AN ADJUSTED LIKELIHOOD-RATIO APPROACH
TO THE BEHRENS-FISHER PROBLEM

Hamparsum Bozdogan and Donald E. Ramirez

Department of Mathematics
University of Virginia
Charlottesville, VA 22903

Key Words and Phrases: adjusted likelihood-ratio; maximum likelihood; Behrens-Fisher problem.

ABSTRACT

In many situations it is necessary to test the equality of the means of two normal populations when the variances are unknown and unequal. This paper studies the celebrated and controversial Behrens-Fisher problem via an adjusted likelihood-ratio test using the maximum likelihood estimates of the parameters under both the null and the alternative models. This procedure allows the significance level to be adjusted in accordance with the degrees of freedom to balance the risk due to the bias in using the maximum likelihood estimates and the risk due to the increase of variance. A large scale Monte Carlo investigation is carried out to show that \( -2 \ln L \) has an empirical chi-square distribution with fractional degrees of freedom instead of a chi-square distribution with one degree of freedom. Also Monte Carlo power curves are investigated under several different conditions to evaluate the performances of several conventional procedures with that of this procedure with respect to control over Type I errors and power.

2405
1. INTRODUCTION AND THE BEHRENS-FISHER PROBLEM

In many practical situations it is necessary to test the equality of the means when neither is known, and we cannot reasonably assume that the unknown variances $\sigma_1^2$ and $\sigma_2^2$ are equal. This is the case in many experimental situations where the researcher often wishes to compare the means, but he is unwilling to assume that $\sigma_1^2 = \sigma_2^2$, or finds it inappropriate to make such an assumption.

The comparison of two normal means, when the variances are unknown and unequal, is called the Behrens-Fisher problem. It is also referred to as the two-means problem, and, in short, Behrens' problem. More formally, suppose we have a random sample of size $n_1, x_{11}, x_{12}, \ldots, x_{1n_1}$, from $N(\mu_1, \sigma_1^2)$ and a second independent sample of size $n_2, x_{21}, x_{22}, \ldots, x_{2n_2}$, from $N(\mu_2, \sigma_2^2)$. We are interested in testing

$$H_0: \mu_1 = \mu_2 = \mu \text{ (say)}$$

against

$$H_a: \mu_1 \neq \mu_2$$

when $\sigma_1^2$ and $\sigma_2^2$ are unknown and not assumed to be equal.

The Behrens-Fisher problem from a theoretical viewpoint can also be formulated as: "To find the similar test which is a function of complete sufficient statistics," which is solved by Linnik (1968).

In this paper we shall study the Behrens-Fisher problem via an adjusted likelihood-ratio test using the maximum likelihood estimates of the parameters under both the null and the alternative models. This procedure allows the significance level to be adjusted in accordance with the degrees of freedom to balance the
risk due to the bias in using the maximum likelihood estimates and the risk due to the increase of variance.

In this paper, we propose a simple iterative likelihood-ratio approach for the Behrens-Fisher problem and show that our proposed approach is essentially equivalent to the Welch's procedure. We also propose an adjusted likelihood-ratio procedure and show that \(-2 \ln \Lambda\) has an empirical chi-squared distribution with fractional degrees of freedom instead of chi-squared distribution with one degree of freedom.

Also, in this paper, via a large scale simulation study, we investigated and compared the Monte Carlo power curves of our procedure with that of several existing procedures to evaluate its performance under several conditions with respect to control over Type I errors and power. Later, we validate our assumption and the simulation study.

2. NOTATION, BACKGROUND MATERIAL AND DERIVATIONS

In this paper we assume that

\[ X_{1i} \sim N(\mu_1, \sigma_1^2), \quad i = 1, 2, \ldots, n_1 \]

and

\[ X_{2i} \sim N(\mu_2, \sigma_2^2), \quad i = 1, 2, \ldots, n_2 \]  

(2.1)

are independent random samples of sizes \(n_1\) and \(n_2\), respectively.

We let \(\bar{X}_1\) and \(\bar{X}_2\) denote the sample means, and \(S_{u1}^2\) and \(S_{u2}^2\) denote the unbiased sample variances. That is,

\[ \bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}, \]

\[ \bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}, \]  

(2.2)
\[ S_{u1}^2 = \frac{1}{n_{1}-1} \sum_{i=1}^{n_1} (X_{1i} - \overline{X}_1)^2, \]

\[ S_{u2}^2 = \frac{1}{n_{2}-1} \sum_{i=1}^{n_2} (X_{2i} - \overline{X}_2)^2. \]

If the population variances are equal \((\sigma_1^2 = \sigma_2^2)\), then the test for equality of means is the usual t-test given by

\[
t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{(n_{1}-1)S_{u1}^2 + (n_{2}-1)S_{u2}^2}{n_{1} + n_{2}-2}\right)\left(\frac{1}{n_{1}} + \frac{1}{n_{2}}\right)}}^{1/2} - t_{n_{1}+n_{2}-2} \quad (2.3)
\]

or

\[
F = t^2 - F_{1, n_{1}+n_{2}-2}. \quad (2.4)
\]

