Sankhyā : The Indian Journal of Statistics
2002, Volume 64, Series A, Pt.1, pp. 139-155

A SIMPLE CONSERVATIVE AND ROBUST SOLUTION OF THE BEHRENS-FISHER PROBLEM

By HAROLD RUBEN
Epping, Essex, U.K.

SUMMARY. A basic and elementary geometrical proposition is used to provide a solution of the Behrens-Fisher problem which is both conservative and robust. An explicit formula is given which allows a strict upper bound of p to be obtained and also allows confidence intervals of the difference in population means to be computed with confidence coefficient strictly greater than $1-\alpha$. (α is a prescribed small positive constant). The main results of this paper are given by (Ia),(Ib), (Ic), (Id) together with (II), in Section 2, and (III) is Section 3.

Notation. we conform to the convention that random variables shall be denoted by capital letters, whilst constants and observed values of random variables be denoted by lowercase letters. If X is a random variable with a continuous and strictly increasing (cumulative) distribution function over its range, then the unique upper γ point of X will be denoted by X_γ, i.e. $Pr\{X \geq X_\gamma\} = \gamma$. The sole exception is that a chi-squared random variable with ν degree of freedom will be denoted by χ^2_ν, and the upper α point of χ^2_ν by $\chi^2_{\nu;\alpha}$. The symbol $\sim$ will denote "is distributed as". The letter Z will denote a standardised normal random variable.

1. The Current Status of the Behrens-Fisher Problem

The Behrens-Fisher problem (Behrens 1929; Fisher 1935, 1941, 1973; Kendall and Start, 1961; Cox and Hinkley, 1996; and Stuart *et al.* 1999) – that of testing and interval estimation of the difference in means of two normal populations with totally unknown means μ_1, μ_2 and totally unknown variances σ_1^2, σ_2^2-has engendered more controversy than perhaps any other statistical problem in the 20-th century. This is hardly surprising in view of the fact that the Behrens-Fisher problem brings into sharp relief the various philosophical currents of 20th century statistical science, personified by the four pairs of statisticians Fisher and Barnard, Neyman and Pearson, Ramsey and de Finetti, Jeffreys and Lindley. At the beginning of the 21st century, it is perhaps opportune to present a critical overview of the various solutions that have been proposed for the problem, and at the same

Paper received November 1999; revised May 2000.
AMS (1991) *subject classification.* 62F03, 62F25.
Key words and phrases. Behrens-Fisher problem, geometrical proposition, conservative and robust, strict upper bound for p, conditional and unconditional pivotals, tail probabilities.

time to offer a new solution to the problem which is easy to use and is at the same time both conservative and robust.

The Behrens-Fisher problem might arise is practice if, for example, one wished to determine whether two varieties of sugar beet had on the average significantly different sugar contents, differences in the variability among individual plants being unimportant. Again, one might wish to test whether a population mean estimated by a new technique differed significantly from the estimate obtained by another and well established technique, e.g. the spectroscopic and chemical techniques of estimating the proportion of carbon in steel. Yet a third example is the situation where a pedagogue might wish to compare two different methods of teaching reading to young children. (In all of these cases, negative observations are meaningless, so that the assumption of normality cannot be exactly true, but that assumption may provide a good approximation if the coefficients of variation σ_i/μ_i are very small, thereby giving an exceedingly low probability of obtaining a negative observation).

Fisher's solution, essentially first given by Behrens in 1929, is to compound the finducial distributions of the two separate population means generated by the data, as summarised by $(\overline{x}_1, \overline{x}_2, s_1^2, s_2^2)$, the observed sample means and sample variances. Two random samples of sized $n_1, n_2 (n_1 \geq 2, n_2 \geq 2)$ are taken from the two populations. Let $\overline{X}_1, \overline{X}_2$, denote the two sample means, and S_1^2, S_2^2 the two (unbiased) sample variances. Then, from the fact that:

$$\frac{\sqrt{n_1}(\overline{X}_1 - \mu_1)}{S_1} \sim T_{\nu_1}, \quad \frac{\sqrt{n_2}(\overline{X}_2 - \mu_2)}{S_2} \sim T_{\nu_2}, \tag{1.1}$$

where $\nu_1 = n_1 - 1, \nu_2 = n_2 - 1$ and T_{ν_1}, T_{ν_2} are independent Student Variables with ν_1 and ν_2 degrees of freedom, one deduces that, conditionally on $(\overline{X}_1, \overline{X}_2, S_1^2, S_2^2)$ being observed to be $(\overline{x}_1, \overline{x}_2, s_1^2, s_2^2)$,

$$\mu_1 \overset{f}{\sim} \overline{x}_1 + \frac{s_1}{\sqrt{n_1}} T_{\nu_1}, \quad \mu_2 \overset{f}{\sim} \overline{x}_2 + \frac{s_2}{\sqrt{n_2}} T_{\nu_2} \tag{1.2}$$

where the symbol $\overset{f}{\sim}$ denotes "is fiducially distributed as". Then, differencing and dividing by the factor $\{(s_1^2/n_1) + (s_2^2/n_2)\}^{1/2}$, one deduces that the fiducial distribution of δ, where $\delta = \mu_1 - \mu_2$, is given by

$$\frac{\delta - (\overline{x}_1 - \overline{x}_2)}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}} \overset{f}{\sim} (\sin\theta) T_{\nu_1} - (\cos\theta) T_{\nu_2} \tag{1.3}$$

where

$$\tan\theta = \frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}} \tag{1.4}$$

In Fisher's reasoning, then, once the pair of samples have been drawn and observed, and in particular, once $(\overline{x}_1, \overline{x}_2, s_1^2, s_2^2)$, the observed value of the minimal sufficient statistics $(\overline{X}_1, \overline{X}_2, S_1^2, S_2^2)$ for the (four dimensional) vector parameter $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ is available, then $\overline{x}_1, \overline{x}_2, s_1^2, s_2^2$ are not to be regarded as random, and therefore θ in (1.4) is not to be regarded as random. Fisher's procedure for

testing the null hypothesis H_0 that $\delta = 0$ is to base oneself on (1.3) and (1.4) and to reject H_0 at significance level α whenever

$$\frac{|\overline{x}_1 - \overline{x}_2|}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}} \geq d_{\nu_1,\nu_2,\theta;\alpha/2} \tag{1.5}$$

where $d_{\nu_1,\nu_2,\theta;\alpha/2}$ is the upper $\alpha/2$ point of the random variable $D_{\nu_1,\nu_2;\theta}$ defined by

$$D_{\nu_1,\nu_2;\theta} = (\sin\theta)T_{\nu_1} - (\cos\theta)T_{\nu_2}. \tag{1.6}$$

