Far East Journal of Theoretical Statistics
Volume 39, Number 1, 2012, Pages 31-52

Published by Pushpa Publishing House, Allahabad, INDIA

ON WELCH'S AND ASPIN'S SERIES SOLUTION OF THE BEHRENS-FISHER PROBLEM

Martin Bachmaier

Technische Universität München
Agricultural Systems Engineering
Am Staudengarten 2, D-85 354 Freising
Germany
e-mail: bachmai@wzw.tum.de

Abstract

Welch [22] stated the Behrens-Fisher problem as a partial differential equation of infinite order and described how to obtain an exact solution of it by a series approach in reciprocal numbers of degrees of freedom. This solution gives the limit of the critical region of the Behrens-Fisher test variable as a function that only depends on the empirical variance ratio. However, Linnik [11] showed that such a function cannot be continuous, and up to now, it has not yet been commented upon that this contradicts Welch's approach, whose solution is postulated to be infinitely often differentiable. This paper tries to dissolve this contradiction on the basis of the Welch-Aspin test, which uses the expansion of Welch's series approach up to the fourth order. It becomes plausible that the convergence radius of Welch's series is zero, so Welch's approach does not provide an exact solution, and this is conform with Linnik's non-existence theorem. The investigation of the error probability of the first kind shows the accuracy of the Welch-Aspin test, but also indicates that developing too high orders could deteriorate the results.

2010 Mathematics Subject Classification: 62F03.

Keywords and phrases: Welch-Aspin test, Behrens-Fisher problem, heterogeneous variances, heteroscedasticity, hypothesis testing.

Received October 18, 2011

1. Introduction

To obtain a similar test of equal means, $H_0 : \mu_1 = \mu_2$, of normally distributed populations, which need not be homoscedastic, Behrens [5] and Fisher [6] proposed to find the probability distribution of

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1}$$

under H_0. Fisher approximated it by ignoring the random variation of the proportion $(s_1^2/n_1)/(s_1^2/n_1 + s_2^2/n_2)$, so the error probabilities of the first kind did not come close to α. The first good approximation came from Welch [21], who worked, like Fisher, at the University College London. The so-called "approximate *t*-solution" of Welch [23], which is most used in statistical packages, goes back to it. Solutions of Bartlett (in Neyman [13, p. 138]) and Scheffé [18, 19], which lead to exactly similar tests are not related to *t* in (1). The main disadvantage of these tests is that they are randomized throughout the sample space; their test result changes when permuting the sample. This and a loss of power of these tests make them undesirable.

Based on the classical and sufficient statistics $\bar{x}_1$, $\bar{x}_2$, s_1^2 and s_2^2, Welch [22] stated the function h for the limit between acceptance and rejection region of the mean difference $\bar{d} = \bar{x}_1 - \bar{x}_2$ as a solution of a partial differential equation of infinite order, claiming that "this, in a very condensed form, is the solution of our problem" (after equation (11)). He also described how to obtain it by a series approach in reciprocal numbers of degrees of freedom. This solution is proportional to the standard deviation of the mean difference, $s_{\bar{d}} = (s_1^2/n_1 + s_2^2/n_2)^{1/2}$, so $h^* := h/s_{\bar{d}}$ gives the boundary of the acceptance region of the Behrens-Fisher test variable *t* in (1). It shall be shown in the present article that h^* can be written as a function that only depends on the empirical variance ratio, and this gives exactly the desired shape of a scale-invariant solution (cf. e.g. Lehmann [10, Section 6.6]). Welch [22] clearly described how to obtain all orders of his series solution, and the Behrens-Fisher problem seemed to be mathematically solved, provided that the sample numbers and thus the degrees of freedom $\nu_i = n_i - 1$ are not too small, so that the series converges. For small ν_i Welch suggested to write his differential equation as an integral equation.

However, without reference to Welch [22], Linnik ([11, Theorem 8.3.1[1]]) showed that such a function cannot be continuous, which contradicts Welch's approach because his solving function h is postulated to be infinitely often differentiable. The fact that Linnik's claim about a non-existing continuous function of this kind is related to $\xi = (\bar{x}_1 - \bar{x}_2)/s_2$ and that Linnik's theorem is based on some weak conditions (Lipschitz continuity and finite derivative in certain intervals) is not essential since ξ and t can be transformed each other. Linnik's assumption $n_2 \geq 4$, or, since the samples can be exchanged, $n_1 \geq 4$, allows the existence of a continuous function if n_1 or n_2 is less than four, however, Welch's [22] approach provides a series solution in reciprocal numbers of degrees of freedom, and if this series converged for small $\nu_i = n_i - 1$, it would also converge for large ν_i and thus provide the exact solution. Therefore, this assumption would not clarify the contradiction between Welch and Linnik.

---

[1] To claim the non-existence of such a function, Pfanzagl [15] referred to Salaevskii's [17] theorem, which is also cited in Linnik [11, Theorem 8.2.1]. However, this theorem refers to the impossibility to find a test variable that is a continuous function of $t = (\bar{x}_1 - \bar{x}_2)/s_{\bar{d}}$ and the ratio s_1/s_2, so that any critical region of the form $[C, \infty)$ (for a one-sided test) gives a similar test. Thus, the Salaevskii theorem requires that the continuous function that is to be found must be the same for any significance level α. Welch [22], however, proposed the concept of how to obtain a continuous and even infinitely often differentiable function h for the boundary of the rejection region of $\bar{d} = \bar{x}_1 - \bar{x}_2$ in dependence of α. Salaevskii's theorem only shows that, contrary to the boundary of the rejection region of $\bar{d}$ in Student's test, Welch's h cannot be factorized into a term independent of α and a quantile value that only depends on α. But this was never intended by Welch and is not necessary to obtain an unbiased test. It is not Salaevskii's [17] theorem, which is cited by Linnik [11, Theorem 8.2.1], but Linnik's [11, Theorem 8.3.1] which refutes the existence of an exact solution in the spirit of Welch [22].

The contradiction between Welch [22] and Linnik [11] has not yet been commented upon. Instead, the Welch-Aspin test, which originates from Welch [22], is only referred to as an approximation of the Behrens-Fisher problem (e.g. Pfanzagl [15]). The tables of Mehta and Srinivasan [12], where Welch's [22] series approach is compared with Banerjee's [4], Fisher's [6], Pagurova's [14] and Wald's [20] approximations, show that for small samples Welch's approach and Mehta's and Srinivasan's improvement of Pagurova's [14] test perform best with respect to the α-error probability. Welch's [22] series is expanded up to the second order, but the Welch-Aspin test is based on Aspin's [1] development up to the fourth order, and its remarkable precision is mentioned by several authors, e.g. by Lee and Gurland [9]. Of course, all these finite series are only approximations, but this does not give us the right to maintain the non-existence of an exact solution and to simultaneously disregard that Welch [22] gave the recipe how to obtain all orders and thus an exact series solution. This discrepancy shall be clarified in the following.

