stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 26/73 DECEMBER

A.H. THOMASSE

PRACTICAL RECIPES TO SOLVE THE BEHRENS-FISHER PROBLEM

Prepublication

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS (MOS) subject classification scheme (1970): 62F05, 62F25

Contents

Practical recipes to solve the Behrens-Fisher problem *)

A.H. Thomasse **)

Abstract

The comparison of the unknown means of two populations with unknown variances is called the Behrens-Fisher problem, if the populations are assumed to be normal and the ratio of the variances is not known. In this paper a summary of recipes is given to solve this problem in practice, as published in the past 35 years by Banerjee, Fisher & Behrens, Pagurova, Wald & Hâjek, Welch, Welch & Aspin, together with two large sample solutions and one solution often used as an approximate one without justification.

The solutions are presented mainly in terms of confidence intervals for the difference of the population means. Some remarks are made concerning the lengths of these intervals and the power of the corresponding tests. The solutions in this paper are dependent on the means and the variances of samples drawn from the two populations only. All solutions discussed, except the disqualified approximate one, are robust against violations of the normality assumptions with respect to the populations and they provide, at least asymptotically, good measures for the difference of the population means if the samples are drawn from populations whatsoever with finite second moment.

*) This paper is not for review; it is meant for publication in a journal.

**) At the moment: Gemeente Energie Bedrijf, Rotterdam.

1.1. Introduction

The comparison of the unknown means μ_1 and μ_2 (say) of two populations with unknown variances σ_1^2 and σ_2^2 (say) is called the Behrens-Fisher problem, if the populations are assumed to be normal and the ratio $\lambda = \sigma_1^2/\sigma_2^2$ is unknown. Below some recipes are given to solve this problem in practice, as published by various authors in the past 35 years. The solutions are presented mainly in terms of confidence intervals for the difference $\delta = \mu_1-\mu_2$ of the population means; tests for the null hypothesis $H_0{:}\delta=\delta_0$ associated in a natural way with them may be derived by accepting H_0 if and only if the corresponding interval covers δ_0. Both intervals and tests are given in the two-sided symmetrical forms; from those the one-sided and asymmetrical forms can be derived by obvious modifications.

1.2. Notations and other preliminaries

Suppose two samples have been drawn independently, one from each normal population. Throughout this article for the sample from the i-th population with parameters $\mu_i,\sigma_i^2(i=1,2)$,n_i denotes the sample size, $\underline{\bar{x}}_i$ the sample mean and $\underline{s}_i^2$ the sample variance with denominator $f_i=n_i-1$ degrees of freedom (d.f.). The statistics $\underline{\bar{x}}_1,\underline{\bar{x}}_2,\underline{s}_1^2,\underline{s}_2^2$ are independently distributed: $\underline{\bar{x}}_i$ is $N(\mu_i,m_i\sigma_i^2)$, with $m_i=n_i^{-1}$, and $f_i\underline{s}_i^2/\sigma_i^2$ is $\chi^2(f_i)$ $(i=1,2)$, where $N(\mu,\tau^2)$ denotes a normal distribution with mean μ and variance τ^2 and $\chi^2(f)$ denotes a chi-square distribution with f d.f.[4]. All solutions of the Behrens-Fisher problem considered below are dependent on $\underline{\bar{x}}_1,\underline{\bar{x}}_2,\underline{s}_1^2,\underline{s}_2^2$ only.

The following (further) notation is used in this article:

(1.2.1) $\delta = \mu_1 - \mu_2, \quad \sigma^2 = m_1\sigma_1^2 + m_2\sigma_2^2, \quad \gamma = m_1\sigma_1^2/\sigma^2;$

(1.2.2) $\underline{d} = \underline{\bar{x}}_1 - \underline{\bar{x}}_2, \quad \underline{s}^2 = m_1\underline{s}_1^2 + m_2\underline{s}_2^2, \quad \underline{c} = m_1\underline{s}_1^2/\underline{s}^2;$

(1.2.3) $f_0 = \min(f_1,f_2), \quad f_3 = f_1 + f_2, \quad \gamma_1 = f_1/f_3.$

In the case σ^2 is known the best solutions of the problem of comparing the two mean values [8] are based on the statistic $(\underline{d}-\delta)/\sigma$, which is

$N(0,1)$. Actually, σ^2 is not known, and then it seems rather natural to try to solve the problem with the help of

$$(1.2.4) \qquad \underline{v}_\delta = (\underline{d}-\delta)/\underline{s},$$

the so-called Behrens-Fisher statistic. The distribution of $\underline{v}_\delta$ depends on λ or, equivalently, since the sample sizes are fixed, on γ. It is not an easy one, but the following inequalities for it can easily be derived [7], [11], [13], [20]:

$$(1.2.5) \qquad P\{|\underline{t}(f_3)|\geq t\} \leq P\{|\underline{v}_\delta|\geq t|\gamma\} \leq P\{|\underline{t}(f_0)|\geq t\}, \qquad \gamma\in[0,1],$$

for $t\geq 0$ and for all values of δ, where $\underline{t}(f)$ denotes a Student's variate with f d.f..