Fisher and Yates, 1975, Table VI, 6th edition, provide tables of $d_{\nu_1,\nu_2,\theta;\alpha/2}$ for various (small) α, various ν_1 and ν_2 and for $\theta = 0^0(15^0)90^0$. (See Ruben[2], 1960, for an investigation of the distribution of $(D_{\nu_1,\nu_2;\theta}.)$ Furthermore, Fisher 1941, provides an asymptotic expansion for the (cumulative) distribution function of $D_{\nu_1,\nu_2;\theta}$, which expresses that function as a series of corrective terms, in powers of $1/\nu_1$ and $1/\nu_2$, to the standardised normal distribution. Fisher's expansion allows the observed p of P, once $|\overline{X}_1 - \overline{X}_2|/S$ has been observed, to take the value $|\overline{x}_1 - \overline{x}_2|/s$, to be computed, where

$$S^2 = S_1^2/n_1 + S_2^2/n_2, \quad s^2 = s_1^2/n_1 + s_2^2/n_2 \tag{1.7}$$

(s is the estimated standard error of $\overline{X}_1 - \overline{X}_2$.) Fisher's solution has been questioned by various statisticians, notably Bartlett, 1936, 1937, 1956, Welch 1947, and Pearson and Hartley, 1966, p. 26, mainly on the grounds that the application of (1.5) does not in the long run lead to a proportion α of samples for which H_0 is rejected when H_0 is valid – in brief that the Fisher test is not similar. Fisher, however, expressed the opinion that the latter requirement is foreign to the logic of the problem and has trenchantly remarked that "the notion of repeated sampling from a fixed population has completed its usefulness when the simultaneous distribution of T_{ν_1} and T_{ν_2} has been obtained". It has been shown by Robinson, 1976, that the Behrens-Fisher test is conservative in the sense that the probability of an error of the first kind cannot exceed the level of significance α.

The first attempted *frequentist* solution of the Behrens-Fisher problem was made by Scheffé, 1943. Scheffé shows that under H_0 (assuming without loss of generality, $n_1 \leq n_2$)

$$\frac{\sqrt{n_1}(\overline{X}_1 - \overline{X}_2)}{S_U} \sim T_{n_1-1}, \tag{1.8}$$

so that one rejects H_0 at significance level α, whenever

$$\frac{\sqrt{n_1}|\overline{x}_1 - \overline{x}_2|}{s_u} \geq t_{n_1-1;\alpha/2}, \tag{1.9}$$

[2]There is a regrettable transcription error in (3.7) of that paper, in that the ratio $\Gamma(-\frac{1}{2}f + \frac{1}{2} + r)/\Gamma(-\frac{1}{2}f + \frac{1}{2})$ should be read as $(-(f-1)/2)_r$ under the summation sign. Note incidehally that the series in (3.7) is finite if f is odd.

Here

$$S_U^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(U_j - \overline{U})^2, \qquad \overline{U} = \frac{1}{n_1}\sum_{j=1}^{n_1} U_j, \tag{1.10}$$

$$U_j = X_{1j} - (n_1/n_2)^{1/2} X_{2j}^*, \quad j = 1, 2, \ldots, n_1, \tag{1.11}$$

where $(X_{11}, X_{12}, \ldots, X_{1n_1})$ is the sample of size n_1 from the first population and $(X_{21}^*, X_{22}^*, \ldots, X_{2n_1}^*)$ is a sub-sample of size n_1, drawn at random from the sample $(X_{21}, X_{22}, \ldots, X_{2n_2})$, of size n_2, from the second population. Scheffé's solution can be criticised on several grounds. First, it discards information on variability from $n_2 - n_1$ observation in the second sample, and this is obviously particularly serious if n_2 is much larger that n_1 (Table VI by Fisher and Yates even provide for the case $n_2 = \infty$, i.e., in practice, n_2 is very large.) Secondly, if $n_1 = n_2 = n$(say), Scheffé's solution (see (1.9), (1.10) and (1.11)) for testing H_0 at significance level α is to reject H_0 whenever

$$\frac{\sqrt{n}|\overline{x}_1 - \overline{x}_2|}{s_u} \geq t_{n-1;\alpha/2}, \tag{1.12}$$

where

$$s_u^2 = \frac{1}{n-1}\sum_{j=1}^{n}(u_j - \overline{u})^2, \quad \overline{u} = \frac{1}{n}\sum_{j=1}^{n} u_j, \tag{1.13}$$

$$u_j = x_{1j} - x_{2j}^*, \quad j = 1, 2, \ldots, n, \tag{1.14}$$

and $(x_{21}^*, x_{22}^*, \ldots, x_{2n}^*)$ is a random permutation of the observed second sample $(x_{21}, x_{22}, \ldots, x_{2n})$. However, if one's plan is to use the distribution of T_{n-1} for a test of H_0, the above is not the *natural* test, but rather a "paired" T-test. Specifically, one adopts a randomised blocks arrangement in which there are n blocks, such that within each block there are two relatively homogenous experimental units (identical twins, adjacent plots of land, etc.). For the i-th block ($i = 1, 2, \ldots, n$), one tosses a fair coin to decide which of the two experimental units shall be given treatment 1 or treatment 2, in order to eliminate as far as possible any residual inherent heterogenerity between the two units. If

$$d_i = x_{i1} - x_{i2} \quad (i = 1, 2, \ldots, n) \tag{1.15}$$

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2, \quad \overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \equiv \overline{x}_1 - \overline{x}_2, \tag{1.16}$$

then one rejects H_0 at significance level α whenever

$$\frac{\sqrt{n}|\overline{x}_1 - \overline{x}_2|}{s_d} \geq t_{n-1;\alpha/2}, \tag{1.17}$$

which differs fundamentally from (1.12). Thirdly, statistical inference is an important component of more general scientific inference (a point frequently stressed by Fisher - see, for example Fisher 1966, 1973) and no serious scientific investigator

could tolerate a situation where his or her conclusions will be governed in part by the outcome of a random extracton of n_1 cards from a pack n_2 cards (when $n_1 < n_2$) or by a random shuffing of n cards (as Scheffé in effect does). Scheffé's solution must therefore be discounted on both practical and philosophical grounds. It should be remarked here that the fundamental priciple of randomisation (as for the "paired" T-test), first proposed by Fisher (see e.g., Fisher 1966) is of an altogether different character, in that its purpose is to eliminate, or at least minimise, the effects of extraneous and uncontrollable factors on the data.