2. The Question of Convergence of Welch's and Aspin's Series Solution

We shall see that Welch's series solution writes as a function that only depends on the empirical variance ratio. If Welch's series uniformly converged, then the limiting function would, as all partial sums, also be continuous, and this would contradict Linnik's [11] Theorem 8.3.1. Therefore, the contradiction between the non-existence of an exact similar solution of the Behrens-Fisher problem and Welch's exact series approach can only be resolved if the uniform convergence of Welch's series fails. Since Linnik's [11] non-existence of a solution is not confined to small-sample cases, this divergence must necessarily even hold if the sample numbers n_i are arbitrarily high. The question arises if this is believable.

Welch [22] described his method in a very condensed form. Aspin [1] did this in a similar way, but with two exchanged differential operators. A more extensive description is given in Bachmaier [2, Chapter 10]. It follows

Aspin [1], but the function to be developed is not related to the mean difference $\bar{x}_1 - \bar{x}_2$, but directly to the Behrens-Fisher test variable *t* in (1), and derivatives with respect to σ_i^2 are replaced by derivatives with respect to $\gamma_i := \sigma_i^2/n_i$. To discuss whether or not Welch's series approach converges, it is helpful to present the result of Aspin's development up to the fourth order. This gives us the possibility to correct a printing mistake[2] in Aspin's [1] paper and to show that Welch's method exclusively provides terms that only depend on the variance ratio $(s_1^2/n_1)/(s_2^2/n_2)$.

2.1. Aspin's development up to the fourth order

Based on the abbreviation

$$V_{lk} := \frac{\dfrac{g_1^l}{\nu_1^k} + \dfrac{g_2^l}{\nu_2^k}}{(g_1 + g_2)^l}, \text{ where } g_i = \frac{s_i^2}{n_i} \tag{2}$$

and the normal quantile $u := \Phi^{-1}(1 - \alpha)$ for a one-sided test, Aspin [1] presented the series development of the boundary *h* that must be exceeded by $\bar{x}_1 - \bar{x}_2$ to reject $H_0 : \mu_1 = \mu_2$ in favor of $H_1 : \mu_1 > \mu_2$. She had calculated it up to the fourth order. It follows the corresponding expansion of the boundary $h^* = h/s_{\bar{d}}$, which is related to the Behrens-Fisher test variable *t*:

$$h_{[4]}^*(\vec{g}) = h_0^*(\vec{g}) + h_1^*(\vec{g}) + h_2^*(\vec{g}) + h_3^*(\vec{g}) + h_4^*(\vec{g}), \tag{3}$$

where $\vec{g} = (g_1, g_2)$ and $h_k^*(\vec{g})$ contains the following terms in $\nu_1^{-k_1}\nu_2^{-k_2}$

---

[2] The summand of the fifth last line on page 90, which corresponds here to the fifth last line of $h_4^*(\vec{g})$, results in $-\frac{1}{64}(\cdots)V_{22}V_{21}^2$ instead of $-\frac{1}{64}(\cdots)V_{42}V_{21}^2$. A further printing mistake concerns the method description. At page 90, equation (11), every exponent of σ_i must be twice as big as the corresponding exponent of ∂_i, thus the expression $\sigma_i^6\partial_i^2$ must be corrected to $\sigma_i^6\partial_i^3$.

with $k_1 + k_2 = k$, which we call terms in ν^{-k}:

$$h_0^*(\vec{g}) = u, \tag{4}$$

$$h_1^*(\vec{g}) = u \cdot \frac{1}{4}(1 + u^2)V_{21}, \tag{5}$$

$$\begin{aligned} h_2^*(\vec{g}) &= u \\ &\cdot \left[-\frac{1}{2}(1 + u^2)V_{22} + \frac{1}{3}(3 + 5u^2 + u^4)V_{32} - \frac{1}{32}(15 + 32u^2 + 9u^4)V_{21}^2 \right], \end{aligned} \tag{6}$$

$$\begin{aligned} h_3^*(\vec{g}) = u \cdot \Big[&(1 + u^2)V_{23} - 2(3 + 5u^2 + u^4)V_{33} \\ &+ \frac{1}{8}(15 + 32u^2 + 9u^4)V_{22}V_{21} \\ &+ \frac{1}{8}(75 + 173u^2 + 63u^4 + 5u^6)V_{43} \\ &- \frac{1}{12}(105 + 298u^2 + 140u^4 + 15u^6)V_{32}V_{21} \\ &+ \frac{1}{384}(945 + 3169u^2 + 1811u^4 + 243u^6)V_{21}^3 \Big], \end{aligned} \tag{7}$$

$$\begin{aligned} h_4^*(\vec{g}) = u \cdot \Big[&- 2(1 + u^2)V_{24} + \frac{28}{3}(3 + 5u^2 + u^4)V_{34} \\ &- \frac{1}{4}(15 + 32u^2 + 9u^4)\left(V_{23}V_{21} + \frac{1}{2}V_{22}^2\right) \\ &- \frac{2}{3}(75 + 173u^2 + 63u^4 + 5u^6)V_{44} \\ &+ \frac{1}{2}(105 + 298u^2 + 140u^4 + 15u^6)\left(\frac{1}{3}V_{22}V_{32} + V_{21}V_{33}\right) \\ &+ \frac{1}{4}(15 + 33u^2 + 11u^4 + u^6)V_{44} \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{5}(735 + 2170u^2 + 1126u^4 + 168u^6 + 7u^8)V_{54}\\
&-\frac{1}{64}(945 + 3169u^2 + 1811u^4 + 243u^6)V_{22}V_{21}^2\\
&-\frac{1}{18}(945 + 3354u^2 + 2166u^4 + 425u^6 + 25u^8)V_{32}^2\\
&-\frac{1}{32}(4725 + 16586u^2 + 10514u^4 + 1974u^6 + 105u^8)V_{21}V_{43}\\
&+\frac{1}{96}(10395 + 42429u^2 + 31938u^4 + 7335u^6 + 495u^8)V_{32}V_{21}^2\\
&-\frac{1}{6144}(135135 + 626144u^2 + 542026u^4\\
&+145320u^6 + 11583u^8)V_{21}^4\Bigg]. \qquad (8)
\end{aligned}
$$

