The Annals of Statistics
1976, Vol. 4, No. 5, 963-971

PROPERTIES OF STUDENT'S *t* AND OF THE BEHRENS-FISHER SOLUTION TO THE TWO MEANS PROBLEM

BY G. K. ROBINSON

University College London

Conditional properties of the usual confidence intervals for the situations referred to in the title are investigated. It is shown that there can be no negatively biased relevant selections in a sense which implies that there can be no negatively biased relevant subsets in the sense of Buehler (1959). The intuitive meaning of these results is that there is no way of betting that the quoted confidence levels are too high which yields positive expected return for all parameter values. In addition it is reported that the coverage probabilities for the Behrens-Fisher intervals are always larger than the nominal significance level would suggest. Thus the Behrens-Fisher and Student's *t* procedures can be considered to be conservative.

1. Introduction. Suppose that we have a sample space Z and a parameter space Θ, typical elements of these being z and θ, respectively. We will take an interval estimator of θ to be a set function $I(z)$, which assigns a subset of Θ to each z in Z, together with a confidence level function $\alpha(z)$, which maps Z into $[0, 1]$. The value of the confidence level function $\alpha(z)$ is intended to express a level of confidence, in some sense, in the possibility that $\theta \in I(z)$. Some writers reserve the word "confidence" for interval estimators in the classical or Neyman–Pearson sense. I will not be doing this. Instead, I will be qualifying the word "confidence" whenever it refers to the ideas of a particular school of thought on statistical inference.

Buehler (1959) contributed to the formalization of a way of investigating the appropriateness of confidence interval functions for their set functions. He introduced two players, Peter and Paul, who make bets in order to question Peter's statistical procedures.

Let us think about a situation where Peter asserts that he has confidence $\alpha(z)$ that $\theta \in I(z)$ and Paul wishes to question the validity of this confidence by betting against him. We could think about betting procedures in which Paul sometimes says that $\alpha(z)$ is too high and sometimes says that it is too low, but our present concern is only with Paul betting that $\alpha(z)$ is too high. We are interested to know whether or not Paul can claim that Peter's level of confidence is too great.

A pure betting strategy for Paul would be to select a subset C of Z in which he thinks that Peter's confidence that $\theta \in I(z)$ is too high. In line with the philosophy of relevant subsets outlined by Buehler, Peter does not accept bets

Received August 1974; revised January 1976.

AMS 1970 *subject classifications*. Primary 62A99, 62F25.

Key words and phrases. Conditional confidence, negatively biased relevant subset, *t* distribution, relevant selections, semirelevant selections, type I error.

whereby he risks $\alpha(z)$ in order to win $1 - \alpha(z)$ if $\theta \in I(z)$. Rather, he only risks $\alpha(z) - \varepsilon$ to win $1 - \alpha(z) + \varepsilon$ for some $\varepsilon > 0$ which does not depend on z.

Let us use the usual characteristic function notation so that $\chi_{I(z)}(\theta)$ is 1 if $\theta \in I(z)$ and is 0 if $\theta \notin I(z)$. If

$$E[\chi_{I(z)}(\theta) - \alpha(z) + \varepsilon \mid z \in C, \theta] \leqq 0$$

for all θ then we call C a negatively biased relevant subset. (This reduces to Buehler's definition when $\alpha(z) \equiv \alpha$.) The inequality says that Peter's expected gain is negative for all θ, supporting Paul's claim that Peter's confidence is too high.

A randomized betting strategy for Paul would be to specify a selection: a function $k(z)$ from Z to the unit interval which has the interpretation that he bets with probability $k(z)$ when z is observed. This idea was introduced by Tukey (1958) and has been used by Wallace (1959) and Pierce (1973). We will call a selection a negatively biased relevant selection if

(i) it is nontrivial in the sense that $E[k(z) \mid \theta] > 0$ for some θ,
(ii) Peter loses for all θ if Paul uses it, i.e.

$$E[\{\chi_{I(z)}(\theta) - \alpha(z) + \varepsilon\}k(z) \mid \theta] \leqq 0 \qquad \text{for some} \quad \varepsilon > 0\,, \quad \text{for all} \quad \theta\,.$$

Wallace (1959) purports to prove results related to the present ones. Stein (1961) showed that his Theorems 2 and 3 are incorrect and this can be seen from the example given by Buehler and Fedderson (1963) which gives a positively biased relevant subset for a Neyman confidence interval based on the t distribution. This provides a counterexample to Wallace's Theorem 2. The complement of that confidence interval is also a Neyman confidence region, although not an interval, and for this confidence region the conditioning set that they have investigated is a negatively biased relevant subset. This is a counterexample to Wallace's Theorem 3.

