DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 1 | Issue 2

Article 55

11-1-2002

Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When σ_1^2≠σ_2^2

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Recommended Citation

Sawilowsky, Shlomo S. (2002) "Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When σ_1^2≠σ_2^2," *Journal of Modern Applied Statistical Methods*: Vol. 1: Iss. 2, Article 55.
DOI: 10.22237/jmasm/1036109940
Available at: http://digitalcommons.wayne.edu/jmasm/vol1/iss2/55

Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 461-472

1538 – 9472/02/$30.00

Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When $\sigma_1^2 \neq \sigma_2^2$

Shlomo S. Sawilowsky
Educational Evaluation and Research
Wayne State University

The history of the Behrens-Fisher problem and some approximate solutions are reviewed. In outlining relevant statistical hypotheses on the probable difference between two means, the importance of the Behrens-Fisher problem from a theoretical perspective is acknowledged, but it is concluded that this problem is irrelevant for applied research in psychology, education, and related disciplines. The focus is better placed on "shift in location" and, more importantly, "shift in location and change in scale" treatment alternatives.

Key words: Behrens-Fisher problem, t test, heterogeneous variances.

Introduction

Simply stated, the Behrens-Fisher problem arises in testing the difference between two means with a t test when the ratio of variances of the two populations from which the data were sampled is not equal to one. This condition is known as heteroscedasticity, which is a violation of one of the underlying assumptions of the t test. The resulting statistic is not distributed as t, and therefore the associated p values based on the entries found in standard t tables are incorrect. Use of tabulated critical values may lead to increased false positives, which are known as Type I errors, or a conservative test that lacks statistical power to detect significant treatment effects.

Development of Student's Distribution For a Unique Sample

Regarding the development of the t test, Fisher (1939) noted,

> To the present generation of statisticians, familiar with 'Student's' distribution..., it has for some time appeared to be a somewhat puzzling historical fact that this advance in simple statistical procedure was not made long before, and was not made rather by a mathematician than a research chemist.
>
> Light is perhaps thrown on this puzzle by the contrast, which has been striking during the last twenty years, between the facility, confidence, and skill with which the new tests have been applied by practical men in research departments, and the embarrassment and confusion of many discussions, in journals devoted to mathematical statistics, by mathematically minded authors lacking contact with practical research (p. 141).

Prior to 'Student' or W. S. Gosset, the mathematician Helmert was able to determine the distribution of the sum of squares $\sum(x-\mu)^2$ (Helmert, 1875) and $\sum(x-\bar{x})^2$ (Helmert, 1876), but indicated no practical value for the results. Subsequent to Gosset, another mathematician, Burnside (1923), used Bayesian methods in rediscovering the t distribution, although the

Shlomo S. Sawilowsky is Professor of Educational Evaluation & Research, & Wayne State University Distinguished Faculty Fellow. His current areas of interest are nonparametric and computer intensive methods, the revival of classical measurement theory, and a recommitment to experimental design in lieu of quasi-experimental design. Email: shlomo@wayne.edu.

inclusion of an à priori distribution for a precision constant resulted in a difference of one degree of freedom. Interestingly, he presented a table of quartiles of the t distribution, prompting Fisher (1941) to remark, "It evidently did not occur to him that a 5 or 1% table would be more useful...[this] may be taken to indicate that he regarded his solution rather as a matter of academic interest than as meeting a need for guidance in practical decisions" (p. 142).

According to Jeffreys (1937), the t distribution was not discovered earlier because it "involves an unstated assumption" (p. 48) that for the sample mean ($\bar{x}$), estimated variance of the mean (s^2), and population mean (μ), then the distribution of

$$t = \frac{\bar{x} - \mu}{s} \qquad (1)$$

depends only on the sample size n. Fisher (1941) added that novel reasoning also left unstated by Gosset was that $\bar{x}$ and s^2 should be unbiased.

The question of bias in s^2 was troublesome indeed. The prepublication title of "The Probable Error of a Mean" (Student, 1908) was "On the Probable Error of a Unique Sample". The uniqueness that worried Gosset was the requirement that s^2 be unbiased. Although Gosset's paper pertained to the difference distribution of paired observations, Fisher (1941) extended this concern to the two independent samples case. Fisher suggested that one of the "difficulties in the way of an early discovery of 'Student's' test" was because of "the application of the same methods to the more intricate problem of the comparison of the means of samples having unequal variances, or more correctly from populations, of which the variance ratio is unknown, and itself constitutes one of the parameters which require to be 'Studentized'"(1941, p. 146).