The next attempt at a frequentist solution of the Behrens-Fisher problem was made by Welch, 1947. Welch seeks a statistic $h(S_1^2, S_2^2; \alpha)$ such that under H_0

$$Pr\{\overline{X}_1 - \overline{X}_2 \geq h(S_1^2, S_2^2; \alpha)\} = \alpha, \tag{1.18}$$

i.e. he seeks for a similar test of H_0. Welch obtains, by means of a Taylor expansion of $h(S_1^2, S_2^2; \alpha)$ about the points (σ_1^2, σ_2^2), proceeding formally,

$$h(S_1^2, S_2^2; \alpha) = Sz_\alpha \left\{ 1 + \frac{(1+z_\alpha)^2}{4} \sum_{i=1}^{2} C_i^2/\nu_i - \frac{1+z_\alpha^2}{2} \sum_{i=1}^{2} C_i^2/\nu_i^2 + \ldots \right\} \tag{1.19}$$

where $\nu_i = n_i - 1, S$ is as in (1.7), and

$$C_i = \frac{S_i^2/n_i}{S^2} \quad (i = 1, 2). \tag{1.20}$$

Table 11 of Pearson and Hartley, 1970, uses Welch's result to provide upper 5% and upper 1% critical values of $(\overline{X}_1 - \overline{X}_2)/S$ under H_0. (See also Fisher, 1956, Bartlett, 1956, and Welch, 1956.) It will be noticed that in writing (1.18) it is tacitly assumed that a function $h(S_2^2, S_2; \alpha)$ satisfying the latter equation *exists* (which is equivalent to the assumption that a solution of a certain integral equation exists). However, Welch's solution cannot provide an exactly similar α-level test of H_0, but at best a *nearly* similar test, if n_1 and n_2 are not too small. This follows from the remarkable finding in a 1966 monopgraph by the distinguished Russian mathematician and statistician Yuri V. Linnik - an Enghlish translation has been available since 1968 - that although similar critical regions for H_0 exist, they are of such a pathological character as to render them totally unsuitable for practical usage; the critical regious are highly irregular and non-nested by α, in that there exist arbitrarily small values of $|x_1 - x_2|/s$ such that H_0 is rejected, and data which are significant at the 1% level are not automatically significant at the 5% level. In this connection, refer also to Barnard, 1982. Furthermore, Barnard, 1984, has provided cogent arguments why the Behrens-Fisher solution is preferable to the Welch solution.

The usual Bayesian solution of the Behrens-Fisher problem is obtained by assuming that $\mu_1, \mu_2, \sigma_1, \sigma_2$ have the joint improper probability element proportional to

$$d\mu_1 d\mu_2 (d\sigma_1/\sigma_1)(d\sigma_2/\sigma_2). \tag{1.21}$$

It is then easy to determine from (1.21), by combining (1.21) with the separate likelihood functions of (μ_1, σ_1) and (μ_2, σ_2), that, conditionally on obtaining $(\overline{x}_1, \overline{x}_2, s_1^2, s_2^2)$,

$$\mu_1 \overset{post}{\sim} \overline{x}_1 + \frac{s_1}{\sqrt{n_i}} T_{\nu_1}, \quad \mu_2 \overset{post}{\sim} \overline{x}_2 + \frac{s_2}{\sqrt{n_2}} T_{\nu_2} \tag{1.22}$$

where $\overset{post}{\sim}$ denotes "has the posterios distribution of". It then follows immediatly from (1.22) that the posterior distribution of δ, given $(\overline{x}_1, \overline{x}_2, s_1^2, s_2^2)$, is such that

$$\frac{\delta - (\overline{x}_1 - \overline{x}_2)}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}} \overset{post}{\sim} D_{\nu_1, \nu_2; \theta} \tag{1.23}$$

with θ as in (1.4). (Cf (1.23) with (1.3) and (1.22) with (1.2).) Thus, the usual Bayesian solution has formally the same appearance as the Behrens-Fisher solution. However, (1.21) can be seriously questioned on the least three grounds. First, it is not clear why if ϕ is a parameter with range $(-\infty, \infty)$ it should be assigned a prior probability element proportional to $d\phi$, whilst if ϕ has the range $(0, \infty)$ it should be assigned a prior probability element proportional to $d\phi/\phi$. Secondly, (1.21) implies that the same prior weights are given for μ_i in any two finite intervals of the same width, *whatever be their locations*, which will often in practice be dubious, whether the Bayesian statistical assumes a prior frequency distribution for μ_i, an objectivistic representation, or a subjective measure. Thirdly, (1.21) states that the prior weights assigned to μ_i and σ_i are independent, but in practice it will often be the case that a high value of σ_i will be associated with a numerically high value of μ_i. The assumpton (1.21) appears therefore to have no justification other than that it is mathematically convenient.

Finally, a frequentist solution of the Behrens-Fisher problem *based on two-stage sampling*, has been provided by Ruben, 1950, pp. 68-127, in his doctorial dissertations for a more accessible and at the same time more compact presentation of the results, see Ruben 1962 (the 1961 paper by Ruben is also relevant). Here the reference set is not that characterised by the two constants n_1 and n_2, as in single-stage sampling, but rather that characterised by the four design constants n_{01}, n_{02}, k_1, k_2 where n_{01}, n_{02} are arbitrary integers not less that 2 and determined in advance, and k_1, k_2, are arbitrary positive constants, also determined in advance. (See also Barnard, 1950.) For $i = 1, 2$, we take an initial sample of size n_{0i} from the i-th population. Let S_{0i}^2 denote the i-th sample variance based on $n_{0i} - 1$ degrees of freedom. We then take a further sample of random size $N_i - n_{0i}$ ($N_i - n_{0i}$ may assume the value zero), where

$$N_i = \max(\{k_i^2 S_{0i}^2\}, n_{0i}) \quad (i = 1, 2), \tag{1.24}$$

$\{k_i^2 S_{0i}^2\}$ denoting the smallest integer not less than $k_i^2 S_{0i}^2$. Let $\overline{X}_i$ denote the i-th sample mean based on N_i observations. It is then shown that, as $\sigma_i^2 \to \infty, \sigma_2^2 \to \infty$ simultaneously,

$$\frac{(\overline{X}_1 - \overline{X}_2) - \delta}{(1/k_1^2 + 1/k_2^2)^{1/2}} \sim D_{\nu_{01}, \nu_{02}; \theta} \tag{1.25}$$

where $\nu_{01} = n_{01} - 1, \nu_{02} = n_{02} - 1$, and

$$\tan\theta = \frac{1/k_1}{1/k_2}. \tag{1.26}$$

We reject the hypothesis H_0 which asserts that $\delta = 0$ at nominal significance level α whenever

$$\frac{|\overline{x}_1 - \overline{x}_2|}{(1/k_1^2 + 1/k_2^2)^{1/2}} \geq d_{\nu_{01},\nu_{02},\theta;\alpha/2}. \tag{1.27}$$

We also obtain central confidence intervals for δ, of *predetermined* and *fixed* width 2ℓ, with nominal confidence coefficient (coverage probability) $1 - \alpha$, namely,

$$(\overline{X}_1 - \overline{X}_2 - \ell, \overline{X}_1 - \overline{X}_2 + \ell) \tag{1.28}$$

provided k_1^2 and k_2^2 satisfy the requirement

$$(1/k_1^2 + 1/k_2^2)^{1/2} d_{\nu_{01},\nu_{02},\theta;\alpha/2} = \ell. \tag{1.29}$$