Brown (1967) has extended Buehler and Fedderson's example to allow the t distribution to have more than one degree of freedom. Olshen (1973) has further extended the ideas to regression problems, but has only found semirelevant subsets (Buehler's notation again), not relevant subsets. At the other end of the line, Buehler and Fedderson's paper is itself an extension of an example of Stein (1961). Stein found a positively biased relevant selection which is rather similar to Buehler and Fedderson's relevant subset.

In order to prove the analytical results in Sections 2 and 3, we shall define $\alpha(z)$ in terms of particular improper Bayesian prior densities. That these prior densities give the same results as the t distribution and the Behrens-Fisher distribution is well known. The reader is referred to Lindley (1965), Fisher (1935) and Jeffreys (1940).

2. Result for the t distribution. Suppose that we make $n\,(\geqq 2)$ observations on a normal population with unkown mean μ and unknown variance θ. Let us denote the vector of observations by $\mathbf{x}$, the mean by $\bar{x}$ and the sample variance by s^2. Suppose that we have a set function $I(\mathbf{x})$ for μ and we quote a confidence level function $\alpha(\mathbf{x})$ which is calculated using the t distribution or, equivalently, using the usual θ^{-1} improper prior density for θ and μ.

THEOREM 1. *Provided that $I(\mathbf{x})$ is always an interval containing $\bar{x}$, there is no negatively biased relevant selection for $\alpha(\mathbf{x})$ as a confidence level function for $I(\mathbf{x})$*; i.e. *there is no function $k(\mathbf{x})$ and $\varepsilon > 0$ such that*

$$\text{(1)}\qquad \begin{aligned} &\text{(i)}\quad 0 \leqq k(\mathbf{x}) \leqq 1,\\ &\text{(ii)}\quad E[k(\mathbf{x})] > 0 \quad \textit{for some}\quad \mu \quad \textit{and some}\quad \theta, \quad \textit{and}\\ &\text{(iii)}\quad E[\{\chi_{I(\mathbf{x})}(\mu) - \alpha(\mathbf{x}) + \varepsilon)k(\mathbf{x}) \mid \mu, \theta] \leqq 0 \quad \textit{for all}\quad \mu \quad \textit{and all}\quad \theta. \end{aligned}$$

PROOF. Define $\alpha_\gamma(\mathbf{x})$ to be the Bayesian confidence level for $I(\mathbf{x})$ based on the improper prior $\theta^{-1+\gamma}$ for μ and θ where $\gamma \geqq 0$. The likelihood of θ and μ given $\mathbf{x}$ is proportional to

$$\theta^{-\frac{1}{2}n} \exp[-(1/2\theta)\{n(\bar{x} - \mu)^2 + (n - 1)s^2\}] .$$

Denoting $\{n(\bar{x} - \mu)^2 + (n - 1)s^2\}$ by S, our posterior density is proportional to

$$\theta^{-\frac{1}{2}n-1+\gamma} \exp(-S/2\theta) ;$$

so the marginal posterior density of μ can be found, by integrating out θ, to be proportional to $S^{-\frac{1}{2}n+\gamma}$ and hence to

$$\left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n+\gamma} \qquad \text{where} \quad t = \frac{\bar{x} - \mu}{s} n^{\frac{1}{2}} .$$

Denoting $\int_{-\infty}^{\infty} \{1 + t^2/(n-1)\}^p \, dt$ by $K(p)$,

$$\begin{aligned} &|\alpha_\gamma(\mathbf{x}) - \alpha(\mathbf{x})|\\ &\quad = \left| \frac{\int_{-\infty}^{\infty} \chi_{I(\mathbf{x})}(\mu)\{1 + t^2/(n-1)\}^{-\frac{1}{2}n+\gamma}\, dt}{K(\gamma - \frac{1}{2}n)} - \frac{\int_{-\infty}^{\infty} \chi_{I(\mathbf{x})}(\mu)\{1 + t^2/(n-1)\}^{-\frac{1}{2}n}\, dt}{K(-\frac{1}{2}n)} \right|\\ &\quad = \left| \int_{-\infty}^{\infty} \chi_{I(\mathbf{x})}(\mu) \left\{ \frac{\{1 + t^2/(n-1)\}^\gamma K(-\frac{1}{2}n)}{K(\gamma - \frac{1}{2}n)} - 1 \right\} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n} dt \right|\\ &\quad \leqq \int_{-\infty}^{\infty} \left| \left(1 + \frac{t^2}{n-1}\right)^\gamma K(-\tfrac{1}{2}n)/K(\gamma - \tfrac{1}{2}n) - 1 \right| \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n} dt \end{aligned}$$