The Behrens-Fisher Problem

The first expression and solution to this problem was by Behrens (1929), and reframed by Fisher (1939a) from a Fisherian perspective as

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{2n_1 + 1} + \dfrac{s_2^2}{2n_2 + 1}}} \qquad (2),$$

where s_1 and s_2 are fixed and σ_1 and σ_2 have fiducial distributions. Tables of critical values were given in Fisher and Yates (1957). This solution was challenged by Bartlett (1936) on the principle of inverse probability from a Bayesian perspective. Fisher responded with his usual tenacious and acrid style: "From a purely historical standpoint it is worth noting that the ideas and nomenclature for which I am responsible were developed only after I had inured myself to the absolute rejection of the postulate of Inverse Probability" (1937a, p. 151; see also 1937b, 1939b). Jeffreys (1940) restored calm by demonstrating that Bartlett's perspective was not a challenge to the Fisherian approach, but rather was another way of starting with the same hypothesis and ending with the same conclusion.

Commonly available solutions implemented in computer software statistics packages have eschewed both of those approaches in favor of a third theoretical perspective. This is the frequentist approach of Neyman-Pearson, where σ_1 and σ_2 are fixed, but s_1 and s_2 are free to vary in (2). The typical solution in statistics packages for solving the two sample problem (k = 2) is the Welch separate variances test, which has become known as the Welch-Aspin test with modified degrees of freedom, given by

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad (3).$$

(Welch, 1937, 1949a, 1949b; Satterthwaite, 1941, 1946; Aspin 1948, 1949). Although the exact distribution of the Welch statistic is known under normality (Ray & Pitman, 1961), it remains an approximate solution to the Behrens-Fisher problem. Welch (1947) also provided a solution for the generalized problem ($k \geq 2$).

The Behrens-Fisher problem continued to attract the attention mathematical statisticians and applied researchers. For example, different perspectives were given by Wald (1955), Banerjee (1960), and Pagurova, (1968). These are but a few of the many solutions published in the literature.

Robustness With Respect To Unequal n's and Population Normality

Eventually, however, questions arose on the robustness with respect to Type I errors for unequal n's. Fisher (1939a) tried to quash this line of research by restating the fact that Gosset's paper (Student, 1908) was on pairs of measurements (height vs length of middle finger for 3,000 criminals), obviating the unequal n problem. Nevertheless, in the context of $k \geq 2$ independent samples, studies indicated that the various solutions were not robust to unequal n's (e.g., Kohr, 1970; Mehta & Srinivasa, 1970; Kohr & Games, 1974; Tomarkin & Serlin, 1986). Solutions to the unequal n situation appeared which preserved nominal alpha (e.g., Scheffé, 1943; McCullough, Gurland, & Rosenberg, 1960), although some of them were subsequently found to be not very powerful.

This line of research was soon overshadowed by the concern of robustness with respect to Type I errors for departures from population normality. Monte Carlo studies showed that the Behrens-Fisher, Bartlett, and Welch-Aspin/Satterthwaite approximate solutions are not robust to departures from normality (e.g., James, 1959; Yuen, 1974). A similar fate awaited many of the other solutions, such as the Brown & Forsythe (1974) test (Clinch & Keselman, 1982), and the H_m test by Wilcox (1990) which had "the tendency to be conservative" (Oshima & Algina, 1992, p. 262) for long-tailed distributions. The inability of these procedures to maintain the Type I error rate at nominal alpha created the opportunity for another round of alternative solutions being published.

Some solutions based on nonparametric or nonparametric-like procedures were unsuccessful. For example, Pratt (1964) showed that the Mann-Whitney U (Mann & Whitney, 1947) and the expected normal scores test (Hájek & Sidák, 1967) resulted in nonrobust Type I error rates. Bradstreet (1997) found the rank transform test (Conover & Iman, 1982) to result in severely inflated Type I error rates. For the case of $k > 2$, Feir-Walsh and Toothaker (1974) and Keselman, Rogan, and Feir-Walsh (1977) found the Kruskal-Wallis test (Kruskal & Wallis, 1952) and expected normal scores test (McSweeney & Penfield, 1969) to be "substantially affected by inhomogeneity of variance" (p. 220).

Other nonparametric solutions met with more success. Yuen (1974) provided a robust solution based on trimmed means and matching sample variances. Tiku and Singh's (1981) solution was based on modified maximum likelihood estimators. Tan and Tabatabai (1985) combined the Tiku and Singh procedure with the Brown-Forsythe test to produce a more powerful procedure than those based only on Huber's M estimator (Huber, 1981; Schrader & Hettmansperger, 1980).

The development of procedures involving the Behrens-Fisher problem is not restricted to the usual $k \geq 2$ independent samples case. Games and Howel (1976) examined pairwise mulitiple comparison solutions. Bozdogan and Rameriz (1986) proposed a likelihood ratio for situations where only subsets respond to a treatment. Johnson and Weerahandi (1988) provided a Bayesian solution to the multivariate problem. Koschat and Weerahandi (1992) developed a class of tests for the problem of inference for structural parameters common to several regressions.

Despite the many approximate solutions published to date, the Behrens-Fisher problem remains actively studied. In the past 35 years, there were 37 doctoral dissertations completed pertaining to some aspect of the Behrens-Fisher problem, including newly proposed approximate solutions (*Dissertation Abstracts Online*, 2000).There was one dissertation completed in the 1960s, six in the 1970s, 16 in the 1980s, and 14 in the 1990s.