All these $h_k^*(\vec{g})$ only depend on V_{lk}, which can be written in a way that makes clear that they only depend on the variance ratio s_1^2/s_2^2 or $(s_1^2/n_1)/(s_2^2/n_2) = g_1/g_2$:

$$V_{lk} = \frac{\dfrac{g_1^l}{\nu_1^k} + \dfrac{g_2^l}{\nu_2^k}}{(g_1 + g_2)^l} = \frac{\left(\dfrac{g_1}{g_2}\right)^l \cdot \dfrac{1}{\nu_1^k} + \dfrac{1}{\nu_2^k}}{\left(\dfrac{g_1}{g_2} + 1\right)}, \text{ where } \frac{g_1}{g_2} = \frac{s_1^2}{s_2^2} \cdot \frac{n_2}{n_1}, \qquad (9)$$

and we may believe that this would not change when developing all other orders h_k^*, $k \in \mathbb{N}$, which are all continuous functions of $\vec{g}$.

Setting $\nu_2 := \infty$ or $s_2 := 0$ in (4)-(8), so that $g_2 = 0$, leads to the one-sample case, where the V_{lk} simplify to $1/\nu_1^k$. The resulting $(h_k^*(g_1))_{k\in\mathbb{N}}$ gives the series development of the *t*-quantile for any number of degrees of freedom (Fisher [7, p. 151]):

$$h_0^*(g_1) = u, \tag{10}$$

$$h_1^*(g_1) = \frac{u}{\nu_1} \cdot \frac{1}{4}(1 + u^2), \tag{11}$$

$$h_2^*(g_1) = \frac{u}{\nu_1^2} \cdot \left(\frac{1}{32} + \frac{1}{6}u^2 + \frac{5}{96}u^4\right), \tag{12}$$

$$h_3^*(g_1) = \frac{u}{\nu_1^3} \cdot \left(-\frac{5}{128} + \frac{17}{384}u^2 + \frac{19}{384}u^4 + \frac{1}{128}u^6\right), \tag{13}$$

$$h_4^*(g_1) = \frac{u}{\nu_1^4} \cdot \left(-\frac{21}{2048} - \frac{1}{48}u^2 + \frac{247}{15360}u^4 + \frac{97}{11520}u^6 + \frac{79}{92160}u^8\right). \tag{14}$$

The coefficients occurring in (4)-(8) are very large and they increase with increasing order *k*, but it seems that the summands in the h_k^* cancel each other nearly completely out when setting $g_2 := 0,$ so that we arrive at the one-sample case; for the coefficients in (10)-(14) are much smaller and they decrease with increasing order *k*, which is necessary for the series' convergence.

By contrast, to avoid a contradiction between Welch's [22] exact concept of a series approach and Linnik's Theorem 8.3.1 about its non-existence, there must exist a ratio g_1/g_2 for which Aspin's series in (4)-(8) does not even converge for arbitrarily large ν_i.

A series $h_0^* + h_1^* + h_2^* + h_3^* + \cdots$ that diverges for any arbitrarily large ν_i corresponds to a one-dimensional power series $a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots$ with a convergence radius of zero, and this requires the ratio of absolute coefficients, $|\, a_{k+1}/a_k \,|,$ to be unbounded, which applies for example if a_k increases with an order of *k*!. Figures of the rejection region boundary developed up to different orders shall help us to judge this question.

2.2. Figures for Aspin's series development

In the following, the degree of sample heteroscedasticity is not measured by variance ratios like s_1^2/s_2^2 or $(s_1^2/n_1)/(s_2^2/n_2)$, but, for reasons of symmetry, by the ratio (R) of difference (D) and sum (S) of the $g_i = s_i^2/n_i$:

$$\mathrm{RDS} = \frac{g_1 - g_2}{g_1 + g_2} = \frac{s_{\overline{X}_1}^2 - s_{\overline{X}_2}^2}{s_{\overline{X}_1}^2 + s_{\overline{X}_2}^2} = \frac{s_1^2/n_1 - s_2^2/n_2}{s_1^2/n_1 + s_2^2/n_2}. \tag{15}$$

The value 0 occurs when the empirical variances of the means, $g_i = s_i^2/n_i$, are equal. The limiting values -1 and $+1$ are obtained if all the variance of the mean difference comes from one of the two samples ($\mathrm{RDS} = -1$ if $\nu_1 = \infty$ or $s_1 = 0$; $\mathrm{RDS} = +1$ if $\nu_2 = \infty$ or $s_2 = 0$), so they correspond to the one-sample case, where Welch's [22] series development gives the *t*-quantile.

Figures 1-4 show the limits of the acceptance region of a one-sided test of $H_0 : \mu_1 \le \mu_2$ (or $H_0 : \mu_1 = \mu_2$) versus $H_1 : \mu_1 > \mu_2$ for $\alpha = 0.05$ as a function of RDS, when developed up to different order $K = 0, 1, 2, 3, 4$ according to Welch [22] and Aspin [1]. It is called $\widetilde{h}^*$, where $\widetilde{h}^*(\mathrm{RDS}) = h^*(\vec{g})$.

At the boundary of the RDS area, $\mathrm{RDS} = \pm 1$, all figures suggest that the series converges. As has been shown by Fisher [7, p. 151], it gives the series of the *t*-quantile, which converges to $t_{\nu_1;1-\alpha}$ if $\mathrm{RDS} = +1$ and to $t_{\nu_2;1-\alpha}$ if $\mathrm{RDS} = -1$. The corresponding power series for $\alpha = 0.05$ results in

$$\widetilde{h}^*(\pm 1) = h^*(g_i) = 1.644854 + \frac{1.523769}{\nu_i} + \frac{1.420203}{\nu_i^2} + \frac{0.983002}{\nu_i^3} + \frac{0.433876}{\nu_i^4} + \cdots. \tag{16}$$

The strong decrease of the latter coefficients also suggests that it converges even for $\nu_i = 1$.