Since $\underline{v}_\delta$ has a Student's distribution with f_2,f_3,f_1 d.f. at the points $\gamma=0,\gamma_1,1$ respectively, the inequalities (1.2.5) cannot be sharpened for constant t. However, if we take instead of the constant t, the function

$$(1.2.6) \qquad b(\underline{c}) = \{t^2_{\alpha/2}(f_1)\underline{c} + t^2_{\alpha/2}(f_2)(1-c)\}^{\frac{1}{2}},$$

where $t_\alpha(f)$ denotes the $(1-\alpha)$-quantile of the Student's distribution with f d.f., it may be proved [2] that

$$(1.2.7) \qquad P\{|\underline{v}_\delta|\geq b(\underline{c})|\gamma\} \leq \alpha, \qquad \gamma\in[0,1],$$

where α is a constant between 0 and 1. The inequality (1.2.7) is fairly more sharp than the inequality on the right-hand side of (1.2.5))) with $t=t_{\alpha/2}(f_0)$, if the sample sizes differ a great amount; they are the same in the case of equal sample sizes.

Since the distribution of $\underline{v}_\delta$ depends on γ, so the (real) probability of type I error of a test for $H_0:\delta=\delta_0$ based on $\underline{v}_\delta$ does. Throughout this article I denote it by $\alpha(\gamma)$. Correspondingly, the probability that the confidence interval for δ associated with a test for H_0 covers the true value of δ equals $1-\alpha(\gamma)$. If α denotes the (nominal) size of a test for

H_0 in the rest of this article, and in agreement with that $1-\alpha$ is the confidence coefficient of the corresponding interval for δ, the following relations are desirable of course:

$$(1.2.8) \qquad \alpha(\gamma) \leq \alpha, \qquad \gamma\in[0,1],$$

$$(1.2.9) \qquad \lim_{f_0\to\infty} \alpha(\gamma) = \alpha, \qquad \gamma\in[0,1],$$

for a given solution. Furthermore it is desirable to make

$$(1.2.10) \qquad \max_{\gamma\in[0,1]} |\alpha-\alpha(\gamma)|$$

as small as possible.

Relation (1.2.8) does not hold in all but two of the given recipes. The latter are considered in section 2.1. Although the magnitude of the maximum in (1.2.10) is important for these solutions, it becomes extremely important of course in the case of solutions which do not obey relation (1.2.8). From those I give two recipes in section 3, for which the maximum in (1.2.10) is relatively small and which are satisfactory practical solutions, especially for samples of intermediate size. Two well-known large sample solutions are considered in section 2.2. For all these six solutions relation (1.2.9) holds. In section 4 I mention briefly a solution, which does not obey relation (1.2.9) but which is heavily used in practice nevertheless, together with two other solutions, which are proposed in the literature but cannot be termed to be really satisfactory in practice, although they have certain theoretical advantages according to the various authors.

The reader who is interested in the theory of the Behrens-Fisher problem is referred to the literature, which has been listed at the end of this article and the literature derived from it.

2.1. Valid solutions

The inequality (1.2.7) suggests the interval solution

$$(2.1.1)\qquad |\delta-\underline{d}| \le b(\underline{c})\underline{s}$$

where $b(\underline{c})$ is defined in (1.2.6). I call it the *Banerjee solution*, because BANERJEE [2] first proved the inequality (1.2.7) fundamental to it and discussed it extensively. In practical applications the form (2.1.1) may be replaced by the equivalent form, due to Banerjee

$$(2.1.2)\qquad |\delta-\underline{d}| \le \{t^2_{\alpha/2}(f_1)m_1\underline{s}_1^2 + t^2_{\alpha/2}(f_2)m_2\underline{s}_2^2\}^{\frac{1}{2}}.$$

The inequality on the right of (1.2.5) suggests the interval solution

$$(2.1.3)\qquad |\delta-\underline{d}| \le t_{\alpha/2}(f_0)\underline{s}.$$

I call it the *Wald-Hājek solution*; WALD [19] discussed it for $n_1=n_2$ and HĀJEK [7] gave a very detailed account of it. Also SCHEFFÉ [18] made some remarks about it. The Banerjee solution and the Wald-Hājek solution coincide if the sample sizes are equal.