Hypothesis Testing

Consider the entries in Table 1. It contains the various hypotheses on the probable error of a mean, and the probable difference between two means. Hypotheses #1-#3 rarely occur in applied studies because they pertain to the Z test which requires σ^2 to be known. It is unusual for a social and behavioral science researcher to have the entire population at her or his disposal, or to know the parameters of the population. Z tests are valuable mainly as a pedagogical tool for introducing inferential statistics to students of data analysis methods.

Table 1. Parametric Nondirectional (Two-Sided) Null (H_o:) And Alternative (H_a:) Hypotheses For One Sample (μ_0) And Two Samples (μ_1, μ_2) Z And t Tests.

<u>Z tests: Hypotheses That Rarely Occur In Applied Studies</u>

#1:	H_o: $\mu_1 = \mu_0$; σ^2 is known
	H_a: $\mu_1 \neq \mu_0$; σ^2 does not change
#2:	H_o: $\mu_1=\mu_2$; $\sigma_1^2=\sigma_2^2$ and known
	H_a: $\mu_1 \neq \mu_2$; σ_1^2 and σ_2^2 do not change
#3:	H_o: $\mu_1=\mu_2$; $\sigma_1^2 \neq \sigma_2^2$, but known
	H_a: $\mu_1 \neq \mu_2$; σ_1^2 and σ_2^2 do not change

<u>t tests: Hypotheses That Occur In Applied Studies - The "Shift in Location Alternative"</u>

#4:	H_o: $\mu_1=\mu_0$; σ^2 is unknown, but assumed to be unbiased
	H_a: $\mu_1 \neq \mu_0$; σ^2 does not change
#5:	H_o: $\mu_1=\mu_2$; σ_1^2 and σ_2^2 are unknown, but assumed to be equal
	H_a: $\mu_1 \neq \mu_2$; σ_1^2 and σ_2^2 do not change

<u>The Two Sample Behrens-Fisher Problem (Fisherian & Bayesian)</u>

#6a:	H_o: $\mu_1=\mu_2$; σ_1^2 and σ_2^2 are unknown, but it is known that $\sigma_1^2 \neq \sigma_2^2$
#6b:	H_o: $\mu_1=\mu_2$; σ_1^2 and σ_2^2 are unknown, but cannot be assumed to be equal

<u>The Two Sample Behrens-Fisher Problem (Neyman-Pearson)</u>

#6c:	H_o: $\mu_1=\mu_2$; σ_1^2 and σ_2^2 are unknown, but it is known that $\sigma_1^2 \neq \sigma_2^2$
	H_a: $\mu_1 \neq \mu_2$; σ_1^2 and σ_2^2 do not change
#6d:	H_o: $\mu_1=\mu_2$; σ_1^2 and σ_2^2 are unknown, but cannot be assumed to be equal
	H_a: $\mu_1 \neq \mu_2$; σ_1^2 and σ_2^2 do not change

<u>Hypotheses That Frequently Occur in Applied Studies: The "Shift in Location and Change in Scale" Alternative</u>

#7:	H_o: $\mu_1=\mu_2$ and $\sigma_1^2=\sigma_2^2$
	H_a: $\mu_1 \neq \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$

Note: Ha: can be expressed as a directional (one-sided) hypothesis by replacing "≠" with either ">" or "<".

Hypotheses #4 and #5 refer to the "shift in location" alternative and are tested by the t test. Although no test can survive violations of independence of observations, under certain commonly occurring conditions (i.e., sample sizes are equal or nearly so and are at least 25 to 30, and tests are two-tailed rather than one-tailed), the t test is remarkably robust with respect to both Type I and II errors for departures from normality (e.g., Sawilowsky, 1990; Sawilowsky & Blair, 1992).

Editors and reviewers challenge the shift alternative as a realistic treatment outcome, which in turn, questions the applicability of Hypotheses #4 and #5 to real world data sets. After studying the histograms of many real treatment vs control and pretest-posttest data sets, I argue that, indeed, shift happens. An example with 714 admit vs discharge Functional Independence Measure scores (Keith, Granger, Hamilton, & Sherwin, 1987), an instrument that is frequently used in the field of rehabilitation counseling, was shown in Nanna and Sawilowsky (1998).

(I would be remiss if I failed to note that numerous Monte Carlo studies have shown that the nonparametric Wilcoxon Rank Sum test can be three to four times more powerful in detecting differences in location parameters when the normality assumption was violated (e.g., Blair & Higgins, 1980a, 1980b, 1985; Blair, Higgins, & Smitley, 1980; Sawilowsky & Blair, 1992). Micceri (1989) found that only about 3% of real data sets in psychology and education are relatively symmetric with light tails. Therefore, the Wilcoxon procedure should be the test of choice. The t test remains a popular test, however, most likely due to the inertia of many generations of classically parametrically trained researchers who continue its use for this situation.)