When the population variances are not equal \((\sigma_1^2 \neq \sigma_2^2)\) and are unknown, then the most common procedure for testing the equality of means is the Welch's procedure [see, for example, Welch (1937, 1947)]. For the Welch procedure, one computes

\[
t_w = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{S_{u1}^2}{n_1} + \frac{S_{u2}^2}{n_2}\right)}}^{1/2}, \quad (2.5)
\]

where

\[
E \left[ \left(\frac{S_{u1}^2}{n_1} + \frac{S_{u2}^2}{n_2}\right) \right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \quad (2.6)
\]

and

\[
\text{Var} \left[ \left(\frac{S_{u1}^2}{n_1} + \frac{S_{u2}^2}{n_2}\right) \right] = \frac{2\sigma_1^4}{n_1(n_1-1)} + \frac{2\sigma_2^4}{n_2(n_2-1)}. \quad (2.7)
\]
AN ADJUSTED LIKELIHOOD-RATIO APPROACH

The Welch procedure assumes that \( S_{u1}^2/n_1 + S_{u2}^2/n_2 \) is approximately distributed as \( (\sigma^2/\nu)\chi^2 \), for some \( \nu > 0 \) and \( \sigma^2 > 0 \). By matching first and second moments, Welch showed that

\[
t_w \sim t_{\nu_w} \text{ (approximately)} \quad (2.8)
\]

with

\[
\nu_w = \frac{\left(\frac{S_{u1}^2}{n_1} + \frac{S_{u2}^2}{n_2}\right)^2}{\frac{(S_{u1}^2/n_1)^2}{n_1-1} + \frac{(S_{u2}^2/n_2)^2}{n_2-1}} \quad (2.9)
\]

degrees of freedom.

With the notation established and the background material given, we now derive the iterative likelihood-ratio procedure, and consequently obtain the maximum likelihood estimators (MLE's) of the parameters. Here the parameter spaces under the alternative and the null hypotheses are:

\[
\Omega = \{ \theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2): -\infty < \mu_1, \mu_2 < \infty, \sigma^2_g > 0, g = 1, 2 \}
\]

\[
\omega = \{ \theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2): -\infty < \mu_1 = \mu_2 < \infty, \sigma^2_g > 0, g = 1, 2 \}
\]

\[
(2.10)
\]

When \( H_0: \mu_1 = \mu_2 \) (=\( \mu \)) holds, then MLE estimators are the roots of the set of three equations:

\[
\begin{align*}
\frac{n_1}{\sigma_1^2} (\bar{X}_1 - \mu) + \frac{n_2}{\sigma_2^2} (\bar{X}_2 - \mu) & = 0 \\
\sigma_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (X_{1i} - \mu)^2 & = S_1^2 + (\bar{X}_1 - \mu)^2 \\
\sigma_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (X_{2i} - \mu)^2 & = S_2^2 + (\bar{X}_2 - \mu)^2
\end{align*}
\]

\[
(2.11)
\]
For this, see, for example, also Kendall and Stuart (1967, p. 226).

From the first equation in (2.11), we obtain very readily the solution of \( \mu \). Thus, under \( H_0 \), the maximum likelihood equations are given by

\[
\hat{\mu} = \frac{\bar{X}_1 + B \bar{X}_2}{A + B} \quad \text{with} \quad A = \left( \frac{\sigma_1^2}{n_1} \right)^{-1}, \quad B = \left( \frac{\sigma_2^2}{n_2} \right)^{-1},
\]

(2.12)

and

\[
\sigma_1^2 = s_1^2 + (\bar{X}_1 - \hat{\mu})^2,
\]

\[
\sigma_2^2 = s_2^2 + (\bar{X}_2 - \hat{\mu})^2.
\]

We note that \( s_1^2 \) and \( s_2^2 \) are the biased sample variances and they need to be distinguished from the unbiased sample variances \( s_{u1}^2 \) and \( s_{u2}^2 \). It is well known that under the null model, the statistic \( T = (\bar{X}_1, \bar{X}_2; s_{u1}^2, s_{u2}^2) \) is sufficient, but it is not complete since \( E(\bar{X}_1 - \bar{X}_2) \) is identically equal to zero as stated in Lehmann (1959, pp. 130-134).

From (2.11), we see that \( \hat{\mu} \) solves the cubic polynomial in \( \mu \):

\[
n_1(s_2^2 + \bar{X}_2^2 - 2\bar{X}_2\mu + \mu^2)(\bar{X}_1 - \mu) + n_2(s_1^2 + \bar{X}_1^2 - 2\bar{X}_1\mu + \mu^2)(\bar{X}_2 - \mu) = 0
\]

(2.13)

Although (2.13) establishes a direct solution for \( \hat{\mu} \), we have found it easy to work with the maximum likelihood equations (2.12) by using the following iterative process to obtain the MLE of \( \mu \) and \( \sigma_2^2 \) under the null model. Recently, however, Sugiuira and Gupta (1985) have shown that the cubic polynomial in (2.13) has either unique solution with large probability, or three solutions with small positive probability.
Proposed Iterative Process:

(i) We start with the initial estimate \( \hat{\sigma}^2_g = s^2_g, \ g = 1, 2 \). That is,

\[
\hat{\sigma}^2_1 = s^2_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} (X_{1i} - \overline{X}_1)^2,
\]

\[
\hat{\sigma}^2_2 = s^2_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (X_{2i} - \overline{X}_2)^2.
\]

(ii) We use these to compute the estimate of \( \mu \). We call this

\[
\hat{\mu}_0 = \frac{n_1 \hat{\sigma}^2_1 \overline{X}_1 + n_2 \hat{\sigma}^2_2 \overline{X}_2}{n_1 \hat{\sigma}^2_1 + n_2 \hat{\sigma}^2_2}
\]

(iii) We use \( \hat{\mu}_0 \) now in (2.12). That is, in

\[
\hat{\sigma}^2_g = s^2_g + (\overline{X}_g - \hat{\mu}_0)^2, \ g = 1, 2.
\]

(iv) We use \( \hat{\sigma}^2_g \) in (iii) and repeat (ii) until convergence takes place.

In our case, convergence to six decimals is usually achieved within ten loops. Hence, in this manner we obtain the iterative MLE's for \( \mu \) and \( \sigma^2_g \) for \( g = 1, 2 \), under \( H_0 \) which we call the Behrens–Fisher (B–F) hypothesis.

Let \( \lambda \) denote the likelihood-ratio for testing the null hypothesis \( H_0: \mu_1 = \mu_2 \), and let \( \lambda = -2 \ln \lambda \). It is easy to show that

\[
\lambda = -2 \ln \lambda = n_1 \ln(\hat{\sigma}^2_1) + n_2 \ln(\hat{\sigma}^2_2) - n_1 \ln(s^2_1) - n_2 \ln(s^2_2).
\]

(2.14)

We now state our main result in the following proposition.

**Proposition 2.1.** Under the null hypothesis \( H_0: \mu_1 = \mu_2 = \mu \), with the assumption that the iteration process gives a consistent
maximum likelihood estimates of order $1/\sqrt{n}$, that is, $\hat{\mu} = \mu + O_p(1/\sqrt{n})$ and $\hat{\sigma}^2 = \sigma^2 + O_p(1/\sqrt{n})$, where $n = n_1 + n_2$, and that the sample sizes $n_1, n_2$ tend to infinity both at the same rate:

$$-2 \ln \lambda - (t_w)^2 \xrightarrow{d} 0,$$

(2.15)

where $d$ stands for convergence in probability, and $t_w$ is Welch's t-statistic given in (2.5).

**Proof:** Using the iterative process, from (2.14) we have

$$\lambda = n_1 \ln \left[ 1 + \frac{(\bar{X}_1 - \hat{\mu})^2}{S_1^2} \right] + n_2 \ln \left[ 1 + \frac{(\bar{X}_2 - \hat{\mu})^2}{S_2^2} \right].$$

(2.16)

Note that $(\bar{X}_2 - \hat{\mu})^2/S_2^2 = O_p(1/n)$ under $H_0$ and that $\hat{\mu}$ satisfies (2.12). From a Taylor's series expansion of $\ln(1+X)$, we get

$$\lambda = n_1 \left[ \frac{1}{n_1} \frac{(\bar{X}_1 - \hat{\mu})^2}{S_1^2} + O_p(1/n_1) \right] + n_2 \left[ \frac{1}{n_2} \frac{(\bar{X}_2 - \hat{\mu})^2}{S_2^2} + O_p(1/n_2) \right]$$

$$= n_1 \frac{(\bar{X}_1 - \hat{\mu})^2}{\sigma_1^2} + n_2 \frac{(\bar{X}_2 - \hat{\mu})^2}{\sigma_2^2} + O_p(1/n)$$

$$= \frac{(n_2 \sigma_1^4/S_1^2 + n_1 \sigma_2^4/S_2^2)}{n_1 \sigma_1^2 + n_2 \sigma_2^2} \cdot \frac{n_1 n_2 (\bar{X}_1 - \bar{X}_2)^2}{n_1 n_2 \sigma_1^2} = O_p(1/n).$$

(2.17)

On the other hand, from (2.5) we have
AN ADJUSTED LIKELIHOOD-RATIO APPROACH

\[ t_w^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} + O_p(1/\sqrt{n}). \]  

(2.18)

This gives

\[ A - t_w^2 = O_p(1/\sqrt{n}), \]  

(2.19)

which implies

\[ A - t_w^2 \rightarrow 0 \text{ in probability}. \]  

(2.20)

Thus, we can assert that

\[ A \equiv -2 \ln \lambda \sim t_w^2 \sim F_{1, n_w}. \]  

(2.21)

3. AN ADJUSTED LIKELIHOOD-RATIO PROCEDURE

To test the B-F hypothesis \( H_0: \mu_1 = \mu_2 \) when \( \sigma_1^2 \) and \( \sigma_2^2 \) are unknown and not assumed to be equal, by the likelihood-ratio test, one would compute \( A \equiv -2 \ln \lambda \) and then compare \( A \) to a chi-square distribution, \( x^2_1 \), with degrees of freedom equal to 1. This procedure requires \( n_1 \) and \( n_2 \) to be quite large.