The adjective "nominal" has been used because the author has in fact shown that the power function corresponding to the procedure (1.27) and the interval estimation (1.28) are both conservative in character in the following sense. The actual unknown power function, which depends on the unknown parameters σ_1^2, σ_2^2, is *better* than the limiting power functon ($\sigma_1^2 \to \infty, \sigma_2^2 \to \infty$ simultaneously), corresponding to (1.27), which depends only on δ. Here we call a power function "good" if it assumes low values for numerically small values of δ and high values for numerically large values of δ. (The limiting power function is symmetrical about the origin and strictly increasing in $|\delta|$, and the test is therefore "unbiased" in the Neyman-Pearson sense.) In particular, then, the probability of an error of the first kind is stricly *less* than the nominal significance level α, whatever be the values of σ_1^2 and σ_2^2. Similarly, the actual unknown confidence coefficient for (1.28), which depends on σ_1^2 and σ_2^2, is strictly *greater* than the nominal confidence coefficient $1-\alpha$, whatever be the values of σ_1^2 and σ_2^2. It will be noticed on comparing (1.26) and (1.27) with (1.4) and (1.5) that the *constant* scale factor $1/k_i$ plays the same role in the two-stage sampling procedure as the studentising role played by the variable estimated standard error $s_i/\sqrt{n_i}$ in single-stage sampling. It will also be noted that even if the pilot sample sizes n_{01}, n_{02} have been chosen and k_1, k_2 have been selected so as to satisfy (1.29), there still remains an infinite number of pairs (k_1, k_2) to choose from. In the author's work in 1950 and 1962, the value α for testing H_0 was predetermined, the problem of choosing k_1 and k_2 was explored in detail, and k_1, k_2 chosen in such a way that the shape of the limiting power function met specified requirements, and also so that $E(N_1 + N_2)$, the expected total amount of sampling subject to the restriction (1.29), was minimised.

In section 2, we revert to the classical *single-stage* sampling situation. A key role will be played by the random variable $W^2_{\nu_1,\nu_2}$, defined by

$$W^2_{\nu_1,\nu_2} = T^2_{\nu_1} + T^2_{\nu_2} \tag{1.30}$$

where T_{ν_1}, T_{ν_2} are (as before) independent Student variables with ν_1 and ν_2 degrees of freedom. We note that the quantity

$$\frac{n_1(\overline{X}_1 - \mu)^2}{S_1^2} + \frac{n_2(\overline{X}_2 - \mu)^2}{S_2^2} \tag{1.31}$$

is a *conditional pivotal* in the following sense. The hypothesis H_0 is equivalent to the assertion that $\mu_1 = \mu, \mu_2 = \mu$ for some unknown value μ. The distribution of the quantity in (1.31) does not depend on $(\mu, \sigma_1^2, \sigma_2^2)$, and is in fact distributed as $W^2_{\nu_1,\nu_2}$, *provided that* H_0 *is valid*. This fact will be exploited to determine a conservative and robust test of H_0. We shall in fact show that the value p for H_0 is *strictly less* than a quantity depending on a number which involves only $|\overline{x}_1 - \overline{x}_2|/s$ (as well as ν_1 and ν_2), whatever be the values of σ_1^2 and σ_2^2, Further, the quantity

$$\frac{n_1(\overline{X}_1 - \mu_1)^2}{S_1^2} + \frac{n_2(\overline{X}_2 - \mu_2)^2}{S_2^2} \tag{1.32}$$

is an *unconditional pivotal* in that, whatever be the values of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, its distribution does not involve these parameters, and is, in fact, also distributed as $W^2_{\nu_1,\nu_2}$. This property will be exploited to obtain conservative confidence intervals for δ with confidence coefficient *strictly greater* than $1 - \alpha$, where α denotes a prescribed small positive quantity, whatever be the values of σ_1^2 and σ_2^2.

2. An Elementary Geometrical Proposition and A Solution of The Behrens-Fisher Problem

PROPOSITION 2.1. *For $i = 1, 2$, let $-\infty < a_i < \infty, 0 < b_i < \infty$ and $c > 0$. Then the following two statements hold.*

(i) *The region in the (v_1, v_2)-plane defined by*

$$|(v_1 - v_2) - (a_1 - a_2)| \geq c(b_1^2 + b_2^2)^{1/2} \tag{2.1}$$

is a proper subset of the region defined by

$$(v_1 - a_1)^2/b_1^2 + (v_2 - a_2)^2/b_2^2 \geq c^2. \tag{2.2}$$

(ii) *The elliptical region in the (v_1, v_2)-plane defined by*

$$(v_1 - a_1)^2/b_1^2 + (v_2 - a_2)^2/b_2^2 < c^2. \tag{2.3}$$

is a proper subset of the infinite strip defined by

$$|(v_1 - v_2) - (a_1 - a_2)| < c(b_1^2 + b_2^2)^{1/2}. \tag{2.4}$$

PROOF. The line

$$v_1 - a_1 = (v_2 - a_2) + h \tag{2.5}$$

in the (v_1, v_2)-plane intersectes the ellipse defined by

$$(v_1 - a_1)^2/b_1^2 + (v_2 - a_2)^2/b_2^2 = c^2. \tag{2.6}$$

at two coincident points, i.e. is a *tangent* to the ellipse in (2.6), if and only if the discriminant of the following quadratic equation in $v_2 - a_2$

$$(v_2 - a_2 + h)^2/b_1^2 + (v_2 - a_2)^2/b_2^2 = c^2. \tag{2.7}$$

vanishes. A simple calculation shows that the discriminant vanishes when

$$h = -c(b_1^2 + b_2^2)^{1/2}, \quad h = c(b_1^2 + b_2^2)^{1/2}. \tag{2.8}$$

The statements in (i) and (ii) then follow directly.

To apply the geometrical proposition to the Behrens-Fisher problem, set for $i = 1, 2$

$$v_i = \overline{X}_i, \quad a_i = \mu, \quad b_i = S_i/\sqrt{n_i}. \tag{2.9}$$

Then (i) says that the (random) region in the $(\overline{X}_1, \overline{X}_2)$-plane defined by

$$|\overline{X}_1 - \overline{X}_2| \geq c(S_1^2/n_1 + S_2^2/n_2)^{1/2} \tag{2.10}$$

is a proper subset of the (random) region defined by

$$\frac{n_1(\overline{X}_1 - \mu)^2}{S_1^2} + \frac{n_2(\overline{X}_2 - \mu)^2}{S_2^2} \geq c^2. \tag{2.11}$$