As noted by #6a - #6d, the hypotheses tested by the Behrens-Fisher problem can be expressed from the Fisherian/Bayesian perspective by the absence of an alternative hypothesis, or in the Neyman-Person frequentist paradigm. In the first example according to both perspectives (i.e., #6a and #6c), it is known that samples were drawn from two different populations (e.g., the first may have been extreme asymmetric such as exponential decay and the second may have been multimodal from a likert scale), but the population parameters remain unknown. Thus, the Behrens-Fisher problem arises because the ratio of population variances is different from one, although neither constituent value is known. The second and more common example, according to both perspectives (i.e., #6b and #6d), indicates that no information is available on the population from which the samples were drawn, and it cannot be safely assumed that the ratio of population variances is equal to one. Now, I discuss two reasons why these situations are important, and two reasons why they are irrelevant to applied researchers.

Two Reasons Why The Behrens-Fisher Problem Is Important

1. The Behrens-Fisher problem is a classic. Many prestigious mathematical statisticians and applied researchers have addressed this problem. For some, their careers began with this problem; for others, their careers ended with this problem. The Behrens-Fisher problem has as much mystique and has received as much fanfare in its discipline as other classical problems that remain unsolved or unfinished in their disciplines, such as these:

- In 1630, Pierre de Fermat, an amateur mathematician, wrote "hanc marginis exigiutas non caperet" - he found a proof that was too large to write in a marginal note in his copy of the ancient Greek Diophantus' *Arithmetica* that $x^n + y^n = z^n$ has no nonzero integer solutions for x, y and z when n>2. In October, 1994, the mathematician Andrew Wiles solved the final aspect of this conjecture. (Fermat's last conjecture is a special case of $x^n + y^n = cz^n$, which remains unproven.) However, Wiles noted, "Fermat couldn't possibly have had this proof. It's a 20th-century proof. There's no way this could have been done before the 20th-century" (Wiles, 1996).Thus, the conjecture remains unproven using 17th century mathematics.

- In 1822, Franz Schubert wrote what was later to be known as the 'Unfinished' Symphony No. 8 (or No. 7 according to some numbering schemes) in B Minor. He worked on it for six years, but only completed the first two movements of an intended four movement symphony. Mysteriously and uncharacteristically, he moved on to other pieces without finishing this symphony. Many musicians have written what they imagine the final two movements might have been if Schubert had finished it.

- In the 20th Century, physicists theorized on the unification of the laws of the universe. However, the solution eluded physicists from Albert Einstein to Stephen Hawkings. (The so-called "Grand Unification Theories" combine the weak, strong, and electromagnetic forces, but leave out gravity.)

2. The second reason that the Behrens-Fisher problem is important is due to the byproducts that have been developed in the course of creating approximate solutions. Some examples include:

- Bartlett's (1937) study of heteroscedasticity culminated in a well known Chi-Squared test on variances, which is useful for testing the underlying assumption of homoscedasticity. Bartlett's test is a logarithmic modification of the Neyman and Pearson (1931) L_1 test for the equality of variances of k groups.

- James' (1959) attempt to improve on the Behrens-Fisher, Welch, and Yates (1939) solutions led to the development of a Cornish-Fisher expansion for a symmetric distribution.

- Statistics were developed throughout the 20th Century based on asymptotic or large sample theory. Many were published based on elegant mathematical statistical theory, but turned out to be invalid for use in applied work. The Behrens-Fisher problem highlighted the importance of conducting robustness and comparative power studies relative to small samples.

(Regarding the last point, my recommendation is that authors of new statistics or procedures publish their work *after* they have conducted studies on the properties of the statistic when underlying assumptions are violated. Note that further study is moot if results for expedient mathematical distributions produce poor results; but if good results are obtained, verification is still required with real data sets.)

Two Reasons Why The Behrens-Fisher Problem Is Irrelevant

1. Howell and Games (1974) suggested that "Educational and psychological researchers often deal with groups that tend to be heterogeneous in variability" (p. 72). This is mitigated by the fact that, "We have spent many years examining large data sets but have never encountered a treatment or other naturally occurring condition that produces heterogeneous variances while leaving population means exactly equal" (Sawilowsky and Blair, 1992, p. 358). None of Micceri's (1989) 440 real psychology and education data sets reflected this condition, nor have I seen an example in the literature. Thus, the issue of heterogeneous variances and their impact on Type I errors is moot.

Zumbo and Coulombe (1997) demurred, and claimed "We could simply counter that in our experience we have seen it occur" (p. 148), but there was no data set in their article. Algina and Olejnik (1984) referred to a data set in Box and Cox from 1964, but the reference is missing from their bibliography. The ratios of minimum (0.0001) to maximum (0.1131) variances given for the 12 entries in their 3×4 layout are impressive; the frequency with which psychological and educational instruments produce variances less than one-twelveth of a single point remains problematic. Koschat and Weerahandi (1992) refer to what appears to be a real data set from business and economics, although they only published summary statistics and not the actual data set. Even if examples can be found, the question remains if the Behrens-Fisher problem surfaces with such frequency that merits the journal space it has been given.