For both intervals (2.1.2) and (2.1.3) relation (1.2.8) holds, so that they are valid in the sense that their probability of covering δ is at least equal to the confidence coefficient for $\gamma\in[0,1]$. Since Student's distribution is asymptotically N(0,1) relation (1.2.9) follows for the Wald-Hājek solution from the left inequality of (1.2.5); however, if $f_0=f_1$ is fixed, then

$$\lim_{f_2\to\infty} P\{|\underline{t}(f_3)|\ge t_{\alpha/2}(f_1)\} = P\{|\underline{u}|\ge t_{\alpha/2}(f_1)\} < \alpha,$$

where $\underline{u}$ is N(0,1), and hence under this condition relation (1.2.9) is not necessarily true for all $\gamma\in[0,1]$ in case of the Wald-Hājek solution. Analogous statements can be proved with respect to the Banerjee solution.

As a consequence of the strict decrease of $t_\alpha(f)$ as a function of f it may be derived from the left inequality of (1.2.5) that for fixed f_3 the maximum in (1.2.10) is as small as possible for interval (2.1.3) if $f_0=[f_3/2]$, where [a] denotes the greatest integer at most equal to a [18]; the same conclusion holds for interval (2.1.2) [2]. For the same reason the length of the Banerjee interval is with probability 1 at most equal to the length of the Wald-Hâjek interval; hence the power of the Banerjee test is at least equal to the power of the Wald-Hâjek test against all alternatives.

The reader might be confused somewhat by the introduction of the Wald-Hâjek solution in addition to the Banerjee solution in this section: the properties of the latter mentioned in the last two paragraphs are at least as good as those of the former. However, some good reasons exist for this introduction. First of all it is desirable for both solutions to be based on samples of equal sizes (or nearly so), as indicated above, and in that case they coincide (or do so almost). Secondly, for all sample sizes, the Wald-Hâjek solution can be used quicker and easier than the Banerjee solution. In the third place, some statisticians are interested, in the case of tests, in the real probability of exceeding the actual value of the test statistic as obtained from the samples drawn. For the Wald-Hâjek test*) it is clear from inequalities (1.2.5) that the greatest lower bound (g.l.b.) and the lowest upper bound (l.u.b.), with respect to $\gamma \epsilon [0,1]$, of this tail probability, as it may be called, may be obtained by interpolating for the actual value of $\underline{v}_\delta$ in a table of two-sided percentage points of the Student's distribution with f_3 and f_0 d.f. respectively (for example, table 12 in [14]). A better approximation of the tail probability of the value of $\underline{v}_\delta$ can be acquired by interpolating for this value in a table of the cumulative distribution function (c.d.f.) (or probability integral) of

*) The same remark as made here for the Wald-Hâjek test, is true of course for every test for H_0: $\delta=\delta_0$ based solely on $\underline{v}_\delta$ as test statistic.

the same Student's distributions (for example, table 9 in [14]): if the tables don't have an entry for f_0 or f_3 d.f. four point interpolation may be needed.

From the left inequality of (1.2.5) it follows for the Wald-Hājek test that g.l.b. $\alpha(\gamma) = P\{|\underline{t}(f_3)| \geq t_{\alpha/2}(f_0)\}$. The analogous probability in case of the Banerjee test is not easily computed, but the g.l.b. $\alpha(\gamma)$ just now obtained for the Wald-Hājek test is of course a lower bound for it. It may be noted that the l.u.b. $\alpha(\gamma)=\alpha$ for both solutions.

2.2. Large sample solutions

The two solutions given in section 2.1 are the only ones I have found which can be termed as practical and, moreover, which are valid. For a discussion of the possibilities (or rather impossibilities) to find valid (similar) solutions of the Behrens-Fisher problem, which are useful in practice for samples of all sizes, the reader is referred to LINNIK [9]. From now on I concentrate on solutions which do not obey relation (1.2.8); hence they are at best approximately valid.

Since $\underline{s}^2$ is an unbiased consistent estimator of σ^2, it is clear that $\underline{v}_\delta$ is asymptotically N(0,1) ($f_0 \to \infty$), for $\gamma \in [0,1]$ and all values of δ. Hence a large sample interval for δ is given by

$$(2.2.1) \qquad |\delta - \underline{d}| \leq u_{\alpha/2}\underline{s},$$

where u_α denotes the $(1-\alpha)$ quantile of the N(0,1) distribution.

