $(j = 1 \ldots n_2)$ the samples of sizes n_1 and n_2 from the normal populations $N(\mu_1, \sigma_1^2)$, and $N(\mu_2, \sigma_2^2)$ respectively, μ_1 and μ_2 being the population means, and σ_1^2 and σ_2^2, the population variances. Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means, and s_1^2 and s_2^2 be the estimates of σ_1^2 and σ_2^2 respectively. Let

$$y = \bar{x}_1 - \bar{x}_2, \quad \sigma^2 = \lambda_1\sigma_1^2 + \lambda_2\sigma_2^2, c_i = \frac{\lambda_i s_i^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2},$$

* The author is indebted to the referee for pointing out the connected work done by Bennett and Hsu.

$$\eta = \mu_1 - \mu_2, \quad \delta = \frac{\lambda_1 \sigma_1^2}{\lambda_1 \sigma_1^2 + \lambda_2 \sigma_2^2}, \quad i = 1, 2;$$

$$f_i = n_i - 1, \quad c_1 = c \quad \text{and} \quad c_2 = 1 - c,$$

$$\lambda_i = \frac{1}{n_i}, \quad \rho = \frac{\eta}{\sigma}, \quad v = \frac{y}{\sqrt{\lambda_1 s_1^2 + \lambda_2 s_2^2}};$$

and the level of significance, $\alpha = 0 \cdot 05$.

To test the hypothesis $H_0: \eta = 0$ against the alternative hypothesis $H_1: \eta = a$ quantity (> 0), say η, the critical region in Welch's test is $v \geqslant V(f_1, f_2, c, \alpha) = V(c)$, where $V(c)$ is the critical value for given f_1, f_2, c and α, so that

$$\text{Prob.}\, [v \geqslant V(c) \mid H_0] = \alpha = 0 \cdot 05,$$

where

$$V(c) = \xi \left[1 + \left(\frac{1+\xi^2}{4}\right) \left(\frac{\sum_i \frac{\lambda_i^2 s_i^4}{f_i}}{\left(\sum_i \lambda_i s_i^2\right)^2} \right) - \left(\frac{1+\xi^2}{2}\right) \left(\frac{\sum_i \frac{\lambda_i^2 s_i^4}{f_i^2}}{\left(\sum_i \lambda_i s_i^2\right)^2} \right) + \left(\frac{3+5\xi^2+\xi^4}{3}\right) \left(\frac{\sum_i \frac{\lambda_i^3 s_i^6}{f_i^2}}{\left(\sum_i \lambda_i s_i^2\right)^3} \right) - \left(\frac{15+32\xi^2+9\xi^4}{32}\right) \left(\frac{\sum_i \frac{\lambda_i^2 s_i^4}{f_i}}{\left(\sum_i \lambda_i s_i^2\right)^2} \right)^2 \right] \qquad (1)$$

and

$$\xi = 1 \cdot 6449 \text{ (Welch, 1947).}$$

(1) can be written as

$$V(c) = \xi \Bigg[1 + \left(\frac{1+\xi^2}{4}\right)\left(\sum_i \frac{c_i^2}{f_i}\right) - \left(\frac{1+\xi^2}{2}\right)\left(\sum_i \frac{c_i^2}{f_i^2}\right)$$
$$+ \left(\frac{3+5\xi^2+\xi^4}{3}\right)\left(\sum_i \frac{c_i^3}{f_i^2}\right)$$
$$- \left(\frac{15+32\xi^2+9\xi^2}{32}\right)\left(\sum_i \frac{c_i^2}{f_i}\right)^2 \Bigg].$$

2.2. *Power function of Welch's test*

Let the second kind of error be equal to β in Welch's test. Then

$$\beta = \text{Prob.}\,[v < V(c) \mid H_1]$$
$$= \text{Prob.}\,[0 < s_1 < \infty,\ 0 < s_2 < \infty,$$
$$-\infty < y < V(c)\sqrt{\lambda_1 s_1^2 + \lambda_2 s_2^2} \mid H_1]$$

$$= \int_0^\infty \int_0^\infty \frac{\left(\frac{1}{2}\right)^{(f_1+f_1)/2}}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)} \left(\frac{f_1 s_1^2}{\sigma_1^2}\right)^{(f_1/2)-1} \left(\frac{f_2 s_2^2}{\sigma_2^2}\right)^{(f_2/2)-1}$$