2. The most prolific treatment outcome in applied studies ***is*** known. It is where a change in scale is concomitant with a shift in means. As an intervention is implemented, the means increase or decrease according to the context. Simultaneously, the treatment group may become more homogeneous on the outcome variable due to sharing the same intervention, method, conditions, etc. Alternatively, the group may become more heterogeneous, as some respond to the treatment while others do not respond, or even regress.

What Is Wrong With Testing For Homogeneity Prior To The t-Test?

A common strategy is to conduct a test on variances prior to the pooled samples t test (e.g., SAS, 1990, p. 25; SPSS, 1993, p. 254-255; SYSTAT, 1990, p. 487). If the F test on variances, for example, is not significant, then the researcher continues with the t test. However, if the F test is significant, then the researcher is advised to conduct the separate variances t test (e.g., Welch-Aspin) with modified degrees of freedom.

There is a serious problem with this approach that is universally overlooked. The sequential nature of testing for homogeneity of variance as a condition of conducting the independent samples t test leads to an inflation of experiment-wise Type I errors. A small Fortran program was written, compiled, and executed to demonstrate this, with the results noted in Table 2.

Table 2. Type I Error And Power For The Pooled-Variances Independent Sample t-test Conducted Unconditionally Or Conditionally On The F Test For Homogeneity Of Variance, $\alpha = 0.050$; $n_1 = n_2 = 5$, 100,000 Repetitions.

	t-test				F-test
	Unconditional		Conditional		Type I Error
	L	R	L	R	
Distribution Normal					
c=0.0	.025	.025	.023	.023	.051
c=0.95	.000	.265	.000	.252	
c=2.0	.000	.790	.000	.750	
Chi-Square (ν=2)					
c=0.0	.023	.019	.015	.013	.172
c=1.5	.000	.252	.000	.202	
c=3.5	.000	.735	.000	.632	

Note: "c" = shift in location to produce approximately small or large Effect Sizes. A study of robustness with respect to Type II errors requires "c" to represent equal Effect Sizes across distributions, which was not done for this illustration. "L" = left tail. "R" = right tail.

An examination of Table 2 highlights a number of important points:

- The experiment-wise Type I error rate, under normality, is .097 (.051+.023+.023) when the t test is conducted conditional on the F test for homogeneity of variance. This is almost twice nominal alpha.

- The experiment-wise Type I error rate when the data were sampled from a Chi-Squared distribution (ν=2) is .200, which is four times nominal alpha!

- The F test on variances, as is well known, is nonrobust to departures from normality. In this case the Type I error rate for Gaussian data of 0.051 ballooned up to .172 for the Chi-Squared (ν=2) data. This inflation level of about 3.5 times nominal alpha means the data analyst will frequently abandon the pooled samples t test in favor of the separate variances test, when in fact, the condition of homoscedasticity holds. This problem can be ameliorated somewhat by using Levene's (1960) test, which is more robust to departures from normality.

- Conducting the t-test conditioned on the F test for variances resulted in a 5% loss of power under normality, which is ill afforded in small samples applied research.

- Conducting the t-test conditioned on the F test for variances resulted in a 20% loss of power under the Chi-Squared (ν=2) distribution for the small Effect Size, and a 14% loss in power for the large Effect Size, which is ill afforded in small samples applied research.

Hyman (1995) opined that methodology articles are less helpful when they are restricted to pointing out errors or deficiencies, and are more helpful when they redirect researchers toward a useful methodology. Given the severity of the problem of pursuing Hypothesis #6 sequentially after a test on variances, it is appropriate to review Hypothesis #7 in more detail.

Refocusing On Treatments That Impact Location And Scale

Hypothesis #7 pertains to the situation where naturally occurring differences or treatment outcomes produce a shift in location and a change in scale. Diamond (1981, p. 73-74) discussed a simple procedure where variances and means are tested separately. What is needed, however, is a test of both parameters simultaneously. Lepage (1971, 1975), Gastwirth and Podgor (1992), and Podgor and Gastwirth (1994) offered some early work and hypothesis tests that depend on location and scale. Two more recently developed statistics for Hypothesis #7 were given by O'Brien (1988) and Brownie, Boos, and Hughes-Oliver (1990). They are discussed below because they are promising for small samples applied research.

(1) O'Brien's (1988) generalized t-test is carried out by ordinary least squares or logistic regression. In terms of the former, a dummy variable of 1, representing group membership, or 0, representing nonmembership, is regressed on the outcome variable, w, as well as w^2:

$$y' = \beta_o + \beta_1 w + \beta_2 w^2 \qquad (4).$$

If β_2 is not near zero, the test for treatment effects is conducted with the 2 degrees of freedom F test of H_o:$\beta_1 = \beta_2 = 0$. If β_2 is near 0, however, (4) is replaced with

$$y' = \beta_o + \beta_1 w \qquad (5),$$

and the one degree of freedom test of H_o: $\beta_o = 0$, an independent samples t test, is conducted. It is called a generalized t-test because of the variety of levels of nominal α which may be selected for testing (4).