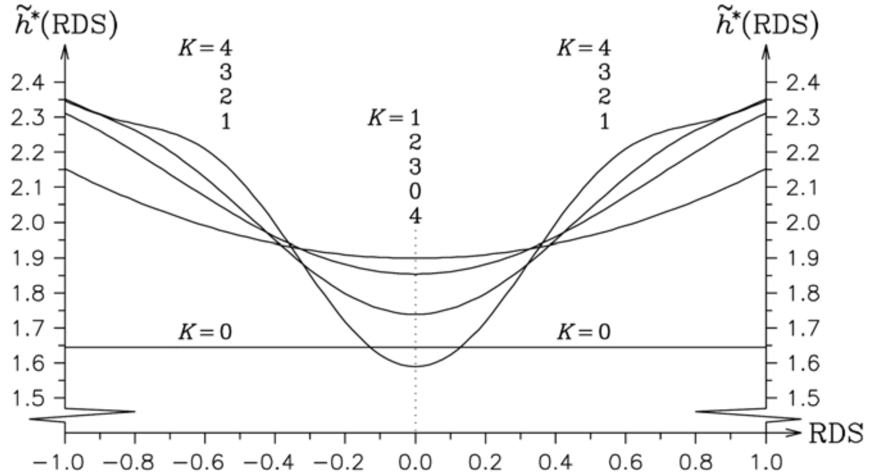


Figure 1. The limits of the rejection region for $\nu_1 = \nu_2 = 3$ and $\alpha = 0.05$ (one-sided test).

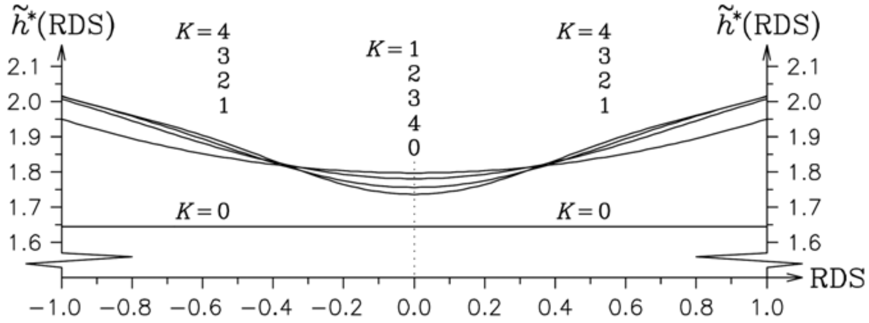


Figure 2. The limits of the rejection region for $\nu_1 = \nu_2 = 5$ and $\alpha = 0.05$ (one-sided test).

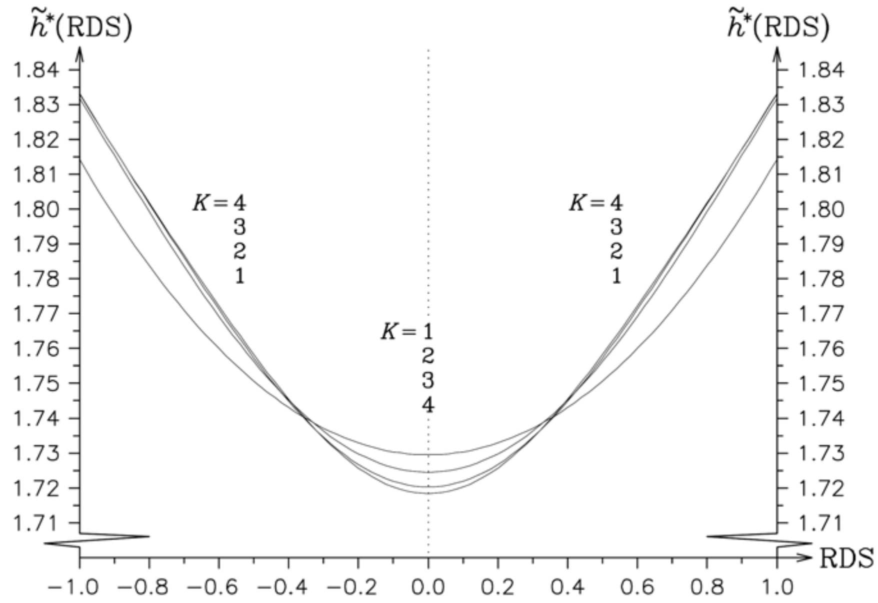


Figure 3. The limits of the rejection region for $\nu_1 = \nu_2 = 9$ and $\alpha = 0.05$ (one-sided test).

Clear doubts that the series converges arise at the mid of the RDS area, where RDS lies around zero. In the balanced case $\nu_1 = \nu_2 = 3$ (Figure 1) the absolute value of $\widetilde{h}_4^*(\text{RDS})$ proves at $\text{RDS} = 0$ greater than that of $\widetilde{h}_3^*(\text{RDS})$, so that a convergence is unbelievable. In the balanced cases with more degrees of freedom (Figures 2-3) and in the unbalanced case $\nu_1 = 4$, $\nu_2 = 8$ (Figure 4), $|\widetilde{h}_4^*(\text{RDS})|$ is at its maximum already less than $|\widetilde{h}_3^*(\text{RDS})|$, where the maximum lies at $\text{RDS} = 0$ in balanced cases and around $\text{RDS} = -0.1$ or -0.2 in the unbalanced case $\nu_1 = 4, \nu_2 = 8$ (Figure 4). Therefore, a convergence seems possible, so the power series needs further investigation. This shall be done for balanced cases $\nu_1 = \nu_2 =: \nu$, where the absolute summands, $|\widetilde{h}_k^*(\text{RDS})|$, of the series have their maximum at $\text{RDS} = 0$, i.e., for $g_1 = g_2 =: g$, for which the V_{lk} in (2) simplify to $V_{lk} = 2^{1-l}\nu^{-k}$. Hence, the following power series at $\text{RDS} = 0$ arises for

$\alpha = 0.05$:

$$\widetilde{h}^*(0) = h^*(g,\ g) = 1.644854 + \frac{0.761885}{\nu} - \frac{0.406834}{\nu^2} - \frac{3.094171}{\nu^3} - \frac{12.160788}{\nu^4} + \cdots. \quad (17)$$

Contrary to the power series in (16), which treats RDS = ±1, the coefficients a_k of the series in (17) do not decrease, but they increase exorbitantly. A convergence for small $\nu = \nu_1 = \nu_2$ seems impossible, but is it plausible that the series converges for no $\nu \in \mathbb{N}$ at all, so that the convergence radius in $1/\nu$ is indeed zero? For this, it is necessary that the ratios of absolute coefficients, $|a_k/a_{k-1}|$, is unbounded. The sequence of these ratios for $k = 1, 2, 3, 4$ results in (0.463, 0.534, 7.605, 3.930). There is a strong increase from the second to the third ratio, but the fourth ratio decreases, so that there might be doubts that this ratio sequence tends to infinity. However, considering that the absolute quadratic coefficient of the series, 0.407, is even smaller than the linear one, one should rather suggest that the absolute coefficients of even order turn out smaller throughout than those of odd order, and, to obtain a convergence radius of zero in $1/\nu$, it suffices that any subsequence of ratios, for example the sequence $|a_k/a_{k-2}|$ for odd or even k (or both), goes to infinity. And this might be supposed when viewing the power series in (17).