which $\rightarrow 0$ as $\gamma \rightarrow 0$ by the dominated convergence theorem.

$$\text{(2)}\qquad \text{Hence}\quad \alpha_\gamma(\mathbf{x}) \rightarrow \alpha(\mathbf{x}) \text{ uniformly in } \mathbf{x} \text{ as } \gamma \rightarrow 0.$$

Now define $\beta_\gamma(\mathbf{x})$, for $\gamma > 0$, to be the Bayesian confidence level for $I(\mathbf{x})$ based upon the improper prior density $\theta^{-1+\gamma}$ truncated to the region $0 < \theta < R$, $-\infty < \mu < \infty$. The value of R is finite but otherwise of no relevance to the calculations.

Let $f_{\mathbf{x}}(\mu)$ and $g_{\mathbf{x}}(\mu)$, respectively, denote the marginal posterior densities for

μ for the nontrucated and truncated formal prior distributions. Then

$$\frac{g_{\mathbf{x}}(\mu)}{f_{\mathbf{x}}(\mu)} = \frac{\int_{S/2R}^{\infty} y^{\frac{1}{2}n-1-\gamma} e^{-y}\, dy}{\int_0^{\infty} y^{\frac{1}{2}-1-\gamma} e^{-y}\, dy}$$

where S denotes $n(\bar{x} - \mu)^2 + (n - 1)s^2$, as before. For fixed $\bar{x}$ and s^2, S is increasing as a function of $|\bar{x} - \mu|$ so the ratio of densities decreases with increasing $|\bar{x} - \mu|$. In other words, the densities have decreasing monotone likelihood ratio in $|\bar{x} - \mu|$. It follows that

$$(3) \qquad \beta_\gamma(\mathbf{x}) = \int_{I(\mathbf{x})} g_{\mathbf{x}}(\mu)\, d\mu \geqq \int_{I(\mathbf{x})} f_{\mathbf{x}}(\mu)\, d\mu = \alpha_\gamma(\mathbf{x})$$

for any interval $I(\mathbf{x})$ containing the point $\mu = \bar{x}$.

Finally, define $\beta_\gamma{}^m(\mathbf{x})$ to be the confidence level for $I(\mathbf{x})$ based on the proper prior density for θ and μ proportional to

$$\theta^{-1+\gamma}(1 + |\mu/m|\}^{-2}$$

over $-\infty < \mu < \infty$, $0 < \theta < R$. Again $\gamma > 0$ and R is finite. We shall make use of the inequalities

$$\begin{aligned} &\{1 + |\bar{x}/m|\}^{-2}\{1 - (2/m)|\bar{x} - \mu|\} \\ &\qquad \leqq \{1 + |\mu/m|\}^{-2} \\ &\qquad \leqq \{1 + |\bar{x}/m|\}^{-2}\{1 + (2/m)|\bar{x} - \mu| + (3/m^2)|\bar{x} - \mu|^2\} \end{aligned}$$

which we shall discuss in the form:

$$\{1 + |a|\}^{-2}(1 - 2|a - b|) \leqq \{1 + |b|\}^{-2} \leqq \{1 + |a|\}^{-2}\{1 + 2|a - b| + 3|a - b|^2\}$$

for all real a and b.