Blair and Morel (1991) examined the experiment-wise Type I error rate of conducting (5) conditional on (4). The sequential conditional testing procedure resulted in inflated Type I errors. Grambsch and O'Brien (1991) provided a "2/3" rule, where approximately correct Type I errors are obtained by reducing alpha to two-thirds of the desired size. Subsequently, a superior solution was made available by Blair (1991), who provided a corrected table of critical values for O'Brien's procedure which results in correct Type I error rates.

(2) Brownie, Boos, and Hughes-Oliver (1990) provided a modification to the t test:

$$t^* = \frac{\overline{x}_1 - \overline{x}_2}{s_1^2 \sqrt{\frac{1}{n_1} \times \frac{1}{n_2}}} \quad (6),$$

where s_1^2 is the sample variance from the control group, and $\nu = n_1 - 1$. Subsequently, Sawilowsky et al. (1991) and Blair and Sawilowsky (1993a, 1993b) demonstrated through Monte Carlo methods that t* is not robust with respect to Type I errors for departures from population normality. In addition, it requires that the change in scale increase, but not decrease. Blair and Sawilowsky (1992a, 1992b, 1993a, 1993b) fixed the Type I error properties by developing two new tests based on t* and F*, the extension based on k >2. In the context of F*, the first test is a permutation analogue (pF*), which does not require à priori knowledge of the expected change (i.e., increase or decrease) in variability relative to the control groups.

The second (pF^*_{min}) designates the group with the smallest variance as the control group, and substitutes s_{min}^2 for s_1^2 in (6). (Both procedures can also be conducted as an approximate randomization test with negligible loss in precision or power.) These tests and other procedures were examined further by Troendle, Blair, Rumsey, and Moke (1997).

Podgor and Gastwirth (1994) compared O'Brien's test with Brownie, Boos, Hughes-Oliver's test in various configurations. However, they did not use Blair's corrected critical values or Blair and Sawilowsky's approximate randomization correction. One of my doctoral students is comparing both procedures with their respective corrections with two nonparametric tests. One statistic is the Savage test for positive random variables (which received some attention by Podgor & Gastwirth, 1994). It assumes that a difference in scale causes a difference in location (see, e.g., Deshpande, Gore, & Shanubhogue, 1995, p. 53-56). The other is the Rosenbaum test for general differences (see, e.g., Neave & Worthington, 1988, p. 144-149).

Conclusion

The Behrens-Fisher problem is a classic, but its many and continuing solutions are perhaps better housed in journals catering to theoretical developments. Sufficient journal space has been given to this problem in comparison with the frequency with which it occurs. Instead, applied researchers should focus on more practical treatment outcomes, such as a treatment or naturally occurring condition that brings about a shift in location and a change in scale. This is the most realistic treatment outcome in applied psychology and education research. It presents an exciting area in which considerable additional research is warranted.

References

Algina, J., & Olejnik, S. F. (1984). Implementing the Welch-James procedure with factorial designs. *Educational and Psychological Measurement*, *44*, 39-48.

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

Aspin, A. A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. *Biometrika*, *36*, 290-296.

Banerjee, S. K. (1960). Approximate confidence intervals for linear functions of means of k populations when the population variances are not equal. *Sankhyā*, *22*, 357-358.

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

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Royal Society of London Proceedings, Series A*, *160*, 268-282.

Behrens, W. -V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwwirtschaftliche Jahrbücher*, *68,* 807-837.

Blair, R. C. (1991). New critical values for the generalized t and generalized rank-sum procedures. *Communications in Statistics*, *20*, 981-994.

Blair, R. C., & Higgins, J. J. (1980a.) A comparison of the power of the t test and the Wilcoxon statistics when samples are drawn from certain mixed normal distributions. *Evaluation Review*, *4*, 645-656.

Blair, R. C., & Higgins, J. J. (1980b). A comparison of the power of the Wilcoxon's rank-sum statistic to that of student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, *5*, 309-335.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, *97*, 119-128.

Blair, R. C., & Morel, J. G. (1991). On the use of the generalized t and generalized rank-sum statistics in medical research. *Statistics in Medicine*, *11*, 491-501.

Blair, R. C., & Sawilowsky, S. S. (1992a). A comparison of the generalized and modified t tests. Annual meeting of the American Educational Research Association, SIG Educational Statisticians, San Francisco, CA.

Blair, R. C., & Sawilowsky, S. S. (1992b). Type I error and power of the modified and generalized t tests. 1992 Abstracts: Joint Statistical Meetings of the American Statistical Association, Biometrics Society, and Institute of Mathematical Statistics. Boston, MA, p. 49.

Blair, R. C., & Sawilowsky, S. S. (1993a). A note on the operating characteristics of the modified F test. *Biometrics*, *49*, 935-939.