Box (1949) and Lawley (1956) have shown that for a certain class of such criteria, it is possible to improve the approximations based on a knowledge of the moments of the criterion which can be developed even for samples of moderate size. Thus, for \( n_1 \) and \( n_2 \) small, one must adjust \( A \equiv -2 \ln \lambda \); or equivalently, one must adjust the distribution of \( x^2_1 \). Therefore, in this section we shall develop a way of adjusting the moments of \( A \equiv -2 \ln \lambda \) based on our results in Section 2.
From Proposition 2.1 in Section 2, we approximate $E[-2 \ln \lambda]$ by $E[F_{1, \nu_w}]$:

$$E[-2 \ln \lambda] = E[F_{1, \nu_w}] = \frac{\nu_w}{\nu_w - 2}$$

(3.1)

Thus, correcting to the first moment, we have that

$$\lambda = \lambda_1 = -2 \ln \lambda - \chi^2_{f_1},$$

(3.2)

with

$$f_1 = \frac{\nu_w}{\nu_w - 2} \text{ or } f_1 = 1 + \frac{2}{\nu_w - 2},$$

and where

$$\nu_w = \frac{(S^2_{u_1}/n_1 + S^2_{u_2}/n_2)^2}{(S^2_{u_1}/n_1)^2 n_1^{-1} + (S^2_{u_2}/n_2)^2 n_2^{-1}}$$

Hence the term $2/(\nu_w - 2)$ can be viewed as a correction for the bias in $\lambda = -2 \ln \lambda$. The simplest improvement consists in multiplying $-2 \ln \lambda$ by a scale factor which results in a statistic having the same first moment as $\chi^2_{f_1}$ ignoring quantities of order $n^{-2}$, where $n$ is the sample size. In the literature, this scaling device was first used by Bartlett (1937). For more details on this, see, for example, Kendall and Stuart (1961, pp. 234–236). Press (1967) has adjusted the likelihood-ratio test in studying the multivariate Behrens-Fisher problem, and Hayakawa (1977) has considered the general asymptotic expansion of the distribution of the likelihood-ratio criterion to improve the approximations in cases where the distribution of $\lambda = -2 \ln \lambda$ is not known in closed form. The technique of correcting the bias in the log-likelihood function using the device of "effective degrees of freedom" has been studied recently by Cox (1964). In the Behrens-Fisher problem, the scale
or the correction factor \( c = (\nu_w - 2)/\nu_w \) for \(-2 \ln \lambda\) to be approximately \( X_1^2\). Certainly one can compare the approximations

(i) \[-2 \ln \lambda \sim F_{1, \nu_w}\]

(ii) \[-2 c \ln \lambda \sim X_1^2\]

(iii) \[-2 \ln \lambda \sim X_{f_1}^2\]

For brevity, we have not pursued this line of investigation, but rather following Cox (1984), we have concentrated on correcting the bias by adjusting the degrees of freedom using method (iii) and compared it to other procedures.

If we wish to adjust \( \Lambda \equiv -2 \ln \lambda \) to the first and second moments jointly, we assume that \( \Lambda \equiv -2 \ln \lambda \sim G(\alpha, \beta) \), a gamma distribution with parameters \( \alpha \) and \( \beta \), where \( \alpha \) is the shape parameter and \( \beta \) is the scale parameter. According to Cox (1985), theoretically, there is no special justification to adjust \( \Lambda \equiv -2 \ln \lambda \) to the first and second moments jointly, but one might hope that by adjusting for both moments, empirically one could achieve some improvement. In this case, the density of a gamma distribution is [see Bury (1975, pp. 290-331)]:

\[
f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha\!-\!1} e^{-x/\beta}, \quad x > 0. \tag{3.3}
\]

Now

\[
E(\Lambda) \equiv E(-2 \ln \lambda) = E(F_{1, \nu_w}) = \frac{\nu_w}{\nu_w - 2},
\]

and

\[
Var(\Lambda) \equiv Var(-2 \ln \lambda) = Var(F_{1, \nu_w}) \tag{3.4}
\]

\[
= \frac{\nu_w(\nu_w - 1)}{(\nu_w - 2)^2 (\nu_w - 4)}, \quad \nu_w > 4.
\]
Thus we set
\[ \alpha \beta = \frac{\nu_w}{\nu_w - 2}. \]  
(3.5)
and
\[ \alpha \beta^2 = \frac{2 \nu_w^2 (\nu_w - 1)}{(\nu_w - 2)^2 (\nu_w - 4)}. \]  
(3.6)
Solving for \( \alpha \) and \( \beta \) yields
\[ \alpha_w = \frac{\nu_w - 4}{2 \nu_w - 2} = \frac{1}{2} - \frac{3}{2 \nu_w - 2}. \]  
(3.6)
and
\[ \beta_w = \frac{2 \nu_w (\nu_w - 1)}{(\nu_w - 2)(\nu_w - 4)} = \frac{2 + 10 \nu_w - 16}{\nu_w^2 - 6 \nu_w + 8}. \]
Note that as \( n_1, n_2 \to \infty \), \( \nu_w \to \infty \) too. Thus \( f_1 \to 1 \), \( \alpha_w \to 1/2 \), and \( \beta_w \to 2 \).