It is straightforward to prove the left-hand inequality by looking at the tangent to the curve $\{1 + |b|\}^{-2}$ at $b = a$. If $|a| \leqq |b|$ the right-hand inequality is trivial. We may as well suppose that a and b are of the same sign so it is sufficient to look at the case $0 < b < a$. The inequality becomes

$$(1 + a)^2 \leqq (1 + b)^2\{1 + 2(a - b) + 3(a - b)^2\}$$

which is established by showing that

$$\begin{aligned} &(1 + b)^2\{1 + 2(a - b) + 3(a - b)^2\} - (1 + a)^2 \\ &\qquad = 2a(a - b) + 2b(3a^2 - 5ab + 4b^2) + 3b^2(a - b)^2 \geqq 0\,. \end{aligned}$$

Now, using $f(\mathbf{x};\, \mu, \theta)$ to denote the probability density function of $\mathbf{x}$,

$$(4) \qquad \begin{aligned} \frac{\beta_\gamma{}^m(\mathbf{x})}{\beta_\gamma(\mathbf{x})} &= \frac{\int_0^R \int_{I(\mathbf{x})} \{1 + |\mu/m|\}^{-2}\theta^{-1+\gamma} f(\mathbf{x};\, \mu, \theta)\, d\mu\, d\theta}{\int_0^R \int_{-\infty}^{\infty} \{1 + |\mu/m|\}^{-2}\theta^{-1+\gamma} f(\mathbf{x};\, \mu, \theta)\, d\mu\, d\theta} \\ &\quad \times \frac{\int_0^R \int_{-\infty}^{\infty} \theta^{-1+\gamma} f(\mathbf{x};\, \mu, \theta)\, d\mu\, d\theta}{\int_0^R \int_{I(\mathbf{x})} \theta^{-1+\gamma} f(\mathbf{x};\, \mu, \theta)\, d\mu\, d\theta} \\ &\leqq \frac{\int_0^R \int_{I(\mathbf{x})} \theta^{-1+\gamma}\{1 + (2/m)|\bar{x} - \mu| + (3/m^2)|\bar{x} - \mu|^2\} f(\mathbf{x};\, \mu, \theta)\, d\mu\, d\theta}{\int_0^R \int_{I(\mathbf{x})} \theta^{-1+\gamma} f(\mathbf{x};\, \mu, \theta)\, d\mu\, d\theta} \\ &\quad \times \frac{\int_0^R \int_{-\infty}^{\infty} \theta^{-1+\gamma} f(\mathbf{x};\, \mu, \theta)\, d\mu\, d\theta}{\int_0^R \int_{-\infty}^{\infty} \theta^{-1+\gamma}\{1 - (2/m)|\bar{x} - \mu|\} f(\mathbf{x};\, \mu, \theta)\, d\mu,\, d\theta}\,. \end{aligned}$$

With k being 1 or 2,

$$\frac{\int_{I(\mathbf{x})} |\bar{x} - \mu|^k f(\mathbf{x}; \mu, \theta)\, d\mu}{\int_{I(\mathbf{x})} f(\mathbf{x}; \mu, \theta)\, d\mu} \leqq \frac{\int_{-\infty}^{\infty} |\bar{x} - \mu|^k f(\mathbf{x}; \mu, \theta)\, d\mu}{\int_{-\infty}^{\infty} f(\mathbf{x}; \mu, \theta)\, d\mu}$$

since $I(\mathbf{x})$ is an interval containing $\bar{x}$. Since $\theta < R$ and each of these last two integrals is independent of $\mathbf{x}$, (4) gives an upper bound for $\beta_\gamma{}^m(\mathbf{x})/\beta_\gamma(\mathbf{x})$ which tends to unity uniformly in $\mathbf{x}$ as m tends to infinity.

We could similarly obtain a lower bound for $\beta_\gamma{}^m(\mathbf{x})/\beta_\gamma(\mathbf{x})$ which tends to 1 uniformly in $\mathbf{x}$ as m tends to ∞. Since $\beta_\gamma(\mathbf{x}) \leqq 1$,

$$(5) \qquad \beta_\gamma{}^m(\mathbf{x}) \to \beta_\gamma(\mathbf{x}) \qquad \text{uniformly in} \quad \mathbf{x} \quad \text{as} \quad m \to \infty\,.$$

Suppose that for some $\varepsilon > 0$ there is a selection $k(\mathbf{x})$ as described in the statement of the theorem. From (2), (3) and (5)

$$\liminf_{(m\to\infty,\gamma\to 0)} \beta_\gamma{}^m(\mathbf{x}) \geqq \alpha(\mathbf{x}) \qquad \text{uniformly in} \quad \mathbf{x}\,.$$

Therefore there are values of m and γ such that

$$\beta_\gamma{}^m(\mathbf{x}) \geqq \alpha(\mathbf{x}) - \tfrac{1}{2}\varepsilon \qquad \text{for all} \quad \mathbf{x}\,.$$