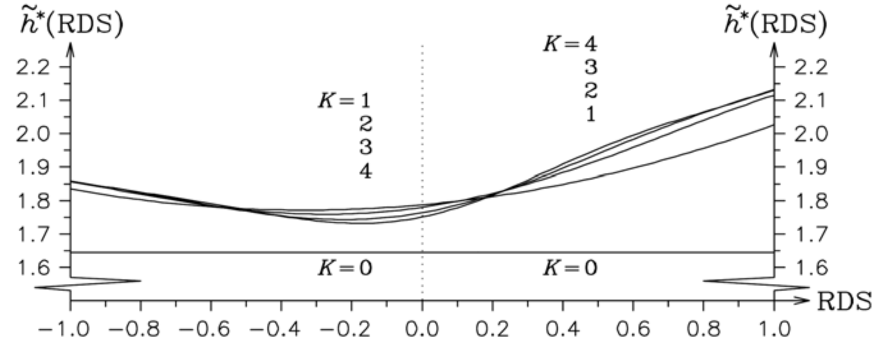


Figure 4. The limits of the rejection region for $\nu_1 = 4$, $\nu_2 = 8$ and $\alpha = 0.05$ (one-sided test).

All in all, the series development up to the fourth order makes it believable throughout that its convergence radius is zero. These means, when generalizing the test to unbalanced cases too, that the series development of Welch [22] and Aspin [1] converges for no pair (ν_1, ν_2), unless at least one of these numbers is infinity.

Welch's and Aspin's development is a series whose partial sums are all continuous. Only if the partial sums converge uniformly, the limit of the series is necessarily a continuous function too, which would contradict Linnik's Theorem 8.3.1. One could claim that the series does not converge uniformly, but pointwise, so that the limiting function need not be continuous, which would be compatible with Linnik's theorem. First of all, a pointwise, but not uniform convergence to a discontinuous function is not very plausible, but if it indeed applied, this limiting discontinuous function would not be proven to be the solution because Welch's differential approach postulates the function to be found to be infinitely often differentiable and thus continuous.

3. The Utility of the Welch-Aspin Test

Welch [22] mentioned that the partial differential equation on which the Welch-Aspin test is based should be replaced by an integral equation if the ν_i are small, and Aspin [1], who solved the differential equation up to the fourth order, produced her tables starting from $\nu_1 = \nu_2 = 6$. Mehta and Srinivasan [12] also warn against "using the asymptotic expansions for too small sample sizes" and investigated Welch's [22] test, which is developed up to the second order $(K = 2)$. In this section, which mainly relates to balanced cases, the Welch-Aspin test $(K = 4)$ is investigated even for $\nu_1 = \nu_2 = 1$, but Figure 5 shows that, at the 5% level, its application for very small ν_i is not meaningful, in particular, for $\nu_1 = \nu_2 = 1$ a *t*-value of zero would suffice to reject the null hypothesis of equal μ_i in favor of a one-sided alternative if RDS is close to zero, and a test versus the two-sided alternative with the corresponding $\alpha = 0.10$ would no longer be possible.

Therefore, for small ν_i the question as to how far the series should be expanded shall be investigated in the following. The limiting function $\widetilde{h}^*$ should be such that it enables the calculation of reasonably small *P*-values. Further, the error probability of the first kind should come as close as possible to α.

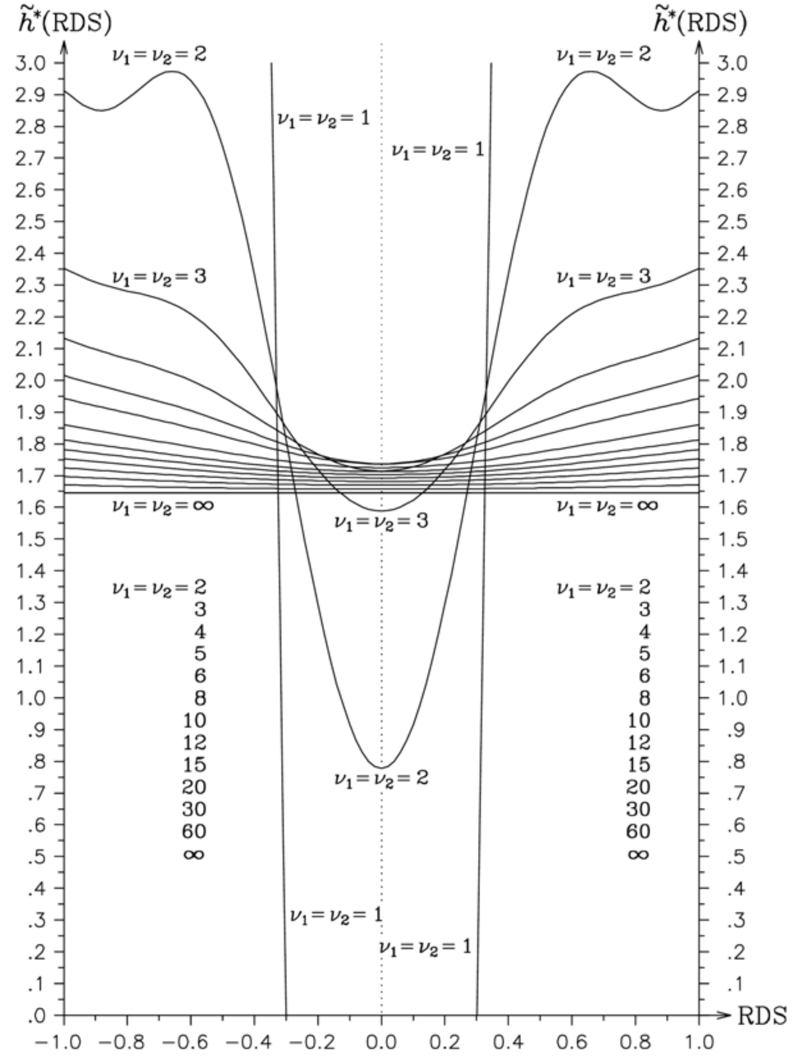


Figure 5. The limits of the rejection region for $K = 4$ and $\alpha = 0.05$ (one-sided test).