$$\times\, e^{-\frac{1}{2}(f_1 s_1^2/\sigma_1^2 + f_2 s_2^2/\sigma_2^2)} \int_{-\infty}^{V(c)\sqrt{\lambda_1 s_1^2+\lambda_2 s_2^2}} \frac{1}{\sqrt{2\pi}\sigma}$$

$$\times\, e^{-(y-\eta)^2/2\sigma^2}\, dy\, d\left(\frac{f_1 s_1^2}{\sigma_1^2}\right) d\left(\frac{f_2 s_2^2}{\sigma_2^2}\right)$$

$$= B \int_0^\infty \int_0^\infty V^{(f_1/2)-1} W^{(f_2/2)-1} e^{-\frac{1}{2}(f_1 V + f_2 W)} dV\, dW$$

$$\times \int_{-\infty}^{V(c)\sqrt{\delta V + (1-\delta)W} - \rho} e^{-u^2/2} du,$$

where

$$B = \frac{\left(\frac{1}{2}\right)^{(f_1+f_2)/2} f_1^{f_1/2} f_2^{f_2/2}}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)\sqrt{2\pi}},$$

and

$$U = \frac{y - n}{\sigma}, \qquad V = \frac{s_1^2}{\sigma_1^2}, \qquad W = \frac{s_2^2}{\sigma_2^2}.$$

Let K be the function of V and W, where

$$K = V(c)\sqrt{\delta V + (1 - \delta) W}.$$

Dividing the last integral into two, we get

$$\beta = B \int_0^\infty \int_0^\infty V^{(f_1/2)-1} W^{(f_2/2)-1} e^{-\frac{1}{2}(f_1 V + f_2 W)}\, dV dW \int_{-\infty}^{-\rho} e^{-u^2/2}\, du$$

$$+ B \int_0^\infty \int_0^\infty V^{(f_1/2)-1} W^{(f_2/2)-1} e^{-\frac{1}{2}(f_1 V + f_2 W)} dV dW \int_{-\rho}^{K-\rho} e^{(-u^2/2)} du. \qquad (2)$$

The first term on r.H.S. of (2) $= \cdot 1587$, for $\rho = 1$. Expanding the term $e^{-u^2/2}$ in the second term of r.H.S. of (2), and integrating it, we get

$$B \int_0^\infty \int_0^\infty V^{(f_1/2)-1} W^{(f_2/2)-1} e^{-\frac{1}{2}(f_1 V + f_2 W)}$$

$$\times \left[K - \frac{(K-\rho)^3 + \rho^3}{6} + \frac{(K-\rho)^5 + \rho^5}{40} - \frac{(K-\rho)^7 + \rho^7}{336} + \frac{(K-\rho)^9 + \rho^9}{3456} \, dvdw. \right] \qquad (3)$$

Let

$$I_r = B \int_0^\infty \int_0^\infty V^{(f_1/2)-1} W^{(f_2/2)-1} e^{-\frac{1}{2}(f_1 V + f_2 W)} K^r dV dW,$$

where r takes the values 1, 2, 3, $\cdots$.

Now solving I_r first and then substituting the values in (3), we get the value of the second term in (2) and ultimately the value of β. We solve I_r by the transformation of V and W into V and t, where $(W/V) = t$, and then integrating with respect to V, and again by the transformation

$$c = \frac{1}{1 + \frac{1-\delta}{\delta} t},$$

we get

$$I_r = \frac{B\left|\dfrac{f_1+f_2+r}{2}\right.}{\left(\dfrac{f_1}{2}\right)^{(f_1+f_2+r)/2}} \, \delta^{r/2} \left(\frac{\delta}{1-\delta}\right)^{f_2/2}$$

$$\times \int_0^1 \frac{[V(c)]^r \, c^{(f_1/2)-1} (1-c)^{(f_2/2)-1} \, dc}{\left[c + \left(\dfrac{f_2}{f_1}\right)\left\{\dfrac{\delta(1-c)}{1-\delta}\right\}\right]^{(f_1+f_2+r)/2}} \, . \tag{4}$$

where

$$V(c) = (a_0 + a_1 c + a_2 c^2 + a_3 c^3 + a_4 c^4),$$

$a_0 \cdots a_4$ being constants calculated from (1).

The integral part of I_r in (4) can be solved by substituting the numerical values of parameters and using the transformation

$$\int_0^1 \frac{x^{m-1}(1-x)^{n-1}}{(a+bx)^{m+n+p}} \, dx$$

$$= \frac{1}{a^{n+p}(a+b)^{m+p}} \int_0^1 x^{m-1}(1-x)^{n-1}(a+b-bx)^p \, dx \tag{5}$$

where m, n, a and $a+b$ are positive constants, with the expansion of the term $(a+b-bx)^p$ in (5).