So, using (1),

$$(6) \qquad E[\{\chi_{I(\mathbf{x})}(\mu) - \beta_\gamma{}^m(\mathbf{x}) + \tfrac{1}{2}\varepsilon\}k(\mathbf{x}) \mid \mu, \theta] \leqq 0\,.$$

But the prior expectation of $E[\{\chi_{I(\mathbf{x})}(\mu) - \beta_\gamma{}^m(\mathbf{x})\}k(\mathbf{x}) \mid \mu, \theta]$ according to the proper prior upon which $\beta_\gamma{}^m(\mathbf{x})$ is based is zero. Therefore the prior expectation of

$$E[\{\chi_{I(\mathbf{x})}(\mu) - \beta_\gamma{}^m(\mathbf{x}) + \tfrac{1}{2}\varepsilon\}k(\mathbf{x}) \mid \mu, \theta]$$

is strictly positive in contradiction with (6). This proves the theorem.

3. Result for the Behrens-Fisher solution to the two means problem. We observe a sample of size n_1 ($\geqq 2$) from a first normal population and a sample of size n_2 ($\geqq 2$) from a second. Let $\mathbf{x}$, $\mathbf{y}$, $\bar{x}$, $\bar{y}$, s_1^2, s_2^2, μ_1, μ_2, θ_1 and θ_2 denote the vectors of observed values, the observed means and variances and the population means and variances.

Suppose that we have a set function $I(\mathbf{x}, \mathbf{y})$ for $\delta = \mu_1 - \mu_2$ which we quote at confidence level $\alpha(\mathbf{x}, \mathbf{y})$ based on the fiducial argument or, equivalently, on a $(\theta_1\theta_2)^{-1}$ improper prior density for μ_1, μ_2, θ_1 and θ_2.

THEOREM 2. *Provided that* $I(\mathbf{x}, \mathbf{y})$ *is always an interval containing the point* $\bar{y} - \bar{x}$ *there is no negatively biased relevant selection.*

OUTLINE OF PROOF. This theorem could be proved in a similar manner to Theorem 1. We shall omit some of the steps here.

First define $\alpha_\gamma(\mathbf{x}, \mathbf{y})$, for $\gamma \geqq 0$, to be the confidence level for $I(\mathbf{x}, \mathbf{y})$ based on the improper prior density $(\theta_1\theta_2)^{-1+\gamma}$ for μ_1, μ_2, θ_1 and θ_2. We could show, analogously to the proof of Theorem 1, that

$$(7) \qquad \alpha_\gamma(\mathbf{x}, \mathbf{y}) \to \alpha(\mathbf{x}, \mathbf{y}) \qquad \text{uniformly as} \quad \gamma \to 0\,.$$

Then define $\beta_\gamma(\mathbf{x}, \mathbf{y})$, for $\gamma > 0$, to be the confidence level for $I(\mathbf{x}, \mathbf{y})$ based on the improper prior density $(\theta_1\theta_2)^{-1+\gamma}$ truncated to $0 < \theta_1 < R$, $0 < \theta_2 < R$. We wish to prove that $\beta_\gamma(\mathbf{x}, \mathbf{y}) \geqq \alpha_\gamma(\mathbf{x}, \mathbf{y})$.

Using a result from the proof of Theorem 1, we have that the marginal posterior distribution for μ_1 for the truncated formal prior is more concentrated about $\bar{x}$ than is the posterior for the untruncated prior in the sense that the densities of these distributions have decreasing monotone likelihood ratio in $|\bar{x} - \mu|$. Similarly, the marginal posterior distribution of μ_2 for the truncated prior is more concentrated about $\bar{y}$ than that for the untruncated prior. This implies that the marginal posterior distribution of $\delta = \mu_2 - \mu_1$ for the truncated prior has decreasing monotone likelihood ratio in $|\mu_2 - \mu_1 - (\bar{y} - \bar{x})|$ with respect to its distribution for the untruncated prior. (See Karlin and Proschan (1960).)