If we denote the likelihood-ratio test statistic without adjustment by \( \Lambda = -2 \ln \Lambda \); the likelihood-ratio test statistic adjusted to the first moment by \( \Lambda_1 \); and the likelihood-ratio test statistic adjusted to the first and second moments jointly by \( \Lambda_{1,2} \). From our Monte Carlo simulation studies, discussed in the next section, it is relevant to report that all or nearly all of our size calculations (\( \Psi = 0 \)) showed \( \Lambda_1 \) as giving values closer to nominal level than \( \Lambda \) and \( \Lambda_{1,2} \).

Therefore, to test for the equality of means without any assumption on the variances, for all practical purposes, it is sufficient to compute \( \Lambda = -2 \ln \Lambda \) by using the iterative procedure we proposed in Section 2, and then compare \( \Lambda \) to \( \Lambda_1 \), i.e., as adjusted \( \chi^2 \) with \( f_1 \) degrees of freedom given by (3.2). Our Monte Carlo simulation studies show that this way of correcting \( \Lambda \) to the first moment will produce excellent results between the nominal
level 2.5% and 10% significance even for $n_1$ and $n_2$ small. As a
representation of our results, we shall report and discuss the case
for only 5% significance in the next section. The results for
$\alpha = 2.5\%$ and 10% are also similar to that of $\alpha = 5\%$.

4. THE MONTE CARLO SIMULATION STUDY

4.1 Design of Simulations

In our Monte Carlo simulations, we have looked at two ranges
for $n$. These ranges are:

\[
\begin{array}{cccccc}
 n_1 & 40 & 24 & 16 & 12 & 10 \\
 n_2 & 10 & 12 & 16 & 24 & 40 \\
\end{array}
\]

\[
\begin{array}{cccc}
 n_1 & 20 & 12 & 8 & 6 \\
 n_2 & 5 & 6 & 8 & 12 & 10 \\
\end{array}
\]

The values for (4.1) and (4.2) are adjusted so that the harmonic
means $\bar{n}$ are 16 and 6, respectively.

The ranges for the variances are:

\[
\begin{array}{cccc}
 \sigma^2_1 & 100 & 60 & 55 \\
 \sigma^2_2 & 100 & 300 & 550 \\
\end{array}
\]

The values for (4.3) are adjusted so that the harmonic means $\bar{s}^2$ are
equal to 100 and have ratios 1:1, 1:5, and 1:10, respectively.

As is well known, the F-test is useful to test the null
hypothesis $H_0: \mu_1 = \mu_2$ (=$\mu$) under the assumption of equal variances
$s^2$. Also, as is well known, the power of the F-test depends on the
distribution of F when the null hypothesis is false. This
distribution is noncentral F with noncentrality parameter $\delta^2$ given
by
\[ \sigma^2 = \frac{(\mu_1 - \mu_2)^2}{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \]  

which is a measure of deviation of the alternative from the null hypothesis. Following Kohn and Games (1974), we compute average bias

\[ \phi^2 = \frac{\phi^2}{2} \]  

for our Monte Carlo power curves for \( \phi = 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, \) and 4.0. In equation (4.4), we use the harmonic mean \( \frac{1}{\sigma^2} \) in place of the common variance \( \sigma^2 \). Note that the harmonic mean \( \frac{1}{\bar{n}} = 2/(1/n_1 + 1/n_2) \). Thus,

\[ \phi^2 = \frac{1}{2} \frac{\bar{n}(\mu_1 - \mu_2)^2}{\sigma^2} \]  

We kept the harmonic means constant in selecting the parameters in an attempt to have the power curves to be compatible throughout.

We first created 18 files of 1000 random numbers each corresponding to the 9 values of \( \phi \) using the required values of \( \mu_1 \) and \( \mu_2 \) such that \( \mu_1 \) and \( \mu_2 \) are equally spaced about 100. and with \( \bar{\sigma}^2 = 100 \), for \( \bar{n} = 16 \) and 8, respectively. From (4.6), we have

\[ \phi^2 = \frac{1}{n} \left[ (\mu_1 - 100)^2 + (\mu_2 - 100)^2 \right] \]  

So, for example, for \( n_1 = 6 \), \( n_2 = 12 \), \( \bar{\sigma}^2 = 100 \), and \( \phi = 0.5 \), we solve (4.7) to get \( \mu_1 = 98.2322 \) and \( \mu_2 = 101.768 \).

The standard normal random variates were generated using the IMSL subroutine GGGNML with DSEED = 1234567.DO. (This is one of the suggested seeds which has been tested for normality by the chi-squared goodness of fit test.)
From each of the 18 files, we randomly chose subsets of sizes \( n_1 \) and \( n_2 \). The experiment was repeated 1000 times for each of the 18 cases. For each sample we computed:

1. \( t \), the t-test statistic given in (2.3) and the corresponding significance level.