3.1. *P*-values

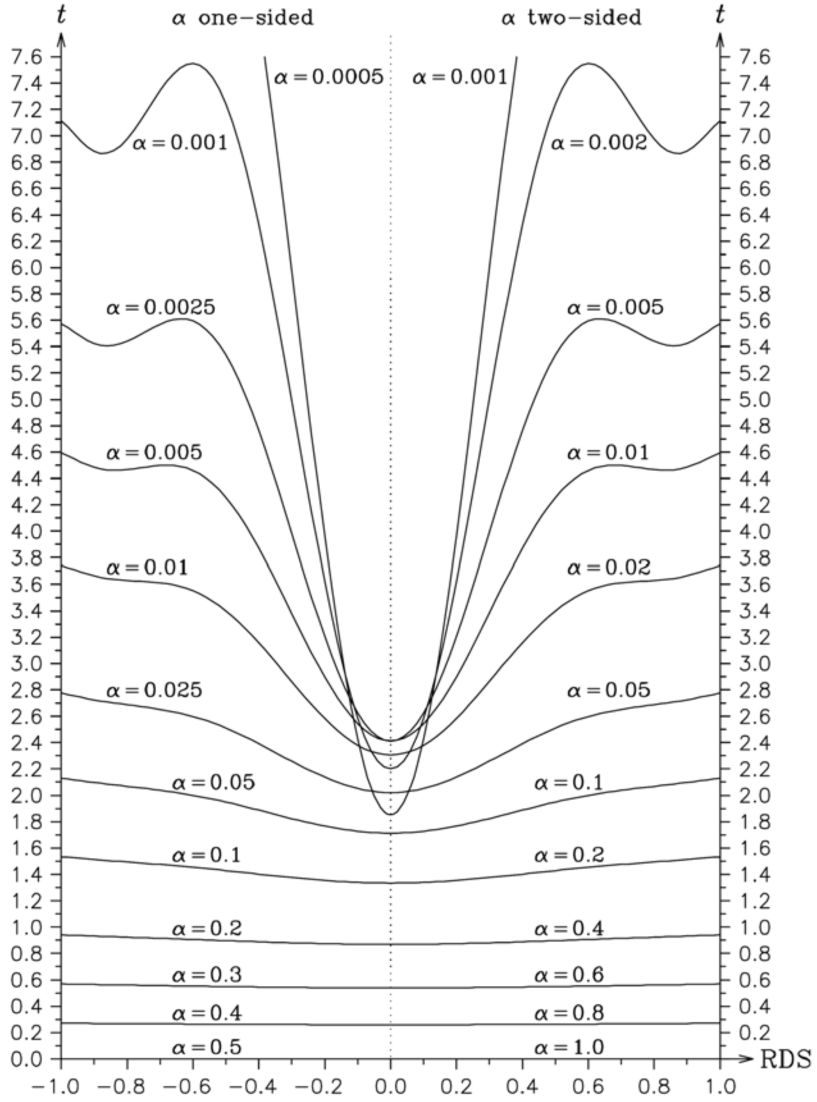


Figure 6. The limits of the rejection region for $\nu_1 = \nu_2 = 4$ and $K = 4$.

Figure 6 treats the case $\nu_1 = \nu_2 = 4$ when all orders $(\text{i.e., } K = 4)$ of the Welch-Aspin test are computed. It shows $\widetilde{h}^*$ for different α down to 0.001 (for one-sided testing). Although $\widetilde{h}^*$ should increase for decreasing α,

$\tilde{h}^*(0)$ begins to decrease when α falls short of 0.003584, where $h^*(0)$ obtains its maximum value $t_{\max} = 2.427755949$. This means in praxis that one should not test at a level of $\alpha < 0.003584$ and that one-sided *P*-values smaller than 0.003584 cannot be obtained. If the *t* value exceeds $t_{\max}$, then it would nevertheless be meaningful to claim that the *P*-value is less than 0.003584. To ensure that *P*-values down to 0.001 can be reached, it turned out that $\tilde{h}^*$ should only be computed up to the third order $(K = 3)$. This equally holds for $\nu_1 = \nu_2 = 3$. For $\nu_1 = \nu_2 = 2$ and $\nu_1 = \nu_2 = 1$ the series should only be computed up to the second order $(K = 2)$. *P*-values down to 0.001 by computing all orders of the Welch-Aspin test $(K = 4)$ can already be obtained if $\nu_1 = \nu_2 = 5$ or greater.

3.2. Error probabilities of the first kind

The probability for the α-error of the Welch-Aspin test has already been much investigated (e.g. Lee and Gurland [9]). In the focus of this section is its dependence of the order *K* up to which the series is developed. In the following, the probability for the α-error is called $\tilde{\alpha}$; it is obtained by numerically computing the following double integral:

$$\tilde{\alpha} = \int_0^\infty \int_0^\infty \left[1 - \Phi\left(\frac{h(\vec{g})}{\sqrt{\gamma_1 + \gamma_2}}\right)\right] f_{\nu_1}(g_1)dg_1 f_{\nu_2}(g_2)dg_2, \qquad (18)$$

where the notation $h(\vec{g}) = h^*(\vec{g}) \cdot \sqrt{g_1 + g_2}$ corresponds to Aspin [1], whose function *h* relates to the test variable $\bar{x}_1 - \bar{x}_2$, and

$$f_{\nu_i}(g_i) = \frac{1}{\Gamma(\nu_i/2)}\left(\frac{\nu_i}{2\gamma_i}\right)^{\nu_i/2} g_i^{\nu_i/2-1} \exp\left(-\frac{\nu_i g_i}{2\gamma_i}\right) \qquad (19)$$

denote the densities of the independent variance estimators, $g_i = s_i^2/n_i$, of the means $\bar{x}_i$. These g_i follow a $(\gamma_i/\nu_i) \cdot \chi^2_{\nu_i}$ distribution, where $\gamma_i = \sigma_i^2/n_i$.

The region of rejection for $a = 2$ is illustrated in dependence of the ratio

RDS of difference and sum of sample variances of the means. Analogously, $\tilde{\alpha}$ and power shall be described in dependence of such a ratio, which now refers to the population variances instead of the sample variances:

$$\mathrm{RaDS} = \frac{\gamma_1 - \gamma_2}{\gamma_1 + \gamma_2} = \frac{\sigma_{\overline{X}}^2 - \sigma_{\overline{Y}}^2}{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2} = \frac{\sigma_x^2/n_x - \sigma_y^2/n_y}{\sigma_x^2/n_x + \sigma_y^2/n_y}. \tag{20}$$

Hence, $\gamma_1 := \frac{1}{2}(1 + \mathrm{RaDS})$ and $\gamma_2 := \frac{1}{2}(1 - \mathrm{RaDS})$ can be set such that $\sqrt{\gamma_1 + \gamma_2} = 1$, which simplifies the computation of $\tilde{\alpha}$ in (18)-(19).