Therefore, since $I(\mathbf{x}, \mathbf{y})$ is an interval containing the point $\delta = \bar{y} - \bar{x}$, the proportion of the posterior for the truncated prior within $\delta \in I(\mathbf{x}, \mathbf{y})$ is greater than the proportion of the posterior for the untrucated prior, i.e.

$$(8) \qquad \beta_\gamma(\mathbf{x}, \mathbf{y}) \geqq \alpha_\gamma(\mathbf{x}, \mathbf{y})\,.$$

Finally, define $\beta_\gamma^m(\mathbf{x}, \mathbf{y})$ to be the confidence level for $I(\mathbf{x}, \mathbf{y})$ based on the proper prior density proportional to

$$(\theta_1\theta_2)^{-1+\gamma}\{1 + |\mu_1/m|\}^{-2}\{1 + |\mu_2/m|\}^{-2}$$

over $0 < \theta_1 < R, 0 < \theta_2 < R$, $-\infty < \mu_1 < \infty$, $-\infty < \mu_2 < \infty$. That $\beta_\gamma^m(\mathbf{x}, \mathbf{y})$ tends to $\beta_\gamma(\mathbf{x}, \mathbf{y})$ uniformly in $\mathbf{x}$ and $\mathbf{y}$ as m tends to infinity for fixed $\gamma > 0$ can be proved in a similar manner to the corresponding result in the proof of Theorem 1.

Now (7), (8) and the last assertion are completely analogous to (2), (3) and (5) in the proof of Theorem 1, so the remaining argument of the proof is the same as that of Theorem 1. Similar notation has been used to enable this to be readily seen.

4. The coverage properties of Behrens-Fisher intervals. Bartlett (1936) was the first to show that the confidence regions of the Behrens-Fisher solution do not cover the true parameter value with probability the nominal confidence level.

A problem equivalent to that of finding coverage probabilities is that of finding type I error rates for the associated tests of the hypothesis $\delta = 0$. Such probabilities of error have been calculated by James (1959), Mehta and Srinivasan (1970) and Wang (1971). James (1959) investigated only the case $n_1 = n_2 = 2$ for which it is easy to find the significance points. Mehta and Srinivasan (1970) looked at three cases: $n_1 = n_2 = 4$; $n_1 = 4$ and $n_2 = 20$; and $n_1 = n_2 = 20$. However they used Fisher's (1941) asymptotic expansion to find their significance points, so only for $n_1 = n_2 = 20$ can their probabilities of type I error be considered reliable. Wang (1971) found probabilities of type I error for the case

$n_1 = 7$, $n_2 = 13$. Presumably he used Sukhatme's (1938) significance points and interpolated between them. In comparison with a table of type I error rates which I have calculated his table appears to contain slight inaccuracies which seem to be attributable to the inaccuracy of such interpolation.

Kempthorne (1966) performed a Monte Carlo experiment to investigate the probabilities of type I error but it was too small to be useful.

These calculations all tended to support the idea that the type I error probabilities are alwavs less than the nominal significance level. I have recently tabulated the probabilities of type I error for one-sided significance levels 0.1, 0.05, 0.025, 0.01, 0.005, 0.001, 0.0005 and 0.0001; for n_1 and n_2 belonging to the set $\{2, 3, 4, 5, 6, 7, 8, 10, 12, 14, 18, 24, 32, 50, 100, \infty\}$ and $\theta_1 n_2/\theta_2 n_1$ belonging to the set $\{\frac{1}{1000}, \frac{1}{100}, \frac{1}{30}, \frac{1}{10}, \frac{1}{5}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1, \frac{3}{2}, 2, 3, 5, 10, 30, 100, 1000\}$. In all cases the probability of type I error was found to be less than the nominal value. The accuracy of the calculation varies but I consider that sufficient accuracy of calculation and fineness of tabulation has been achieved to infer with reasonable certainty that the type I error rate is always less than the nominal significance level for Behrens–Fisher tests.

The actual calculation has little intrinsic interst. Firstly, significance points were found using either Sukhatme's method involving numerical integration or using a numerical adaption of Fisher and Healy's (1956) method. This part of the calculation was accurate to at least 10 significant figures. Secondly, the significance points were fitted by polynomials in the angle, usually denoted θ, upon which they depend. Thirdly the probabilities of type I error were found by using numerical integration like that used by Wang (1971). Approximate significance points were calculated from the fitted polynomials to greatly speed this step.

5. Should we use the Behrens–Fisher solution to the two means problem in statistical practice? The two means problem has long been of interest as a testing ground for theories of statistical inference. Consequently, whether or not one believes that the two means problem is an important one from a practical point of view, it is important to decide what statistical procedure we should use in a two means situation.

The Behrens–Fisher solution to the two means problem has two properties which seem very desirable:

(i) Its confidence regions for δ cover the true value of δ with probability always larger than the nominal confidence level.