Blair, R. C., & Sawilowsky, S. S. (1993b). Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls. *Statistics in Medicine*, *12*, 2233-2243.

Blair, R. C., Higgins, J. J., & Smitley, W. D. S. (1980). On the relative power of the U and t tests. *British Journal of Mathematical and Statistical Psychology*, *33*, 114-120.

Bozdogan, H., & Ramirez, D. E. (1986). An adjusted likelihood-ratio approach to the Behrens-Fisher test. *Communications in Statistics*, *15*, 2405-2433.

Bradstreet, T. E. (1997). A Monte Carlo study of type I error rates for the two-sample Behrens-Fisher problem with and without rank transformation. *Computational Statistics and Data Analysis*, *25*, 167-179.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*, 129-132.

Brownie, C., Boos, D. D., & Hughes-Oliver, J. (1990). Modyifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls. *Biometrics*, *46*, 259-266.

Burnside, W. (1923). On errors of observation. *Proceedings of the Cambridge Philosophical Society*, *21*, 482-487.

Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics*, *7*, 207-214.

Conover, W. J., & Iman, R. I. (1982). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*, 124-129.

Deshpande, J. V., Gore, A. P., & Shanubhogue, A. (1995). *Statistical analysis of nonnormal data*. NY: John Wiley & Sons.

Diamond, W. J. (1981). *Practical experiment designs for engineers and scientists*. Belmont, CA: Lifetime Learning Publications, Wadsworth.

Dissertation Abstracts Online. (2000). http://firstsearch.oclc.org.

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement*, *34*, 789-799.

Fisher, R. A. (1937a). Editorial note. *Annals of Eugenics*, *7*, 146-151.

Fisher, R. A. (1937b). On a point raised by M. S. Bartlett on fiducial probability. *Annals of Eugenics*, *7*, 370-375.

Fisher, R. A. (1939a). The comparison of samples with possibly unequal variances. *Annals of Eugenics*, *9*, 174-180.

Fisher, R. A. (1939b). A note on fiducial inference. *The Annals of Mathematical Statistics*, *10*, 383-388.

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

Fisher, R. A., & Yates, F. (1957). *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver & Boyd.

Games, P. A., Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, *1*, 113-125.

Gastwirth, J. L., & Podgor, M. J. (1992). Efficient robust rank tests for the location-scale problem. In Saleh, A. K. Md. E. (ed.) *Nonparametric statistics and related topics*. Amsterdam: Elsevier, p. 17-31.

Grambsch, P., & O'Brien, P. (1991). The effects of transformations and preliminary tests for non-linearity in regression. *Statistics in Medicine*, *10*, 697-709.

Hájek, J., & Sidák, F. (1967). *Theory of rank tests*. Prague: Academic Press and Academia.

Helmert, C. F. (1875). Über die Berechnung des wahrscheinlichen Fehlers aus einer endlichen Anzahl wahrer Beobachtungsfehler. *Zeitschrift für Mathematik und Physik*, *20*, 300-303.

Helmert, C. F. (1876). Über die Wahrscheinliehkeit der Potenzsummen der Beobachtungsfehler und über einige damit in Zusammenhang stehende Fragen, *Zeitschrift für Mathematik und Physik*, *21*, 192-219.

Howell, J. F., & Games, P. A. (1974). The effects of variance heterogeneity on simultaneous multiple-comparison procedures with equal sample size. *British Journal of Mathematical and Statistical Psychology*, *27*, 72-81.

Huber, P. J. (1981). *Robust statistics*. NY: Wiley.

Hyman, R. (1995). How to critique a published article. *Psychological Bulletin*, *118*, 178-182.

James, G. S. (1959). The Behrens-Fisher distribution and weighted means. *Journal of the Royal Statistical Society*, *21*, 73-80.

Jeffreys, H. (1937). On the relation between direct and inverse methods in statistics. *Royal Society of London Proceedings, Series A, 160*, 325-348.

Jeffreys, H. (1940). Note on the Behrens-Fisher formula. *Annals of Eugenics*, *10*, 48-51.

Johnson, R. A., & Weerahandi, S. (1988). A Bayesian solution to the multivariate Behrens-Fisher problem. *Journal of the American Statistical Association*, *83*, 145-149.

Keith, R. A., Granger, C. V., Hamilton, B. B., & Sherwin, F. S. (1987). *The Functional Independence Measure*: A new tool for rehabilitation. In M. G. Eisenberg & R. C. Grzesiak (Eds.), *Advances in clinical rehabilitation* (Vol. 1). NY: Springer, p. 6-18.

Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of some non-parametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, *30*, 213-221.

Kohr, R. L. (1970). A comparison of procedures for testing $\mu_1 = \mu_2$ with unequal n's and variances. Unpublished doctoral dissertation, The Pennsylvania State University.

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogeneous variances. *The Journal of Experimental Education*, *43*, 61-69.