2. the nonparametric Wilcoxon-Mann-Whitney (W-M-W) test statistic and the corresponding significance level (based on the asymptotic normality of the test statistic).

3. \( t_w \), the Welch-test statistic given in (2.5) and the corresponding significance level.

4. \( A \equiv -2 \ln \lambda \), the likelihood-ratio test statistic without adjustment, that is, the unadjusted likelihood-ratio test and the corresponding \( \chi^2 \) significance level.

5. \( A_{1} \), the first moment adjustment of \(-2 \ln \lambda \) given in (3.2) and the corresponding significance level, and finally

6. \( A_{1.2} \), the first and second moment adjustments of \(-2 \ln \lambda \) given in (3.6) and the corresponding significance level.

All computer programs were run on one of the University of Virginia's Prime 750 computers. The programs were written in FORTRAN 77 and compiled under FORTRAN 77, Version 17.3. The IMSL subroutines refer to Edition 9. Uniform random numbers were generated using the Prime FORTRAN subroutine RND. The IMSL subroutines required were MTD for computing the probabilities for the t-distribution, WDCH for computing the probabilities for the gamma distribution, and NRWRST for computing the probabilities for the Wilcoxon-Mann-Whitney test.

A listing of these interactive programs is available from the authors upon request.

We report and discuss our results obtained from this large-scale Monte Carlo simulation study next.

4.2 Discussion and Report of the Results

For each of the 18 cases and each of the 6 procedures, we recorded the number of times the null hypothesis \( H_0: \mu_1 = \mu_2 \) was
FIG. 4.1 Monte Carlo Power Curves for t-test for $n = 12, 6$ at $\alpha = 5\%$.

- $\sigma_1^2 : \sigma_2^2 = 1:1$
- $\sigma_1^2 : \sigma_2^2 = 1:5$
- $\sigma_1^2 : \sigma_2^2 = 1:10$
rejected at the nominal level of significance $\alpha = 5\%$. These values
determine the Monte Carlo power curves given in Figures 4.1 to 4.7.
For $n = 1000$, the 95% acceptable range for the observed
significance level is from .036 to .064.

The Monte Carlo power curves for $\alpha = 5.0\%$ with $\Psi = 0.0$ in
Figures 4.1 and 4.2, clearly show the nonrobustness inherent in the
classical t-test. The t-test performs well when $\sigma^2_1 : \sigma^2_2 = 1:1$, but
for $\sigma^2_1 : \sigma^2_2 = 1:5$ and $\sigma^2_1 : \sigma^2_2 = 1:10$, the significance level $\alpha$ jumps
from 12.1% to 15.6%, respectively. Therefore, comparing these
Monte Carlo power curves here, one might seriously question the
widespread use of the ordinary two-sample t-test, especially since
there is often no strong evidence on behalf of the assumption of
equal variances. These findings are also supported by Leaverton
and Birch (1969). The Monte Carlo power curves for $\alpha = 5\%$ for the
nonparametric Wilcoxon-Mann-Whitney test are shown in Figure 4.3.
It performs better than the ordinary t-test, although it, too, is
also nonrobust. In Figure 4.5, the unadjusted likelihood-ratio
procedure based on a $\chi^2_1$-distribution is clearly biased. The Welch
procedure in Figure 4.4, and the likelihood-ratio test adjusted to
the first moment in Figure 4.6, are more accurate and robust tests.
Comparing the Monte Carlo power curves in these figures, we clearly
see that both of these tests almost always hit the target at $\alpha = 5\%$
which we originally had set for these power curves. The likelihood-
ratio test adjusted to the first moment (Figure 4.6), with smaller
sample sizes, is superior to the likelihood-ratio test adjusted for
the first and second moments in Figure 4.7.

From these results, we note that for the Behrens-Fisher
problem the primary importance is how to control the level of
significance which we are able to do by our adjusted likelihood-
ratio test. The comparison of powers, on the other hand, is
secondary, since without controlling the size of tests one cannot
compare the powers.
FIG. 4.2 Monte Carlo Power Curves for $t$-test for $n_g = 6, 12$ at $\alpha = 5\%$.

- $\sigma_1^2 : \sigma_2^2 = 1:1$
- $\sigma_1^2 : \sigma_2^2 = 1:5$
- $\sigma_1^2 : \sigma_2^2 = 1:10$
FIG. 4.3 Monte Carlo Power Curves for Wilcoxon-Mann-Whitney (W-M-W) Nonparametric Test for $n_g = 12,6$ at $\alpha = 5\%$.

- $\sigma_1^2 : \sigma_2^2 = 1:1$
- $\sigma_1^2 : \sigma_2^2 = 1:5$
- $\sigma_1^2 : \sigma_2^2 = 1:10$
FIG. 4.4 Monte Carlo Power Curves for Welch's Test for $n_g = 12.6$, at $\alpha = 5\%$.