(ii) There are no negatively biased relevant selections.

Thus it seems to yield procedures which are conservative in an intuitive sense. I would like to propose that the word "conservative" be used in a technical sense to mean that these two properties hold.

As was remarked by Buehler (1959), Fisher (1956) shows that there is a negatively biased relevant subset for a Neyman confidence interval statement devised by Welch (1947) for the two means problem. I agree with Fisher (1956), Buehler (1959) and Wallace (1959) that the existence of such a subset seems to undermine the validity of a confidence statement. Certainly Welch's test is not conservative in the technical sense suggested above.

Perhaps the Behrens–Fisher test is optimal in some sense amongst the class of procedures which are conservative. I, personally, suspect that it will eventually be regarded as the correct test to use except when a proper Bayesian test is considered appropriate. I believe that research towards finding a test with approximately constant probability of type I error is misguided. In the two means situation it seems that such a test would inevitably have unacceptable conditional properties.

Acknowledgments. A result similar to Theorem 1 was proved by Charles Stein in the second of the Wald Lectures for 1961. The proof here is presented with his permission.

The computing referred to in Section 4 was done using the facilities of the University of London and University College London computing centres.

The constructive criticism of two referees has been very helpful.

REFERENCES

BARTLETT, M. S. (1936). The information available in small samples. *Proc. Cambridge Philos. Soc.* **32** 560–566.

BROWN, L. (1967). The conditional level of Student's *t* test. *Ann. Math. Statist.* **38** 1068–1071.

BUEHLER, R. J. (1959). Some validity criteria for statistical inferences. *Ann. Math. Statist.* **30** 845–863.

BUEHLER, R. J. and FEDDERSON, A. P. (1963). Note on a conditional property of Student's *t*. *Ann. Math. Statist.* **34** 1098–1100.

FISHER, R. A. (1935). The fiducial argument in statistical inference. *Ann. Eugenics* **6** 391–398.

FISHER, R. A. (1941). The asymptotic approach to Behrens' integral with further tables for the *d*-test of significance. *Ann. Eugenics* **11** 141–172.

FISHER, R. A. (1956). On a test of significance in Pearson's Biometrika Tables (No. 11). *J. Roy. Statist. Soc. Ser. B* **18** 56–60.

FISHER, R. A. and HEALY, M. J. R. (1956). New tables of Behrens' test of significance. *J. Roy. Statist. Soc. Ser. B* **18** 212–216.

JAMES, G. S. (1959). The Behrens-Fisher distribution and weighted means. *J. Roy. Statist. Soc. Ser. B* **21** 73–90.

JEFFREYS, H. (1940). Note on the Behrens-Fisher formula. *Ann. Fugenics* **10** 48–51.

KARLIN, S. and PROSCHAN, F. (1960). Pólya type distributions of convolutions. *Ann. Math. Statist.* **31** 721–736.

KEMPTHORNE, O. (1966). Some aspects of experimental inference. *J. Amer. Statist. Assoc.* **61** 11–34.

LINDLEY, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference.* Cambridge Univ. Press.

MEHTA, J. S. and SRINIVASAN, R. (1970). On the Behrens-Fisher problem. *Biometrika* **57** 649–655.

OLSHEN, R. A. (1973). The conditional level of the *F*-test. *J. Amer. Statist. Assoc.* **68** 692–698.

PIERCE, D. A. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.* **1** 241–250.

STEIN, C. (1961). Approximation of prior measures by probability measures. *Inst. Math. Statist.* Wald Lectures. Unpublished manuscript.

SUKHATME, P. V. (1938). On Fisher and Behrens test of significance for the difference in means of two normal samples. *Sankhyā* **4** 39–48.

TUKEY, J. W. (1958). Fiducial inference. *Inst. Math. Statist.* Wald Lectures. Unpublished manuscript.

WALLACE, D. L. (1959). Conditional confidence level properties. *Ann. Math. Statist.* **30** 864–876.

WANG, Y. Y. (1971). Probabilities of type I errors of the Welch tests for the Behrens-Fisher problem. *J. Amer. Statist. Assoc* **66** 605-608.

WELCH, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika* **34** 28-35.

AUSTRALIAN ROAD RESEARCH BOARD
P.O. BOX 156 (BAG 4)
NUNAWADING 3131
VICORIA, AUSTRALIA