The values of β calculated for a few particular cases are given in Table I. It has not been possible to make Table I more comprehensive as the calculations are very lengthy.

The second kind of error for Scheffe's test for $\rho = 1$ is $\cdot 775$ when smaller of $(f_1, f_2) = 6$. It can be seen for the cases considered in Table I that the power of Welch's test is slightly larger than the power of Scheffe's test, which may be explained as the effect of the fact that Scheffe's first kind of error is exactly equal to $0 \cdot 05$, whereas Welch takes a slight freedom at this point by approximating it nearly equal to $0 \cdot 05$.

TABLE I

f_1	f_2	ρ	δ	β
6	6	1	·1	·77
10	6	1	·1	·77
16	6	1	·1	·77
6	6	1	·5	·76
26	6	1	·2	·76
6	26	1	·1	·75

3. COMPARISON WITH WALD'S SOLUTION

Wald has given a solution to this problem (posthumously published work) for a particular case when $n_1 = n_2 = n$. The statistic is same as that in Welch's test.

$$\text{Prob.}\left[\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\lambda_1 s_1^2 + \lambda_2 s_2^2}} \geqslant \phi(l)\right] = \alpha,$$

where $\phi(l)$ is the function of l, and $l = (s_2^2/s_1^2)$. He finds out the value of $\phi(l) = K - K'[l.(1 + l)^2]$, for small values of n; and $\phi(l) = K$, for large values of n; where K and K' are positive constants, determined by the probability statements prob. $[t_f \geqslant K] = \alpha$, where t_f is student's t-variable with f degrees of freedom; and prob. $(t_{2f} \geqslant K - (K'/4)] = \alpha$, where t_{2f} is student's t-variable with $2f$ degrees of freedom. Wald claims that no appreciably better critical region exists though it is not the best possible one.

$\phi(l)$ can be put in the series form in increasing powers of $1/f$ in terms of c_i's.

From the two probability statements given above, we get

$$K = \xi\left[1 + \frac{1 + \xi^2}{4f} + \frac{3 + 16\xi^2 + 5\xi^4}{96f^2} + \text{etc.}\right], \text{ (Fisher, 1941)}$$

and

$$K - \frac{K'}{4} = \xi\left[1 + \frac{1 + \xi^2}{8f} + \frac{3 + 16\xi^2 + 5\xi^4}{384f^2} + \text{etc.}\right],$$

and

$$\phi(l) = \xi\left[1 + \frac{1+\xi^2}{4f}(c_1{}^2 + c_2{}^2) + \frac{3+16\xi^2+5\xi^4}{96f^2}(c_1{}^3 + c_2{}^3)\right].$$

For Welch's test the critical value $V(c)$ is given by

$$V(c) = \xi\left[1 + \frac{1+\xi^2}{4f}(c_1{}^2 + c_2{}^2) + \frac{3+5\xi^2+\xi^4}{3f^2}(c_1{}^3 + c_2{}^3) - \frac{1+\xi^2}{2f^2}(c_1{}^2 + c_2{}^2) - \frac{15+32\xi^2+9\xi^4}{32f^2}(c_1{}^2 + c_2{}^2)^2\right].$$

For Fisher's method (1941) the critical value is given by

$$d(c) = \xi\left[1 + \frac{1+\xi^2}{4f}(c_1{}^2 + c_2{}^2) + \frac{2}{f}c_1c_2\right]$$

to the order of $1/f$. The first two solutons, namely of Wald and Welch, are same to the order of $1/f$ but differ in the terms of $1/f^2$. Fisher's solution differs from the first two also in the term containing $1/f$.

The numerical values for these solutions are given in Table II for $f_1 = f_2 = 8$, $\alpha = 0\cdot05$ and prob. $[|v| \geqslant \phi(l)] = \alpha$. Wald's $\phi(l)$ is given for the corresponding l or c.

TABLE II (*a*)

l ..	0	·1	·2	·3	·4	·5	·6	·7	·8	·9	1·0
c ..	1	·909	·833	·769	·714	·666	·625	·588	·555	·526	·500
$\phi(l)$	2·306	2·245	2·203	2·174	2·155	2·141	2·132	2·126	2·122	2·121	2·120

Welch's and Wald's critical values for the same case:

TABLE II (*b*)

c	0 or 1	·1 or ·9	·2 or ·8	·3 or ·7	·4 or ·6	·5
$V(c)$	2·31⁻	2·25	2·20	2·14	2·10	2·08
$\phi(l)$	2·306	2·239	2·187	2·150	2·128	2·120

Range of Fisher's values of $d(c)$ is $2 \cdot 306$ to 2.292 as l changes from 0 to 1 for the above values of the parameters.