Thus, under H_0, which asserts that $\mu_1 = \mu$, $\mu_2 = \mu$, with μ denoting the common unknown value of μ_1 and μ_2,

$$\begin{aligned} Pr\left\{ \frac{|\overline{X}_1 - \overline{X}_2|}{S} \geq c \;\middle|\; H_0 \right\} &< Pr\left\{ \frac{n_1(\overline{X}_1 - \mu)^2}{S_1^2} + \frac{n_2(\overline{X}_2 - \mu)^2}{S_2^2} \geq c^2 \;\middle|\; H_0 \right\} \\ &= Pr\{W^2_{\nu_1,\nu_2} \geq c^2\}, \end{aligned} \tag{2.12}$$

where S is as in (1.7). Write

$$Y = \frac{|\overline{X}_1 - \overline{X}_2|}{S}, \; c = \frac{|\overline{x}_1 - \overline{x}_2|}{s} = y \tag{Ia}$$

i.e. c is the observed value of Y, and s is also as in (1.7). Then the test statistic for H_0 is Y, and

$$p = Pr\{Y \geq c | H_0\} < Pr\{W^2_{\nu_1,\nu_2} \geq c^2\}. \tag{Ib}$$

In the Appendix, we give an outline of a proof that, for arbitrary $t \geq 0$,

$$Pr\{W^2_{\nu_1,\nu_2} \geq t\} = \exp\left(-\frac{t}{2}\right) \left\{ \begin{array}{c} 1 + g_{10}(t)\left(\frac{1}{\nu_1} + \frac{1}{\nu_2}\right) + g_{20}(t)\left(\frac{1}{\nu_1^2} + \frac{1}{\nu_2^2}\right) + g_{11}(t)\frac{1}{\nu_1\nu_2} + \\ g_{30}(t)\left(\frac{1}{\nu_1^3} + \frac{1}{\nu_2^3}\right) + g_{21}(t)\left(\frac{1}{\nu_1^2\nu_2} + \frac{1}{\nu_1\nu_2^2}\right) + \ldots \end{array} \right\} \tag{Ic}$$

with

$$\begin{array}{lcl} g_{10}(t) & = & \frac{1}{16}(3t^2+4t) \\ g_{20}(t) & = & \frac{1}{12288}(105t^4-280t^3-240t^2-192t) \\ g_{11}(t) & = & \frac{1}{2048}(3t^4-8t^3-80t^2-8t) \\ g_{30}(t) & = & \frac{1}{393210}(231t^6-2772t^5+3920t^4+1920t^3-1152t^2-7680t) \\ g_{21}(t) & = & \frac{1}{393210}(21t^6-252t^5-4000t^4+2688t^3+2688t^2+1536t) \end{array} \tag{Id}$$

(Ia), (Ib), (Ic) and (Id) supply a conservative test for H_0 with a strict upper bound for p, with $t = c^2$ in (Ic) and (Id).

To obtain a confidence interval for S, set in (ii), for $i = 1, 2$,

$$v_i = \overline{X}_i,\ a_i = \mu_2,\ b_i = S_i/\sqrt{n_i},\ c = \sqrt{w^2_{v_1,v_2,;\alpha}}, \tag{2.13}$$

$w^2_{v_1,v_2;\alpha}$ denoting the upper α point of W_{v_1,v_2}. Then the random elliptical region in the $(\overline{X}_1, \overline{X}_2)$-plane defined by

$$\frac{n_1(\overline{X}_1-\mu_1)^2}{S_1^2}+\frac{n_2(\overline{X}_2-\mu_2)^2}{S_2^2} < w^2_{v_1,v_2;\alpha} \tag{2.14}$$

is a proper subset of the ranodm strip defined by

$$|(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)| < \sqrt{w^2_{\nu_1,\nu_2;\alpha}}(S_1^2/n_1+S_2^2/n_2)^{1/2}. \tag{2.15}$$

Hence

$$Pr\left\{\frac{n_1(\overline{X}_1-\mu_1)^2}{S_1^2}+\frac{n_2(\overline{X}_2-\mu_2)^2}{S_2^2}<w^2_{v_1,v_2;\alpha}\right\}<Pr\left\{|\overline{X}_1-\overline{X}_2-\delta|<\sqrt{w^2_{v_1,v_2;\alpha}}S\right\} \tag{2.16}$$

i.e.

$$Pr\{W^2_{v_1,v_2} < w^2_{v_1,v_2;\alpha}\} < Pr\left\{|\delta-(\overline{X}_1-\overline{X}_2)| < \sqrt{w^2_{v_1,v_2;\alpha}}S\right\} \tag{2.17}$$

or

$$1-\alpha < Pr\left\{|\delta-(\overline{X}_1-\overline{X}_2)| < \sqrt{w^2_{v_1,v_2;\alpha}}S\right\} \tag{2.18}$$

which we can write in the form

$$Pr\left\{\overline{X}_1-\overline{X}_2-\sqrt{w^2_{\nu_1,\nu_2;\alpha}}S<\delta<\overline{X}_1-\overline{X}_2+\sqrt{w^2_{\nu_1,\nu_2;\alpha}}S\right\} > 1-\alpha. \tag{2.19}$$

In other words, the intervals

$$\left(\overline{X}_1-\overline{X}_2-\sqrt{w^2_{v_1,v_2;\alpha}}S, \overline{X}_1-\overline{X}_2+\sqrt{w^2_{v_1,v_2;\alpha}}S\right) \tag{II}$$

are confidence intervals for δ with confidence coefficient strictly greater that $1-\alpha$. It remains only to determine the percentage points $w^2_{v_1,v_2;\alpha}$ of $W^2_{v_1,v_2}$. From (1.30)

$$W^2_{\infty,\infty} \sim \chi^2_2 \tag{2.20}$$

(note that χ^2_2 has an exponential distribution with mean 2 and with probability density function $1/2\exp(-x/2)$ for $x \geq 0$.) Since a χ^2_2 distribution is a very special case of a Pearson Type III distribution, we follow Bartlett, 1954, and approximate $W^2_{\nu_1,\nu_2}$ by a *multiple* of a χ^2_2 random variable, i.e. we write

$$T^2_{\nu_1} + T^2_{\nu_2} \equiv W^2_{\nu_1,\nu_2} \underset{\bullet}{\overset{\bullet}{\sim}} \lambda\chi^2_2 \tag{2.21}$$

where $\underset{\bullet}{\overset{\bullet}{\sim}}$ denotes "is approximately distributed as" and λ is a positive constant. Since $ET^2_\nu = \nu/(\nu-2)$, we find, on equating the means of the first and third random variables in (2.21),

$$\frac{\nu_1}{\nu_1 - 2} + \frac{\nu_2}{\nu_2 - 2} = 2\lambda \tag{2.22}$$

so that our approximation is

$$W^2_{\nu_1,\nu_2} \underset{\bullet}{\overset{\bullet}{\sim}} \frac{1}{2}\left(\frac{\nu_1}{\nu_1 - 2} + \frac{\nu_2}{\nu_2 - 2}\right)\chi^2_2, \tag{2.23}$$

and therefore

$$w^2_{\nu_1,\nu_2;\alpha} \underset{\bullet}{\overset{\bullet}{\sim}} \frac{1}{2}\left(\frac{\nu_1}{\nu_1 - 2} + \frac{\nu_2}{\nu_2 - 2}\right)\chi^2_{2;\alpha}, \tag{2.24}$$

One method of computing $w^2_{\nu_1,\nu_2;\alpha}$ is a direct graphical one, i.e. one constructs an accurate graph of $Pr\{W^2_{\nu_1,\nu_2} \geq t\}$ against t, with the probability on the vertical axis, using (Ic) and (Id), in the neighbourhood of $t = 1/2\{\nu_1(\nu_1-2)^{-1} + \nu_2(\nu_2 - 2)^{-1}\}\chi^2_{2;\alpha} = t_0$ (say). One then draws a horizontal line above the t-axis distant α from that axis. The abscissa of the point of intersection of the line with the graph the gives the value of $w^2_{\nu_1,\nu_2;\alpha}$. Another method is to use two values of t, namely $t = t_0$ and another value close to it. We then compute, using (Ic) and (Id), together with linear graduation, the value $w^2_{\nu_1,\nu_2;\alpha}$.