Koschat, M. A., & Weerahandi, S. (1992). Chow-type tests under heteroscedasticity. *Journal of Business and Economic Statistics*, *10*, 221-228.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583-621.

Lepage, Y. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika, 58*, 213-217.

Lepage, Y. (1975). Asymptotically optimum rank tests for contiguous location and scale alternatives. *Communications in Statistics*, *4*, 671-687.

Levene, H. (1960). Robust tests for equality of variance. In I. Olkin (Ed.) *Contributions to probability and statistics*. Palo Alto, CA: Stanford University Press, p. 278-292.

Mann, H. B., & Whitney, d. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50-60.

Maxwell, S. E., & Cole, D. A. (1995). Tips for writing (and reading) methodological articles. *Psychological Bulletin*, *118*, 193-198.

McCullough, R. S., Gurland, J., & Rosenberg, L. (1960). Small sample behaviour of certain tests of the hypothesis of equal means under variance heterogeneity. *Bimoetrika*, *47*, 345-353.

McSweeney, M., & Penfield, D. (1969). The normal scores test for the c-sample problem. *The British Journal of Mathematical and Statistical Psychology*, *20*, 187-204.

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Nanna, M., & Sawilowsky, S. (1998). Analysis of Likert scale data in disability and medical rehabilitation evaluation. *Psychological Methods*, *3*, 55-67.

Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests*. London: Unwin Hyman Ltd.

Neyman, J., & Pearson, E. S. (1931). On the problem of k samples. *Bulletin international de l'Académie Polonaise des Sciences et des lettres (Cracovié), Sciences mathématiques, Série A,* 460-481.

O'Brien, P. C. (1988). Comparing two samples: extension of the t, rank-sum, and log-rank tests. *Journal of the American Statistical Association*, *83*, 52-61.

Oshima, T. C., & Algina, J. (1992). Type I error rates for James's second-order test and Wilcox's H_m test under heteroscedasticity and non-normality. *British Journal of Mathematical and Statistical Psychology*, *45*, 255-263.

Pagurova, V. I. (1968). On comparison of mean values based on two normal samples. *Teoriya Veroyatnostei i Ee Primeneniya*, *13*, 561-569.

Podgor, M.. J., & Gastwirth, J. L. (1994). On non-parametric and generalized tests for the two-sample problem with location and scale change alternatives. *Statistics in Medicine*, *14*, 747-758.

Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, *59*, 665-680.

Ray, W. D., & Pitman, A. E. N. T. (1961). An exact distribution of the Fisher-Behrens-Welch statistic for testing the difference between the means of two normal populations with unknown variances. *Journal of the Royal Statistical Society*, *23*, 377-384.

SAS (1990). *SAS/STAT user's guide, Vol. 1.* (4th ed.) Cary, NC: SAS Institute.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*, 309-316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110-114.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, *60*, 91-126.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111*, 353-360.

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples t test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology*, *60*, 240-243.

Sawilowsky, S. S., Baerg, P., Boza, L. A. D., Kallmannsohn, M., Spencer, B., & Vollhardt, L. T. (April, 1991). Power analysis of the Brownie-Boos-Oliver t test for expected increases in treatment variability. Annual meeting of the American Educational Research Association, SIG/Educational Statisticians. Chicago, IL.

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

Schrader, R. M., & Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, *67*, 93-101.

SPSS (1993). *SPSS for Windows: Base system user's guide release 6.0.* Chicago: SPSS.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, *5*, 1055-1098.

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1-25.

SYSTAT (1990). *SYSTAT: The system for statistics*: Evanston, IL: SYSTAT.

Tan, W. Y., & Tabatabai, M. A. (1985). Some robust ANOVA procedures under heteroscedasticity and nonnormality. *Communications in Statistics*, *14*, 1007-1026.

Tiku, M. L., & Singh, M. (1981). Robust test for means when population variances are unequal. *Communications in Statistics*, 10, 2057-2071.

Tomarkin, A. J., & Serlin, R. c. (1986). Comparison of ANOVA under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90-99.

Troendle, J. F., Blair, R. C., Rumsey, D., & Moke, P. (1997). Parametric and non-parametric tests for the overall comparison of several treatments to a control when treatment is expected to increase variability. *Statistics in Medicine*, *16*, 2729-2739.

Wald, A. (1955). Testing for differences between the means of two normal populations with unknown standard deviations. *Selected papers in statistics and probability*. NY: McGraw-Hill.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-62.

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

Welch, B. L. (1949a). Further notes on Mrs. Aspin's tables. *Biometrika*, *36*, 243-246.

Welch, B. L. (1949b). Appendix to A. A. Aspin's tables. *Biometrika*, *36*, 293-296.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal*, *32*, 771-780.

Wiles, A. (January 15, 1996). J. Lynch (Ed.) Interview: Fermat's last theorem. BBC Horizon.

Yates, F. (1939). An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters. *Proceedings of the Cambridge Philosophical Society*, *35*, 579-591.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*, 165-170.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology*, *51*, 139-150.