Wald suggests that a slightly higher value of K' would give better results. Then it can be seen that Wald's critical values will go nearer to Welch's values around the value of $c = \cdot 5$ but not when c tends to 0 or 1. It may be noted that Welch's critical region is better than Wald's in controlling the first kind of error. The probability of rejecting the hypothesis, when true, lies in between $\cdot 045$ to $\cdot 051$ for Wald's critical region ($f_1 = f_2 = 5$), while for Welch's critical region ($f_1 = f_2 = 6$) it is accurate at least for the first three decimal places.

Power function for Wald's test can be found out exactly in the same way as given in Section 2.2 by considering $\phi(l)$ instead of $V(c)$. But as the level of significance for Wald's test is approximate, we may find its approximate power function by considering the critical value $\phi(l)$, only to the order of $1/f$, which is same as Welch's critical value to that order.

Now to order $1/f$, Welch's test is known to be equivalent to a procedure which uses "effective" degrees of freedom somewhere between $f_1 + f_2$ and the smaller of f_1 and f_2. The second kind of error for Welch's test to order $1/f$ is in fact given by the non-central t-distribution with the degrees of freedom F satisfying (Welch, 1947).

$$\frac{1}{F} = \frac{\delta^2}{f_1} + \frac{(1-\delta)^2}{f_2}.$$

It may be seen that the second kind of error thus calculated is same as given in Table I for the two decimal places.

4. Summary

Welch's and Scheffe's solutions for the Behrens-Fisher problem are compared by finding their power functions. Wald's solution is put in a form which can be compared with Welch's solution.

5. Acknowledgment

The author is grateful to Dr. M. N. Ghosh, for his guidance in the preparation of this paper. The author also thanks Dr. V. G. Panse and Dr. G. R. Seth for their encouragement and the interest they have shown during the course of the work.

6. References

1. Aspin, A. A. .. "Tables for use in comparisons whose accuracy involves two variances, separately estimated," *Biometrika*, 1949, **36**, 290–93.

2. Bartlett, N. S. .. "The information available in small samples," *Proc. Camb. Phil. Soc.*, 1936, **32**, 560–66.

——— .. "Complete simultaneous fiducial distributions," *Ann. of Math. Stat.*, 1939, **10**, 129–38.

3. Behrens, W. V. .. *Landw. J.*, 1929, **68**, 807–37.

4. Bennett, B. M. and Hsu, P. "Sampling studies on the Behrens-Fisher problem," *Metrika*, 1961, **4**, 89–104.

5. Fisher, R. A. .. "Fiducial argument in statistical inference," *Ann. Eugen. Lond.*, 1935, **6**, 391-98.

——— .. "The asymptotic approach to Behrens' integral with further tables for the *d*-test of significance," *Ibid.*, 1941, **11**, 141–72.

6. James, G. S. .. "Behrens-Fisher distribution and weighted means," *J.R.S.S.*, 1959, **21 B**, 73–90.

7. Neyman, J. .. "Statistical problems in agricultural experimentation," *Ibid.*, 1935, **2 B**, 131.

——— .. "Fiducial argument and theory of confidence interval," *Biometrika*, 1941, **32**, 128–50.

8. Scheffe, H. .. "On solution of the Behrens-Fisher problem based on *t*-distribution," *Ann. of Math. Stat.*, 1943, **14**, 35–44.

9. Sukhatme, P. V. .. "On Fisher and Behrens' test of significance for the difference in means of two normal samples," *Sankhya*, 1938, **4**, 39–48.

10. Trickett, W. H., Welch, B. L. and James, G. S. "Further critical values for two mean problems," *Biometrika*, 1956, **43**, 203–05.

11. Trickett, W. H. and Welch, B. L. .. "On the comparison of two means and further discussion of iterative methods for calculating tables," *Ibid.*, 1954, **41**, 361–74.

12. Wald, A. .. *Selected Papers in Statistics and Probability*, pp. 669–95.

13. Welch, B. L. .. "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, 1947, **34**, 28–35.

——— .. "Further note on Mrs. Aspin's tables and on certain approximations to the tabled function", *Ibid.*, 1949, **36**, 293–96.