It is interesting to compare (II) with the "usual" approximate confidence intervals when ν_1 and ν_2 are large. In the "usual" procedure, one usues the fact that

$$\frac{(\overline{X}_1 - \overline{X}_2) - \delta}{S} \underset{\bullet}{\overset{\bullet}{\sim}} Z \tag{2.25}$$

when ν_1 and ν_2 are both large, so that approximate confidence intervals with confidence coefficient $1 - \alpha$, are

$$(\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}S, \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}S). \tag{2.26}$$

Now $Z^2 \sim \chi^2_1$, so that clearly $z_{\alpha/2} = \sqrt{\chi^2_{1;\alpha}}$, and (2.26) becomes

$$\left(\overline{X}_1 - \overline{X}_2 - \sqrt{\chi^2_{1;\alpha}}S, \overline{X}_1 - \overline{X}_2 + \sqrt{\chi^2_{1;\alpha}}S\right). \tag{2.27}$$

On the other hand, from (2.20),

$$\sqrt{w^2_{\infty,\infty;\alpha}} = \sqrt{\chi^2_{2,\alpha}},$$

so that approximately our confidence intervals (II) become

$$(\overline{X}_1 - \overline{X}_2 - \sqrt{\chi^2_{2;\alpha}}S, \overline{X}_1 - \overline{X}_2 + \sqrt{\chi^2_{2;\alpha}}S) \tag{2.28}$$

$(n_1 \to \infty,\ n_2 \to \infty)$ with confidence coefficient strictly greater that $1-\alpha$. Now χ^2_2 is stochastically larger than χ^2_1, so that clearly $\chi^2_{2;\alpha} > \chi^2_{1;\alpha}$, and the ratio, $r(\alpha)$, of the width of the confidence intervals in (2.28) to that of the confidence intervals in (2.27) is given by

$$r(\alpha) = \sqrt{\chi^2_{2;\alpha}/\chi^2_{1;\alpha}}. \tag{2.29}$$

The fact that $r(\alpha)$ is greater that 1 reflects the fact that whereas the confidence coefficient of the intervals (2.27) is (approximately, when n_1 and n_2 are both large) *equal* to $1-\alpha$, the confidence coefficient of the intervals (2.28) is *greater* that $1-\alpha$. We give below a short table of $r(\alpha)$.

α	.1	.05	.025	.01	.005	.001
$r(\alpha)$	1.29	1.25	1.21	1.20	1.16	1.12

3. **Concluding Remarks: Robustness and Similarity**

(a). It has long been known (see Bartlett, 1935, and Gayen, 1949) that, under H_0, the T-distributions of $\sqrt{n_1}(\overline{X}_1 - \mu)/S_1$ and $\sqrt{n_2}(\overline{X}_2 - \mu)/S_2$ in (1.31) are not materially affected by non-drastic departures from normality of the two parent distribution, and similarly the T-distributions of $\sqrt{n_1}(\overline{X}_1 - \mu_1)/S_1$ and $\sqrt{n_2}(\overline{X}_2 - \mu_2)/S_2$ in (1.32) are not materially affected by non-drastic departures from normality of the two parent distributions. It follows that our test procedure for H_0, based on (Ib) and $W^2_{\nu_1,\nu_2}$, is *robust*, and similarly our interval estimation procedure for δ, based on (II) and $W^2_{\nu_1,\nu_2}$, is likewise *robust*, in that the two procedures remain valid, provided only that there is no violent change from normality of the two parent distributions. This has an obvious bearing on the *generalised Behrens-Fisher problem* (GBF), posed by Barnard, 1995. Barnard considers two populations with probability density functions

$$\frac{1}{\tau_1}\phi_1\left(\frac{x-\lambda_1}{\tau_1}\right) \quad \text{and} \quad \frac{1}{\tau_2}\phi_2\left(\frac{x-\lambda_2}{\tau_2}\right) \tag{3.1}$$

where λ_1 and λ_2 are unknown location parameters $(-\infty < \lambda_1, \lambda_2 < \infty)$, whilst τ_1 and τ_2 are unknown scale parameters $(0 < \tau_1, \tau_2 < \infty)$. The GBF problem is that of testing the composite hypothesis H_0, which asserts that $\lambda_1 = \lambda_2$, with probability of an error of the first kind not exceeding α. Barnard assumes that the functions ϕ_1 and ϕ_2 are known *approximately*, but it now becomes apparent from our previous remark that that assumption need not hold, i.e., we can still test H_0 with the required specification for the risk of an error of the first kind, even if we are

quite ignorant of the functional froms ϕ_1 and ϕ_2, provided only that the underlying parent distributions do not depart dramatically from normality.

(b). In Neyman and Pearson's seminal 1933a paper – but see also the second reference 1933b to that paper which contains on pp. 67–72 an interesting historical review by E.L. Lehmann – the authors, 1933b, assert in effect on p. 98 that it is "evident" that the test of a composite hypothesis should have a constant risk of error α for all values of the nuisance parameters (in our case σ_1^2 and σ_2^2). They proceed to assert on p. 98; "The fundamental position from which we start should be noted at this point. It is assumed that the only possible critical regions are similar; that is to say regions such that[3] $P(w) = \epsilon$ for every simple hypothesis of the subset w" An important by-product of the present paper is the demonstration, by means of a concrete example, of the artificiality and lack of a logical basis of the notion of similarity in testing composite hypothesis (a view maintained by the present author for some 45 years). Neyman and Pearson's requirements are in fact neither "evident" nor "fundamental". We can illustrate the point by the following considerations. In (i) of Section 2, for the normal Behrens-Fisher problem, let $v_i = \overline{X}_i, a_i = \mu, b_i = S_i/\sqrt{n_1} (i = 1, 2)$, and let $c^2 = w^2_{\nu_1,\nu_2;\alpha}$. Then the random region in the $(\overline{X}_1, \overline{X}_2)$-pane defined by