- $\sigma_1^2 : \sigma_2^2 = 1:1$
- $\sigma_1^2 : \sigma_2^2 = 1:5$
- $\sigma_1^2 : \sigma_2^2 = 1:10$
FIG. 4.5 Monte Carlo Power Curves for Unadjusted Likelihood-Ratio Test \( \Lambda = -2 \ln \lambda - \chi^2 \) for \( n_g = 12,6 \) at \( \alpha = 5\% \).

- \( \sigma_1^2 : \sigma_2^2 = 1:1 \)
- \( \sigma_1^2 : \sigma_2^2 = 1:5 \)
- \( \sigma_1^2 : \sigma_2^2 = 1:10 \)
FIG. 4.6 Monte Carlo Power Curves for Adjusted Likelihood-Ratio Test to the First Moment ($\Lambda_1 = -2 \ln \lambda - x^2_{f_1}$) for $n_g = 12.6$ at $\alpha = 5\%$.

- $\sigma_1^2 : \sigma_2^2 = 1:1$
- $\sigma_1^2 : \sigma_2^2 = 1:5$
- $\sigma_1^2 : \sigma_2^2 = 1:10$
FIG. 4.7 Monte Carlo Power Curves for Adjusted Likelihood-Ratio Test to the First and Second Moments \([\Lambda_{1,2} = -2\ln \lambda \cdot G(\alpha, \beta)]\) for \(n_g = 12.6\) at \(\alpha = 5\%\).

- \(\sigma_1^2 : \sigma_2^2 = 1:1\)
- \(\sigma_1^2 : \sigma_2^2 = 1:5\)
- \(\sigma_1^2 : \sigma_2^2 = 1:10\)
TABLE I

Maximum Likelihood Values with $\lambda = -2 \ln \lambda$
as a Gamma Distribution

<table>
<thead>
<tr>
<th>$n_1,n_2$</th>
<th>$\sigma^2_{1,2}$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
<th>$\hat{\mu} = \hat{\alpha}\hat{\beta}$</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>30,5</td>
<td>1:1</td>
<td>0.516</td>
<td>2.178</td>
<td>1.228</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>1:5</td>
<td>0.519</td>
<td>2.390</td>
<td>1.244</td>
<td>16%</td>
</tr>
<tr>
<td></td>
<td>1:10</td>
<td>0.523</td>
<td>2.648</td>
<td>1.384</td>
<td>40%</td>
</tr>
<tr>
<td>12,6</td>
<td>1:1</td>
<td>0.513</td>
<td>2.229</td>
<td>1.145</td>
<td>13%</td>
</tr>
<tr>
<td></td>
<td>1:5</td>
<td>0.532</td>
<td>2.342</td>
<td>1.246</td>
<td>64%</td>
</tr>
<tr>
<td></td>
<td>1:10</td>
<td>0.512</td>
<td>2.532</td>
<td>1.297</td>
<td>73%</td>
</tr>
<tr>
<td>8,8</td>
<td>1:1</td>
<td>0.530</td>
<td>2.232</td>
<td>1.197</td>
<td>60%</td>
</tr>
<tr>
<td></td>
<td>1:5</td>
<td>0.522</td>
<td>2.280</td>
<td>1.191</td>
<td>32%</td>
</tr>
<tr>
<td></td>
<td>1:10</td>
<td>0.517</td>
<td>2.296</td>
<td>1.152</td>
<td>21%</td>
</tr>
<tr>
<td>6,12</td>
<td>1:1</td>
<td>0.484</td>
<td>2.548</td>
<td>1.233</td>
<td>03%</td>
</tr>
<tr>
<td></td>
<td>1:5</td>
<td>0.499</td>
<td>2.246</td>
<td>1.119</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>1:10</td>
<td>0.517</td>
<td>2.088</td>
<td>1.080</td>
<td>24%</td>
</tr>
<tr>
<td>5,20</td>
<td>1:1</td>
<td>0.491</td>
<td>2.846</td>
<td>1.398</td>
<td>38%</td>
</tr>
<tr>
<td></td>
<td>1:5</td>
<td>0.504</td>
<td>2.317</td>
<td>1.117</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td>1:10</td>
<td>0.520</td>
<td>1.889</td>
<td>0.982</td>
<td>37%</td>
</tr>
</tbody>
</table>

5. VALIDATION OF THE MONTE CARLO SIMULATION STUDY

The Welch procedure is based on the assumption that $t_w$ in
(2.5) is approximately distributed as a $\chi^2$-distribution when $n_1$ and
$n_2$ are large. To correct to the first and second moments jointly,
we assumed that $\lambda = -2 \ln \lambda$ in (2.15) is approximately a gamma
distribution with parameters $\alpha$ and $\beta$ in (3.3). To validate our
assumption, we fitted the $\lambda = -2 \ln \lambda$ values with $\Phi = 0.0$ to gamma
distributions. The results are summarized in Tables I and II. The
"significance" in the last column of both tables, denotes the
significance level for the $\chi^2$-goodness-of-fit test using 20
intervals with equal probabilities.
The Monte Carlo study shows that the assumption that $A = -2 \ln \lambda$ as a gamma distribution is valid. We point out that the estimators of $\alpha$, $\beta$, and $\mu$ are sensitive to outliers since the gamma distribution is unbounded, that is, since $\hat{\lambda} < 1$. The estimator of the mean, $\hat{\mu}$, gives the computed value of the biasness or the amount at which the correction should be made in $A = -2 \ln \lambda$. For this, recall (3.2).