The left inequality of (1.2.5) suggests the large sample interval for δ

$$(2.2.2) \qquad |\delta - \underline{d}| \leq t_{\alpha/2}(f_3)\underline{s},$$

which is somewhat more conservative than (2.2.1), since $u_\alpha < t_\alpha(f)$ for all d.f. f and all sizes α with $0 < \alpha < \frac{1}{2}$. For both solutions based

on confidence intervals (2.2.1) and (2.2.2) for δ, it follows directly that $\alpha(\gamma) \geq \alpha$ for $\gamma \epsilon [0,1]$, so that these solutions are invalid. For large samples however, say $f_0 \geq 50$, the difference between $\alpha(\gamma)$ and α becomes rather small. Clearly an analogous remark can be made for the valid solutions of section 2.1, but these solutions are rather obscure in statistical practice, and they are not used as large sample solutions, in distinction of the ones discussed in this section.
Of course both solutions based on intervals (2.2.1) and (2.2.2) obey relation (1.2.9). It may be proved that the maximum in (1.2.10) for solution (2.2.1) is greater than for solution (2.2.2), and that both are minimal if $f_0 = [f_3/2]$. The length of interval (2.2.1) is smaller than that of interval (2.2.2), which in turn is smaller than the length of the Banerjee interval (2.1.2); all those inequalities hold with probability 1 and thus the power of the tests based on these intervals, neglecting their possible invalidity, gets smaller according to the increasing length of the intervals.

3. Approximate solutions for samples of intermediate sizes

In this section I give two recipes of approximate solutions, which are satisfactory in practice, if not good, for problems with samples of intermediate sizes, between 7 and 50 say; they are described below. For problems with samples of small sizes, say 6 or less, I am not aware of (approximate) solutions, based on classical probability theory, which really deserve the qualifications to be satisfactory or good in practice; hence for small sample problems only the valid solutions of section 2.1 remain at the moment. However, the reader might judge the approximate solutions of this section to satisfy his needs in these cases too.

3.1. Welch's APDF solution

Inequalities (1.2.5) suggest a possibility to approximate the distribution of $\underline{v}_\delta$ by a Student's distribution with a suitable number ϕ of d.f..

This approximation has been made by WELCH [22]: he approximated $\underline{s}^2$ as $k\sigma^2\underline{\chi}^2(\phi)/\phi$, where k is a constant, by equating the first two moments of these statistics. Welch found, that k=1 and

$$(3.1.1) \qquad \phi = \{\gamma^2/f_1 + (1-\gamma)^2/f_2\}^{-1}.$$

So $\underline{v}_\delta$ is approximately distributed as $\underline{t}(\phi)$, where of course ϕ is not an integer in general. Replacing γ by $\underline{c}$ in (3.1.1) gives an estimator $\underline{F}$ of ϕ as

$$(3.1.2) \qquad \underline{F} = \{\underline{c}^2/f_1 + (1-\underline{c})^2/f_2\}^{-1}.$$

If $t_\alpha(\underline{F})$ denotes the $(1-\alpha)$ quantile of a Student's distribution with $\underline{F}$ d.f. (obtained by interpolating for F in a table of two-sided percentage points of the Student's distribution, where F denotes the actual value of $\underline{F}$ as calculated from the sample statistics), *Welch's approximate degrees of freedom solution*, or *APDF solution* for short, can be given in the form of an interval solution by

$$(3.1.3) \qquad |\delta-\underline{d}| \leq t_{\alpha/2}(\underline{F})\underline{s},$$

where $\underline{F}$ is defined by (3.1.2).

Since $\lim_{f_0\to\infty}\underline{F}=\infty$ with probability 1 for $y\in[0,1]$, relation (1.2.9) holds for the APDF solution, but relation (1.2.8) does not: the APDF solution is not a valid one. To get an impression of the magnitude of the maximum in (1.2.10) numerical verifications of $\alpha(\gamma)$ are needed in case of the APDF solution. Such verifications are reported in [20] and [22] for d.f. between 4 and 20 and α=.10, .05, .01 and various values of γ. In the examples in [20] it turns out that $|\alpha(\gamma)-\alpha|\leq.0020$ for α=.10 (with the maximum deviation for f_1=4, f_2=8),$\leq$.0028 for α=.05 (maximum deviation for f_1=4, f_2=20) and $\leq$.0035 for α=.01 (maximum deviation for f_1=4, f_2=20) (values correct within .0001). In the example reported in [22] it is claimed that $|\alpha(\gamma)-\alpha|\leq.004$ in the case $f_1=f_2=6$, α=.10.