$$|\overline{X}_1 - \overline{X}_2| \geq \sqrt{w^2_{\nu_1,\nu_2;\alpha}}(S_1^2/n_1 + S_2^2/n_2)^{1/2} \tag{3.2}$$

is a proper subset of the random region defined by

$$n_1(\overline{X}_1 - \mu)^2/S_1^2 + n_2(\overline{X}_2 - \mu)^2/S_2^2 \geq w^2_{\nu_1,\nu_2;\alpha}, \tag{3.3}$$

and accordingly

$$\begin{aligned} Pr\left\{ \frac{|\overline{X}_1 - \overline{X}_2|}{(S_1^2/n_1 + S_2/n_2)^{1/2}} \geq \sqrt{W^2_{\nu_1,\nu_2;\alpha}} \,\middle|\, H_0 \right\} \\ < \quad Pr\left\{ \frac{n_1(\overline{X}_1 - \mu)^2}{S_1^2} + \frac{n_2(\overline{X}_2 - \mu)^2}{S_2^2} \geq w^2_{\nu_1,\nu_2;\alpha} \,\middle|\, H_0 \right\} \\ = \quad Pr\{T^2_{\nu_1} + T^2_{\nu_2} \geq w^2_{\nu_1,\nu_2;\alpha}\} \quad = \quad Pr\{W^2_{\nu_1,\nu_2;\alpha} \geq w^2_{\nu_1,\nu_2;\alpha}\} \quad = \quad \alpha \end{aligned} \tag{3.4}$$

Thus, if we reject H_0 whenever

$$|\overline{x}_1 - \overline{x}_2|(s_1^2/n_1 + s_2^2/n_2)^{-1/2} \geq \sqrt{w^2_{\nu_1,\nu_2;\alpha}}, \tag{III}$$

then the probability of an error of the first kind is strictly *less* than α, whatever be tbe values of σ_1^2 and σ_2^2. We are then confronted with an apparently bizarre paradoxical situation. Linnik and Barnard (quoted earlier) have shown that the sought for acceptable test of H_0 with the probability of an error of the first kind *equal* to α, whatever be the values of σ_1^2 and σ_2^2, *does not exist*. However, on the other hand, a *better* test of H_0 than the sought for and unattainable test, in the

[3]Neymann and Pearson's ϵ is our α

specific sense that the probability of an error of the frist kind is *strictly less* than α, whatever be the values of σ_1^2 and σ_2^2, *does* exist. That test is specified by (III). The paradox arises purely from the unreasonable restrictions on tests demanded by similarity.

Appendix: Tail Probability of $W^2_{\nu_1,\nu_2}$

Our starting point is Fisher's 1925 result that the probability density function, evaluated at the point x, of a Student T_ν variable is

$$f_\nu(x) = (2\pi)^{-1/2}\exp(-x^2/2)\left\{1 + a_1(x)/\nu + a_2(x)/\nu^2 + a_3(x)/\nu^3 + \ldots\right\} \quad (A1)$$

where

$$\begin{array}{rcl} a_1(x) & = & (1/4)(x^4 - 2x^2 - 1) \\ a_2(x) & = & (1/96)(3x^8 - 28x^6 + 30x^4 + 12x^2 + 3) \\ a_3(x) & = & (1/384)(x^{12} - 22x^{10} + 113x^8 - 92x^6 - 33x^4 - 6x^2 + 15). \end{array} \quad (A2)$$

Hence the joint probability density function of two independent Student variables T_{ν_1}, T_{ν_2}, evaluated at the point (x, y), is

$$\begin{array}{c} f_{\nu_1}(x)f_{\nu_2}(y) = (2\pi)^{-1}\exp(-(x^2+y^2)/2) \\ \times\left[\begin{array}{c} 1 + \left\{a_1(x)\frac{1}{\nu_1} + a_1(y)\frac{1}{\nu_2}\right\} + \left\{a_2(x)\frac{1}{\nu_1^2} + a_1(x)a_1(y)\frac{1}{\nu_1\nu_2} + a_2(y)\frac{1}{\nu_2^2}\right\} \\ + \left\{a_3(x)\frac{1}{\nu_1^3} + a_2(x)a_1(y)\frac{1}{\nu_2^2\nu_2} + a_1(x)a_2(y)\frac{1}{\nu_1\nu_2^2} + a_3(y)\frac{1}{\nu_2^3}\right\} + \ldots \end{array}\right] \end{array} \quad (A3)$$

We find

$$\begin{array}{rcl} a_1(x)a_1(y) & = & \frac{1}{16}(x^4y^4 - 2x^4y^2 - x^4 - 2x^2y^4 + 4x^2y^2 + 2x^2 - y^4 + 2y^2 + 1) \\ a_2(x)a_1(y) & = & \frac{1}{384}\left(\begin{array}{l} 3x^8y^4 - 28x^6y^4 + 30x^4y^4 + 12x^2y^4 + 3y^4 - 6x^8y^2 + 56x^6y^2 \\ -60x^4y^2 - 24x^2y^2 - 6y^2 - 3x^8 + 28x^6 - 30x^4 - 12x^2 - 3 \end{array}\right) \end{array} \quad (A4)$$

and $a_1(x)a_2(y)$ is obtained from $a_2(x)a_1(y)$ by interchanging x and y. We need to evaluate, for arbitrary $c \geq 0$

$$Pr\{W^2_{\nu_1,\nu_2} \geq c^2\} = Pr\{T^2_{\nu_1} + T^2_{\nu_2} \geq c^2\} = \int\int_{x^2+y^2\geq c^2} f_{\nu_1}(x)f_{\nu_2}(y)dxdy. \quad (A5)$$

From (A3), (A4), (A5) it is apparent that one needs to determine the basic function $J_{\ell,m}(c)$, defined, for non-negative integers ℓ, m by the double integral

$$\int\int_{x^2+y^2\geq c^2} (2\pi)^{-1}\exp\left(-\frac{x^2+y^2}{2}\right)x^{2\ell}y^{2m}dxdy. \quad (A6)$$

Transforming the Cartesian co-ordinates (x, y) to polar co-ordinates (r, θ) (when the area element $dxdy$ becomes the area element $rdrd\theta$), (A6) reduces to the product of two unvariate integrals, namely

$$J_{\ell,m}(c) = \int_0^{2\pi} (2\pi)^{-1} \cos^{2\ell}\theta \sin^{2m}\theta d\theta \times \int_c^{\infty} \exp\left(-\frac{r^2}{2}\right) r^{2(\ell+m)+1} dr$$

or, on setting $u = r^2/2$,

$$J_{\ell,m}(c) = (2\pi)^{-1} 2B(\ell + 1/2, m + 1/2) \times 2^{\ell+m} \int_{c^2/2}^{\infty} \exp(-u)\, u^{\ell+m} du. \qquad (A7)$$