6. **A NUMERICAL EXAMPLE**

As an example of our methodology, we simulated a random sample of $n_1 = 5$ observations from $N(100,300)$, and $n_2 = 20$ observations from $N(100,60)$. The data set is as follows:
TABLE III

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>131.7</th>
<th>88.9</th>
<th>112.6</th>
<th>103.0</th>
<th>117.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_2 )</td>
<td>86.8</td>
<td>97.7</td>
<td>92.6</td>
<td>97.9</td>
<td>109.9</td>
</tr>
<tr>
<td></td>
<td>117.3</td>
<td>83.4</td>
<td>93.8</td>
<td>96.9</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td>102.0</td>
<td>101.7</td>
<td>93.4</td>
<td>91.7</td>
<td>109.9</td>
</tr>
<tr>
<td></td>
<td>109.8</td>
<td>99.8</td>
<td>89.6</td>
<td>101.3</td>
<td>94.2</td>
</tr>
</tbody>
</table>

Under the B-F null hypothesis \( H_0: \mu_1 = \mu_2 \), the maximum likelihood estimators (2.13) for \( \mu \) and \( \sigma^2 \) (g = 1.2) are:

\[
\begin{align*}
\hat{\sigma}_1^2 &= 346.9 \\
\hat{\sigma}_2^2 &= 69.1 \\
\hat{\mu} &= 98.9 \\
\end{align*}
\]  

(6.1)

Under the alternative hypothesis \( H_a: \mu_1 \neq \mu_2 \), the maximum likelihood estimators for \( \mu \) and \( \sigma^2 \) (g = 1.2) are:

\[
\begin{align*}
\hat{\mu}_1 &= 110.8 \\
\hat{\sigma}_1^2 &= 205.3 \\
\hat{\mu}_2 &= 98.3 \\
\hat{\sigma}_2^2 &= 68.8 \\
\end{align*}
\]  

(6.2)

For each of the six methods we have discussed, in Table IV we present our results for this particular example.

7. SUMMARY

In this paper we have shown that the often-used Welch's procedure can be viewed as an approximation to the likelihood ratio
### TABLE IV

Results of the Numerical Example

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Statistics</th>
<th>Parameters</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-test</td>
<td>$t = 2.44$</td>
<td>$df = 23.00$</td>
<td>$.023$</td>
</tr>
<tr>
<td>W-M-W</td>
<td>$W = 92.0$</td>
<td></td>
<td>$.071$</td>
</tr>
<tr>
<td>Welch</td>
<td>$t_W = 1.68$</td>
<td>$df = \nu_W = 4.58$</td>
<td>$.158$</td>
</tr>
<tr>
<td>Likelihood-Ratio (Unadjusted)</td>
<td>$A = 2.72$</td>
<td>$df = 1.00$</td>
<td>$.099$</td>
</tr>
<tr>
<td>Likelihood-Ratio (Adj. to first moment)</td>
<td>$A_1 = 2.72$</td>
<td>$df = f_1 = 1.78$</td>
<td>$.218$</td>
</tr>
<tr>
<td>Likelihood-Ratio (Adj. to first and second moments)</td>
<td>$A_{1,2} = 2.72$</td>
<td>$\alpha = .0809$</td>
<td>$.127$</td>
</tr>
</tbody>
</table>

We have used this fact to derive an adjusted likelihood-ratio test with respect to the first moment, and with respect to the first and second moments simultaneously. These procedures allow the significance level to be adjusted in accordance with the degrees of freedom to balance the risk due to the bias in using the maximum likelihood estimates and the risk due to the increase of variance.

Our Monte Carlo simulation studies clearly show that the adjusted likelihood-ratio procedure yield excellent power curves which are robust and which do not depend on any arbitrary and dubious assumptions for the equality of variances, or for the group sample sizes.

As a result of this paper, we will show in forthcoming articles how to modify this adjusted likelihood-ratio procedure to
the Behrens-Fisher problem from a model selection viewpoint using an information-theoretic approach via Akaike's Information Criterion (AIC).

ACKNOWLEDGEMENTS

The authors extend their appreciation to Professor Nariaki Sugiuira and the referee for the detailed comments and suggestions which have resulted in substantial improvements of the paper. We also wish to thank Professor David R. Cox for his personal correspondence and suggestions, and to Professors Marvin Rosenblum and Eugene C. Paige for their encouragement during the preparation of this paper. We also extend our thanks to Ms. Julie Riddleberger for her excellent typing.

BIBLIOGRAPHY


Received by Editorial Board member June, 1984; Revised March, 1986.

Recommended by Naoki Sugiura, University of Tsukuba, Sakura-mura Niihari-Gun Ibaraki, Japan.

Recommended by Yoshinada Shibata, National Institute for Minamata Disease, Kumamoto, Japan.