Figure 7 illustrates the error probability of the first kind for $\nu_1 = 4$ and $\nu_2 = 8$. It shows that the development up to the fourth order gives the best result and that the level $\alpha = 0.05$ is reached exactly for even six values of RaDS.

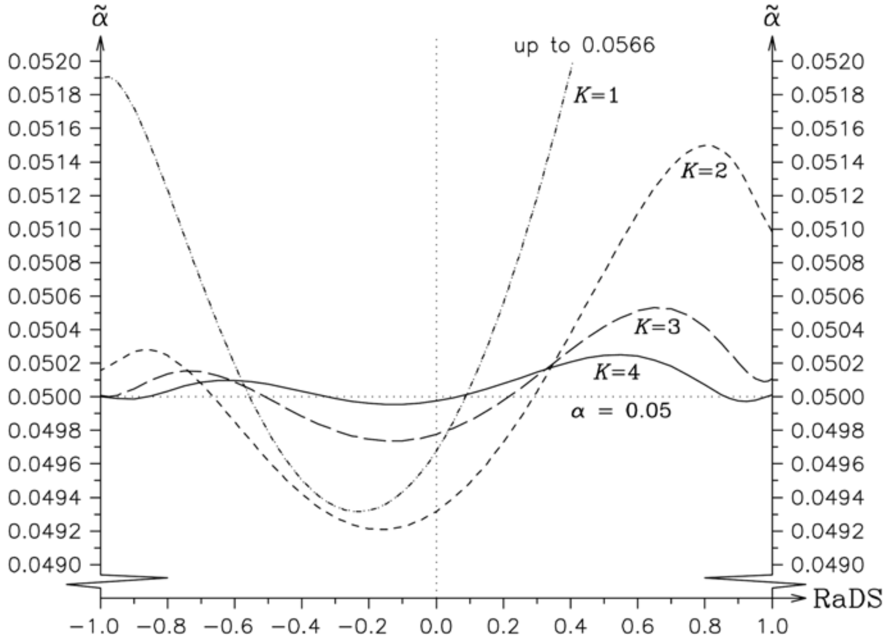
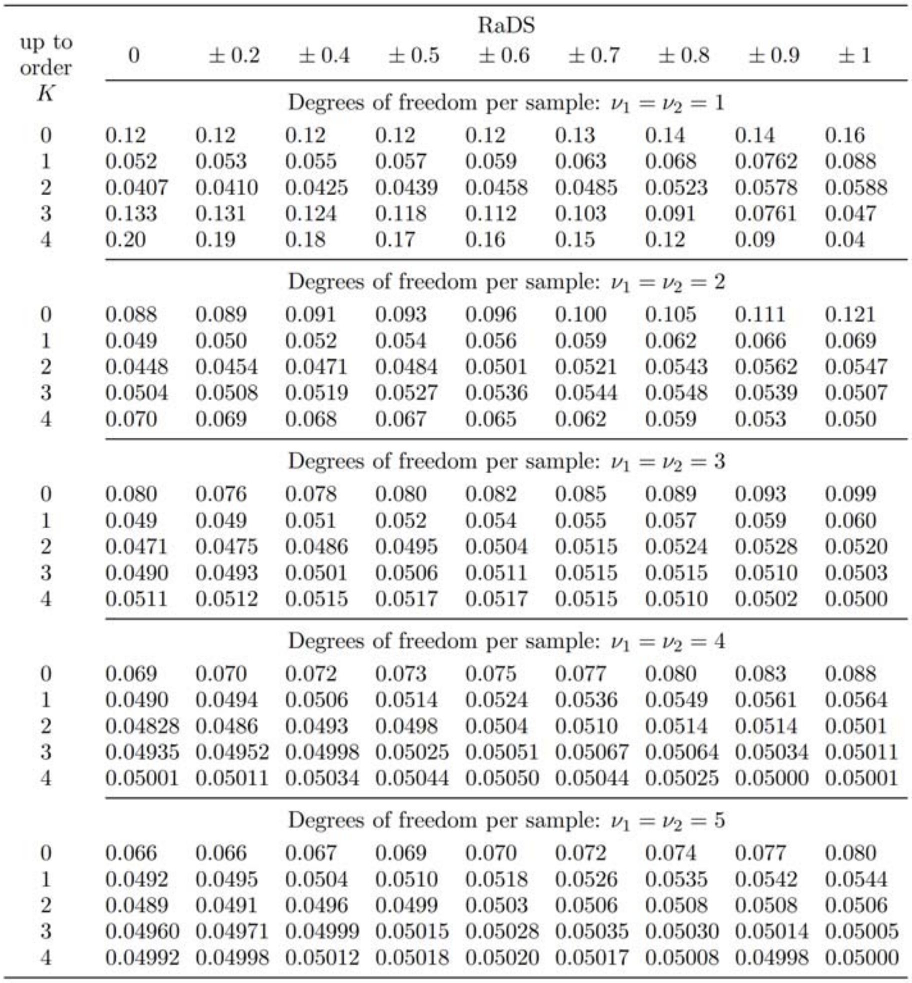


Figure 7. The error probability of the first kind, $\tilde{\alpha}$, for $\nu_1 = 4$, $\nu_2 = 8$ and $\alpha = 0.05$ (one-sided test).

The actual error probabilities of the first kind, $\tilde{\alpha}$, in Tables 1 and 2 are given such that, within the range of RaDS values (i.e., within each line), the maximal deviation from the nominal level $\alpha = 0.05$ has a two-digit precision. This also visualizes the accuracy of the Welch-Aspin test with respect to the maximal order of development.

Table 1 shows that Welch's [22] series approach is even applicable in the case of minimal samples, i.e., if $\nu_1 = \nu_2 = 1$. The best result is obtained when Welch's series is developed up to the second order $(K = 2)$, where $\tilde{\alpha}$ ranges between 0.04 and 0.06. For $\nu_1 = \nu_2 = 2$ and $\nu_1 = \nu_2 = 3$ the series should be developed up to the third order $(K = 3)$ to obtain the most accurate $\tilde{\alpha}$. The fourth order $(K = 4)$ should only be developed if $\nu_1 = \nu_2 = 4$ or greater. These results indicate that there also exists for greater ν_i an order K up to which the series development is optimal with respect to $\tilde{\alpha}$. It seems that there exist no $\nu_i < \infty$ for which the development of all orders would be optimal, and this would mean that the infinite series cannot be the exact solution, and Linnik's [11] Theorem 8.3.1 of the non-existence of a continuous solution would be confirmed again.