From the values given in [20] one gets the further impression that $\max_{\gamma\in[0,1]}|\alpha(\gamma)-\alpha|$ becomes smaller with increasing f_0, and that $\max_{\gamma\in[0,1]}|\alpha(\gamma)-\alpha|$ does not vary much with α for $\alpha=.10$, $.05$, $.01$. The first statement is in accordance with the first line of this paragraph, the second statement indicates that the approximation of α by $\alpha(\gamma)$ is better in a relative sense for larger α than for smaller α.
Because $f_0\leq\underline{F}\leq f_3$ with probability 1 for $\gamma\in[0,1]$, the APDF interval (3.1.3) is shorter than the Wald-Hājek interval (2.1.3) in the ratio $t_{\alpha/2}(\underline{F})/t_{\alpha/2}(f_0)$ which varies between $t_{\alpha/2}(f_3)/t_{\alpha/2}(f_0)$ and 1, but the APDF interval is longer than the large sample intervals (2.2.1) and (2.2.2) in the ratios $t_{\alpha/2}(\underline{F})/u_{\alpha/2}$ and $t_{\alpha/2}(\underline{F})/t_{\alpha/2}(f_3)$ respectively; the latter varies between 1 and $t_{\alpha/2}(f_0)/t_{\alpha/2}(f_3)$. The APDF test associated with (3.1.3) has thus a greater power than the Wald-Hājek test, but it has a smaller power than the large sample tests neglecting their possible invalidity; computations [10] indicate that the power of Welch's APDF test is (much) greater than the power of Banerjee's test too.
One can get a reasonable approximation of the real tail probability of $\underline{v}_\delta$ by means of four point interpolation for the actual values of $\underline{v}_\delta$ and $\underline{F}$ in a table of the c.d.f. of Student's distribution, on account of (3.1.1) and (3.1.2); it should be emphasized however that the thus obtained value *approximates* the real tail probability of $\underline{v}_\delta$: if this approximation for $\delta=\delta_0$ is close to α, the size of the test for $H_0:\delta=\delta_0$, it seems to be rather risky therefore to draw inferences about H_0 from this value.

3.2. Pagurova's solution

PAGUROVA [13] derived a solution, which is an extension of a solution given by WALD [19] for $n_1=n_2$ in a certain sense. She considers on [0,1] an interpolating third degree polynomial $p(\gamma)$ of Lagrange type, which takes on the values $t_{\alpha/2}(f_2)$, $t_{\alpha/2}(f_3)$ and $t_{\alpha/2}(f_1)$ at the points $\gamma=0$, γ_1, 1 respectively:

$$(3.2.1)\qquad p(\gamma) = \frac{(\gamma_1-\gamma)^2(1-\gamma)}{\gamma_1^2}\,t_{\alpha/2}(f_2) + \frac{\{\gamma_1(1-\gamma_1)+(2\gamma_1-1)(\gamma-\gamma_1)\}\gamma(1-\gamma)}{\gamma_1^2(1-\gamma_1)^2}\,t_{\alpha/2}(f_3) + \frac{(\gamma_1-\gamma)^2\gamma}{(1-\gamma_1)^2}\,t_{\alpha/2}(f_1).$$

Then the unknown γ in (3.2.1) is estimated by an estimator, for example $\underline{c}$, as is done in the case of Welch's APDF solution. Actually Pagurova takes an estimator $\underline{c}_1$ with skewness of order f_0^{-2} for $f_0 \to \infty$:

$$(3.2.2)\qquad \underline{c}_1 = \underline{c}-2\underline{c}(1-\underline{c})\{(1-\underline{c})/f_2-\underline{c}/f_1\} = \underline{c}(1-2/f_2)+2\underline{c}^2(1/f_1+2/f_2)-2\underline{c}^3(1/f_1+1/f_2).$$

With the help of (3.2.1) and (3.2.2) the interval form of the *Pagurova solution* can be stated to be

$$(3.2.3)\qquad |\delta-\underline{d}| \le p(\underline{c}_1)\underline{s}.$$

Since $\lim_{f_0\to\infty} p(\underline{c}_1) = u_{\alpha/2}$ with probability 1 for $\gamma\in[0,1]$, the Pagurova solution obeys relation (1.2.9), but it is not a valid solution: relation (1.2.8) does not hold for it. Pagurova proves for her solution that

$$\alpha(\gamma) = \alpha + O(f_0^{-2}), \qquad \gamma\in[0,1],$$

for $f_0 \to \infty$.*)

In [13] Pagurova gives a table of values of $p(\underline{c}_1)$ for $\alpha = .05$, which

*) The same statement holds for Welch's APDF solution, either with $\underline{c}$ or with $\underline{c}_1$ as an estimator of γ.

is entered with f_1, f_2 and the value c of $\underline{c}$, as computed from the sample statistics, and which requires $f_0 \geq 4$. Furthermore, Pagurova gives a table which contains maximum and minimum values of $\alpha(\gamma)$ with respect to $\gamma \in [0,1]$ for her solution in the case $\alpha = .05$ and which is entered with f_1, f_2, where $f_1 \geq f_2 \geq 2$ (values correct within .0002). A short version of this table is given below: the first line of it contains the values of f_1, the second line the maximum value and the third line the minimum value of $\alpha(\gamma)$, both with respect to $f_2 \in [2, f_1]$ and $\gamma \in [0,1]$; the latter two values must be multiplied with 10^{-4}:

2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
585	531	515	508	505	504	503	502	501	501	500	500	500	500	500	500	500	500
333	372	419	445	461	471	477	482	485	490	493	496	497	498	499	499	499	500

It is seen from this table, that already for such small sample sizes as $f_0 = 4$ the over-estimation of $\alpha = .05$ by $\alpha(\gamma)$ for some γ is small, and as it seems to be, smaller than for the APDF solution, at the cost however of a greater underestimation of $\alpha = .05$ by $\alpha(\gamma)$ for some γ, as it seems, than for the APDF solution. In [10] computations of $\alpha(\gamma)$ for $\alpha = .05$, $f_0 \geq 2$ bear out the same conclusions.
Since $t_{\alpha/2}(f_3) \leq p(\underline{c}_1) \leq t_{\alpha/2}(f_0)$ with probability 1 for $\gamma \in [0,1]$ analogous remarks with respect to length and power as made for the APDF interval (3.1.3) and associated test can be made for the Pagurova interval (3.2.3) and associated test. Calculations show that the Pagurova interval is somewhat longer than the APDF interval, but the differences are small, so the associated tests have about the same power [10]. The replacement of $\underline{c}_1$ by $\underline{c}$ tends to shorten the interval length somewhat, but of course at the price of diminishing the confidence coefficient possibly. An analogous remark is true for the APDF intervals.

Both Welch's APDF solution and Pagurova's solution require only tables of the Student's distribution, which are readily available. Both solutions have about the same properties with respect to the probability of type I error and the power of the tests based on them or the corresponding properties of the associated confidence intervals, and these properties seem to be satisfactory in practice, if not good. Welch's APDF solution can be computed possibly somewhat easier than Pagurova's solution, but the latter is more accurate than the former with respect to the probability of type I error, as it seems to be; the APDF solution allows a reasonable approximation of the real tail probability of the value of $\underline{v}_\delta$.

A reasonable conclusion is, I think, that Welch's APDF solution is a satisfactory practical solution for common use in the Behrens-Fisher problem, if the sample sizes are intermediate, and that Pagurova's solution is very useful as a satisfactory practical solution of this problem for the same sample sizes if a good protection against overestimation of α by $\alpha(\gamma)$ is desirable, of course in both cases under the assumption that a reasonable power (or interval length) is needed.

4. Other solutions

If λ is a known number, say λ_0, the statistic

$$(4.1) \qquad \underline{a}_\delta(\lambda_0) = (m_1+m_2/\lambda_0)^{-\frac{1}{2}}(\underline{d}-\delta)/\underline{S}, \text{ where } \underline{S} = (f_1\underline{s}_1^2+\lambda_0 f_2\underline{s}_2^2)/f_3,$$

may be proved to have a Student's distribution with f_3 d.f.. Tests and confidence intervals concerning δ based on it are the best among all related unbiased procedures [8]. In statistical practice it is often assumed that $\underline{a}_\delta(\lambda_0)$ is distributed, at least approximately, as $\underline{t}(f_3)$, even if λ is not known to equal λ_0. The variable obtained in this manner depends on λ of course and I denote it by $\underline{a}_\delta(\lambda|\lambda_0)$ from here. It is used to obtain "solutions" of the Behrens-Fisher problem in the interval form

$$(4.2) \qquad |\delta-\underline{d}| \leq t_{\alpha/2}(f_3)\ (m_1+m_2/\lambda_0)^{\frac{1}{2}}\underline{S},$$

where the quotation marks indicate, that (4.2) is not intended as a solution of the Behrens-Fisher problem; nevertheless it is used heavily in such a way in practice. If λ_0 is the true value of λ, this solution

is justified as outlined above, but otherwise not.
If $\lambda_0 = n_1 f_1 (n_2 f_2)^{-1}$ it is easily seen that $\underline{v}_\delta$ and $\underline{a}_\delta(\lambda|\lambda_0)$ have the same distribution, so that the solution (4.2) and the large sample solution (2.2.2) coincide; hence the former is asymptotically correct for this special choice of λ_0, depending on the sample sizes. For all constant values of λ_0 however, solution (4.2) is asymptotically incorrect, for it can be proved by elementary calculations, that for every constant λ_0, every $\alpha\epsilon(0,1)$ and every $\alpha^1\epsilon(0,1)$ a triplet (λ, n_1, n_2) exists such that $P\{|\underline{a}_\delta(\lambda|\lambda_0)| \geq t_{\alpha/2}(f_3)\} = \alpha^1$. In practical applications (4.2) is mostly used with $\lambda_0=1$ as a "solution" of the Behrens-Fisher problem. From above the conclusion may be drawn that it should never be used that way, unless $n_1=n_2$ or, only if n_1 and n_2 are large, nearly so.