Now, for arbitrary non-negative integral q,

$$\Gamma\left(q + \frac{1}{2}\right) = \frac{q^*\sqrt{\pi}}{2^q} \qquad (A8)$$

where

$$q^* = \begin{cases} 1 \cdot 3 \cdots (2q-1) & \text{if } q = 1, 2, \ldots, \\ 1 & \text{if } q = 0, \end{cases} \qquad (A9)$$

so that the first term on the right of (A7) becomes

$$(2\pi)^{-1} 2 \frac{\ell^*\sqrt{\pi}}{2^\ell} \frac{m^*\sqrt{\pi}}{2^m} = \frac{\ell^* m^*}{2^{\ell+m}}. \qquad (A10)$$

Since

$$\Gamma(\ell + m + 1) = (\ell + m)! \qquad (A11)$$

(A7) reduces to

$$J_{\ell,m}(c) = \ell^* m^* \times \frac{1}{(\ell+m)!} \int_a^{\infty} \exp(-u) u^{\ell+m} du \qquad (A12)$$

where $a = c^2/2$. To evaluate the second term on the right of (A12) integrate sucessively by parts (or, alternatively, recall the well known elementary result which relates the incomplete gamma function ratio to the cumulative sum of Poisson probabilities), giving

$$J_{\ell,m}(c) = \ell^* m^* \pi_{\ell+m}(a) \qquad (A13)$$

where

$$\pi_q(a) = e^{-1}\left(1 + a + \frac{a^2}{2!} + \ldots + \frac{a^q}{q!}\right). \qquad (A14)$$

The use of (A13) and (A14) in (A3)–(A7) gives, after completely straightforward (though rather lengthy and tedious) elementary algebraic and numerical evaluations, the results (Ic) and (Id). (Recall here that $a = c^2/2 = t/2$.)

References

Barnard, G.A. (1950). On the Fisher-Behrens test, *Biometrika*, **37**, 203-207.

– – ––– (1982). A new approach to the Behrens-Fisher problem, *Utilitas Mathematica*, **XXIB**, 261-271.

– – ––– (1984). Comparing the means of two independent samples, *J. Roy. Stat. Soc.*, Series C, **33**, 266–271.

– – ––– (1995). Neyman and the Behrens-Fisher problem – an anecdote, *Probab. Math. Stat.*, **15**, 67-71.

Bartlett, M.S. (1935). The effect of non-normality on the t-distribution, *Procedings of the Cambridge Philosophical Society*, **31**, 223-231.

– – ––– (1936). The information available in small samples, *Proc. Roy. Soc.*, **32**, 560-566.

– – ––– (1937). Properties of sufficiency and statistical tests, *Proc. Roy. Soc.*, A, **160**, 266-282.

– – ––– (1954). A note on the multiplying factors for various χ^2 approximation, *J. Roy. Stat. Soc.* Series B, **16**, 296-298.

– – ––– (1956). Comment on Sir Ronald's paper 'On a test of significance in Pearson's Biometrika Tables (No.11)' *J. Roy. Stat. Soc.* Series B, **18**, 295-296.

Behrens, W.V. (1929). Ein beitrag zur Fehlerberechnung bei wenigen Beobachtungen, *Landwirischaftliche Jahrburcher*, **68**, 807-837.

Cox, D.R. and Hinkley, D.V. (1996). *Theoretical Statistics*. Chapman and Hall, London.

Fisher, R.A. (1925). Expansion of "Student's" integral in powers of n^{-1}, *Metron,* **5**, 109-112.

– – ––– (1935). The fiducial argument in statistical inference, *Annals Engenics*, **6**, 391-398.

– – ––– (1941). The asymptotic approach to Behrens' integral with further tables for the d-test of significance, *Ann. Eugenics*, **11**, 141-173.

– – ––– (1956). On a test of significance in Pearson's Biometrika tables (no.11), *J. Roy. Stat. Soc.* Series B, **18**, 56-60.

– – ––– (1966). *Design of Experiments*, 8th edition. Oliver and Boyd, Edinburgh.

– – ––– (1973). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

Fisher, R.A. and Yates, F. (1975). *Statistical Tables*, 6th. Edn. Longman, London.

Gayen, A.K. (1949). The distribution of "Student's" t in random sample of any size drawn from non-normal universes, *Biometrika*, **36**. 353-369.

Kendall, M.G. and Stuart, A. (1961). *The Advanced Theory of Statistics*, Vol. **2**. London, Griffin.

Linnik, Ya, V. (1968). *Statistical Problems with Nuisance Parameters*. Translations of mathematical monographs, No. **20**, American Mathematical Society, New York.

Neyman, J. and Pearson, E.S. (1933a). On the problem of the most efficient tests of statistical hypotheses, *Phil. Trans. Roy. Soc., Series A*, **231**, 289-337.

– – ––– (1933b). On the problem of the most efficient tests of statistical hypotheses, *in Breakthroughs in Statistics*, Vol **1**, Samuel Kotz and Normal L. Johnson, eds. (1993), 73-108. London, Griffin.vspace-.2em

Pearson E.S. and Hartley, H.O. (1970). *Biometrika Tables for Statisticians*, Vol **I**, 3rd Edn. Cambridge University Press, Cambridge.

Robinson G.K. (1976). Properties of Student's t and of the Behrens-Fisher solution to the two means problem, *Ann. Statist.*, **4** 963-971.

Ruben, H. (1950). *Sequential Studentisation of Sampling Means from Unknown Normal Populations*, Ph.D. dissertation, London University.

– – ––– (1960). On the distribution of the weighted difference of two independent Student variables *J. Roy. Stat. Soc.* Series B, **22**, 188-194.

– – ––– (1961). Studentisation of the two-stage sampling means from normal populations with unknown common variance, *Sankhyā Series A*, **23**, 231-250.

– – –– (1962)[4]. Studentisation of two-stage sample means from normal populations with unknown variances I. General theory and application to the confidence estimation and testing of the difference in population means. *Sankhyā Series A*, **24**, 157–180.

Scheffé H. (1943). On solutions of the Behrens-Fisher problem based on the t-distribution, *Ann. Math. Statist.*, **14**, 35-44.

Stuart A, Ord J.K. and Arnold S. (1999) *Kendall's Advanced Theory of Statistics.* Vol **2A**, 6th Edn. Arnold, London.

Welch, B.L. (1947). The generalisation of Student's problem when several different population variances are involved, *Biometrika*, **6**, 28-35.

– – –– (1956). Note on some criticisms made by Sir Ronald Fisher. *J. Roy. Stat. Soc.* Series B, **18**, 297-302.

Harold Ruben
35 Theypun Bower
Bower Hill
Epping
Essex, CM16 7AB, U.K.

[4]There are two printer's omisson. In (6.22), on p. 179, $1-\psi[*]-\psi[**]$ should read $1-\{\psi[*]-\psi[**]\}$ In (6.24), on p. 180, $1-\Phi[*]-\Phi[**]$ should read $1-\{\Phi[*]-\Phi[**]\}$.