Another solution sometimes used in practice is the one which has been derived by WELCH [21] and which has been developed further by ASPIN [1]; it has the form

$$(4.3) \qquad |\delta-\underline{d}| \leq w(\underline{c})\underline{s},$$

where the value of $w(\underline{c})$ is given by table 11 in [14] for $\alpha = .10, .05, .02, .01$. For the execution of solution (4.3), generally called the *Welch-Aspin solution*, these tables are necessary; they are entered with f_1, f_2 and the value c of $\underline{c}$, computed from the samples drawn, and they require f_0 to be $\geq 6, 8, 10, 10$ when $\alpha = .10, .05, .02, .01$ respectively.
The Welch-Aspin solution is not valid, but it obeys relation (1.2.9), since $\lim_{f_0\to\infty} w(\underline{c}) = u_{\alpha/2}$ with probability 1 for $\gamma\epsilon[0,1]$. Numerical verification of $\alpha(\gamma)$ is needed because of the purely formal character of the derivation of the Welch-Aspin solution. Examples of it may be found in [20] and [22] for d.f. between 6 and 20 and $\alpha = .10, .05, .01$ and various values of γ. For these examples $|\alpha(\gamma)-\alpha| \leq .0009$ (correct within .0002), so the Welch-Aspin solution makes the maximum in (1.2.10) small indeed. Interval length of (4.3), and accordingly the power of the associated test, are about the same as those of Welch's APDF solution [18].

It is a direct consequence of inequalities (1.2.5) that it is reasonable to expect the value of $w(\underline{c})$ to lie between $t_{\alpha/2}(f_3)$ and $t_{\alpha/2}(f_0)$. However from the tables it is seen to fall between $u_{\alpha/2}$ and $t_{\alpha/2}(f_0)$; hence the nominal value α is certainly overestimated by $\alpha(\gamma)$ for some values of γ, especially if the sample sizes are not to great. Although the Welch-Aspin solution is intended to have the mean value of $\alpha(\gamma)$ equal to α, at least in the long run, it is not really a practical and satisfactory test for reason of this overestimation of α by $\alpha(\gamma)$ for some γ, certainly if the needfulness of the tables of the values of $w(\underline{c})$ is taken into account.

The solution used by supporters of fiducial probability has been derived by FISHER [5]; it has the form

$$(4.4) \qquad |\delta-\underline{d}| \leq F(\underline{c})\underline{s},$$

where the value of $F(\underline{c})$ is given by tables VI and VI_1 in [6] and by tables in [23], [24], [25]. For the execution of solution (4.4), which is called the *Fisher-Behrens solution* in honour of the statisticians who worked first on it, these tables are necessary; they are entered with f_1, f_2 and θ, where

$$(4.5) \qquad \theta = \arctan\{m_1\underline{s}_1^2(m_2\underline{s}_2^2)^{-1}\} = \arcsin \underline{c}^{\frac{1}{2}}$$

and the n_i and $\underline{s}_i$ of the tables and text denote our f_i and $m_i^{\frac{1}{2}}\underline{s}_i$ respectively. With respect to [6] table VI is for $f_0 \geq 6$, $\alpha = .05, .01$ and table VI_1 is for all odd f_1, $f_2 \leq 7$ and $\alpha = .10, .05, .02, .01$. The tables in [23] and [24] are for $f_0 \geq 6$ and $\alpha = .001$ and $.002$ respectively. The table in [25] contains some revised figures of the other tables. Here α does not denote the size of the test corresponding with (4.4), but this symbol is used rather in a fiducial probability sense.

The Fisher-Behrens solution is not valid, at least in a classical probability sense [10], but it obeys relation (1.2.9), since

$\lim_{f_0 \to \infty} F(\underline{c}) = u_{\alpha/2}$ with probability 1 for $\gamma \in [0,1]$; in this paragraph α denotes the true size of the test associated with (4.4) again. In [20] it is computed with respect to the maximum in (1.2.10) that $|\alpha(\gamma)-\alpha| \leq .0144$ for $f_1=6$, $f_2=12$, $\alpha=.05$ and various values of γ and from computations in [10] it is seen, that the Fisher-Behrens solution is comparable with the Banerjee solution so far as probability of type I error and power of the corresponding tests are concerned; the latter however is a valid solution. Welch [21] calculated that the Fisher-Behrens interval (4.4) should always include the Welch-Aspin interval (4.3) and the tables of both solutions bear this out, in accordance with the computations in [10].
Because of the necessary tables, the conceptual difficulties and the relatively unfavourable properties of the Fisher-Behrens solution, I think it is reasonable to judge that this solution is not a satisfactory practical solution of the Behrens-Fisher problem, at least not within the framework of the classical frequency interpolation of probability.

In conclusion I remark that all solutions discussed in sections 2, 3 and 4 except the solution given by (4.2) are robust against violations of the normality assumption with respect to the underlying distributions, since all solutions are based on $\underline{v}_\delta$, which depends mainly on the difference $\underline{d}$ of sample means and where in $\underline{s}^2$ is an unbiased consistent estimator of σ^2. The same arguments assure that $\underline{v}_\delta$ is asymptotically N(0,1), if the samples are drawn from populations whatsoever with finite second moment. Hence the discussed solutions, except (4.2), will provide, at least asymptotically, good measures for the difference of the population means in these cases too.

Acknowledgement. I would like to thank prof.dr. J. Hemelrijk for his stimulating criticisms and valuable suggestions during the preparation of this paper.

References

[1] Aspin, A.A. *An examination and further development of a formula arising in the problem of comparing two mean values*, Biometrika 35 (1948) 88-96.

[2] Banerjee, S. *On confidence interval for two-means problem based on separate estimates of variances and tabulated values of t-table*, Sankhyā A 23 (1961) 359-378.

[3] Bartlett, M.S. *The information available in small samples*, Proceedings of the Cambridge Philosophical Society 32 (1936) 560-566.

[4] Cramér, H. *Mathematical methods of statistics*, Princeton Mathematical Series, volume 9, Princeton University Press, Princeton, N.J., 1946.

[5] Fisher, R.A. *The fiducial argument in statistical inference*, Annals of Eugenics 6 (1935) 391-398 (Reprinted as paper 25 in R.A. Fisher - *Contributions to mathematical statistics*, Wiley, New York and Chapman & Hall, London, 1950).

[6] Fisher, R.A. & Yates, F. *Statistical tables for biological, agricultural and medical research*, Oliver and Boyd, Edinburgh, 1957.

[7] Hâjek, J. *Inequalities for the generalized Student's distribution and their applications*, Selected translations in mathematical statistics and probability, volume 2 (1962) 63-74, American Mathematical Society, Providence, Rhode Island.

[8] Lehmann, E.L. *Testing statistical hypotheses*, Wiley, New York, 1959.

[9] Linnik, Ju.V. *Statistical problems with nuisance parameters*, Izdva "Nauka", Moscow, 1966; Translations of

Mathematical Monographs, volume 20; American Mathematical Society, Providence, Rhode Island, 1968.

[10] Mehta, J.S. & Srinivasan, R. *On the Behrens-Fisher problem*, Biometrika 57 (1970) 649-655.

[11] Mickey, M.R. & Brown, M.B. *Bounds on the distribution functions of the Behrens-Fisher statistic*, The Annals of Mathematical Statistics 37 (1966) 639-642.

[12] Owen, D.B. *The power of Student's t-test*, Journal of the American Statistical Association 60 (1965) 320-333.

[13] Pagurova, V.I. *On a comparison of means of two normal samples*, Teoriya Veroyatnostei i ee Primeneniya 13 (1968) 561-569; Theory of probability and its applications 13 (1968) 527-534.

[14] Pearson, E.S. & Hartley, H.O. *Biometrika tables for statisticians*, Cambridge University Press, Cambridge, England, 1966.

[15] Resnikoff, G.J. & Lieberman, G.J. *Tables of the non-central t-distribution*, Stanford University Press, Stanford, California, 1957.

[16] Scheffé, H. *On solutions of the Behrens-Fisher problem, based on the t-distribution*, The Annals of Mathematical Statistics 14 (1943) 35-44.

[17] Scheffé, H. *A note on the Behrens-Fisher problem*, The Annals of Mathematical Statistics 15 (1944) 430-432.

[18] Scheffé, H. *Practical solutions of the Behrens-Fisher problem*, Journal of the American Statistical Association 65 (1970) 1501-1508.

[19] Wald, A. *Testing the difference between the means of two normal populations with unknown standard deviations*, Selected papers in statistics and proba-

bility by Abraham Wald, 669-695; McGraw Hill, New York, 1955.

[20] Wang, Y.Y. *The probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem*, Journal of the American Statistical Association 66 (1971) 605-608.

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

[22] Welch, B.L. *Further note on Mrs. Aspin's tables and on certain approximations to the tabled function*, Biometrika 36 (1949) 293-296.

[23] Weir, J.B. de V. *Table of 0.1 percentage points of Behrens' d*, Biometrika 53 (1966) 267-268.

[24] Weir, J.B. de V. *Table of 0.2 percentage points of Behrens' d*, Sankhyā B 31 (1969) 103-104.

[25] Weir, J.B. de V. *Revised end-figures of Behrens' d*, Sankhyā B 31 (1969), 105-106.