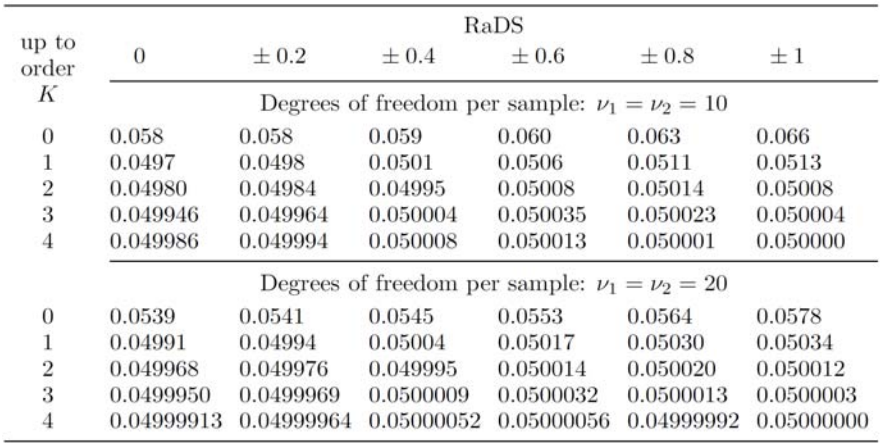
Table 1. Probability of α-error of the Welch-Aspin test at $\alpha = 0.05$



up to order K	RaDS								
	0	± 0.2	± 0.4	± 0.5	± 0.6	± 0.7	± 0.8	± 0.9	± 1
	Degrees of freedom per sample: $\nu_1 = \nu_2 = 1$								
0	0.12	0.12	0.12	0.12	0.12	0.13	0.14	0.14	0.16
1	0.052	0.053	0.055	0.057	0.059	0.063	0.068	0.0762	0.088
2	0.0407	0.0410	0.0425	0.0439	0.0458	0.0485	0.0523	0.0578	0.0588
3	0.133	0.131	0.124	0.118	0.112	0.103	0.091	0.0761	0.047
4	0.20	0.19	0.18	0.17	0.16	0.15	0.12	0.09	0.04
	Degrees of freedom per sample: $\nu_1 = \nu_2 = 2$								
0	0.088	0.089	0.091	0.093	0.096	0.100	0.105	0.111	0.121
1	0.049	0.050	0.052	0.054	0.056	0.059	0.062	0.066	0.069
2	0.0448	0.0454	0.0471	0.0484	0.0501	0.0521	0.0543	0.0562	0.0547
3	0.0504	0.0508	0.0519	0.0527	0.0536	0.0544	0.0548	0.0539	0.0507
4	0.070	0.069	0.068	0.067	0.065	0.062	0.059	0.053	0.050
	Degrees of freedom per sample: $\nu_1 = \nu_2 = 3$								
0	0.080	0.076	0.078	0.080	0.082	0.085	0.089	0.093	0.099
1	0.049	0.049	0.051	0.052	0.054	0.055	0.057	0.059	0.060
2	0.0471	0.0475	0.0486	0.0495	0.0504	0.0515	0.0524	0.0528	0.0520
3	0.0490	0.0493	0.0501	0.0506	0.0511	0.0515	0.0515	0.0510	0.0503
4	0.0511	0.0512	0.0515	0.0517	0.0517	0.0515	0.0510	0.0502	0.0500
	Degrees of freedom per sample: $\nu_1 = \nu_2 = 4$								
0	0.069	0.070	0.072	0.073	0.075	0.077	0.080	0.083	0.088
1	0.0490	0.0494	0.0506	0.0514	0.0524	0.0536	0.0549	0.0561	0.0564
2	0.04828	0.0486	0.0493	0.0498	0.0504	0.0510	0.0514	0.0514	0.0501
3	0.04935	0.04952	0.04998	0.05025	0.05051	0.05067	0.05064	0.05034	0.05011
4	0.05001	0.05011	0.05034	0.05044	0.05050	0.05044	0.05025	0.05000	0.05001
	Degrees of freedom per sample: $\nu_1 = \nu_2 = 5$								
0	0.066	0.066	0.067	0.069	0.070	0.072	0.074	0.077	0.080
1	0.0492	0.0495	0.0504	0.0510	0.0518	0.0526	0.0535	0.0542	0.0544
2	0.0489	0.0491	0.0496	0.0499	0.0503	0.0506	0.0508	0.0508	0.0506
3	0.04960	0.04971	0.04999	0.05015	0.05028	0.05035	0.05030	0.05014	0.05005
4	0.04992	0.04998	0.05012	0.05018	0.05020	0.05017	0.05008	0.04998	0.05000

Table 2 exhibits the exactness of the Welch-Aspin test for greater ν_i like $\nu_1 = \nu_2 = 10$ and $\nu_1 = \nu_2 = 20$, which could make one believes that the exact solution must also exist, but this nevertheless does not imply that the development of any additional order improves the accuracy of the test.

Table 2. Probability of α-error of the Welch-Aspin test at $\alpha = 0.05$



up to order K	RaDS 0	± 0.2	± 0.4	± 0.6	± 0.8	± 1
	Degrees of freedom per sample: $\nu_1 = \nu_2 = 10$					
0	0.058	0.058	0.059	0.060	0.063	0.066
1	0.0497	0.0498	0.0501	0.0506	0.0511	0.0513
2	0.04980	0.04984	0.04995	0.05008	0.05014	0.05008
3	0.049946	0.049964	0.050004	0.050035	0.050023	0.050004
4	0.049986	0.049994	0.050008	0.050013	0.050001	0.050000
	Degrees of freedom per sample: $\nu_1 = \nu_2 = 20$					
0	0.0539	0.0541	0.0545	0.0553	0.0564	0.0578
1	0.04991	0.04994	0.05004	0.05017	0.05030	0.05034
2	0.049968	0.049976	0.049995	0.050014	0.050020	0.050012
3	0.0499950	0.0499969	0.0500009	0.0500032	0.0500013	0.0500003
4	0.04999913	0.04999964	0.05000052	0.05000056	0.04999992	0.05000000

4. Summary of Results and Conclusions

Welch [22] stated the Behrens-Fisher problem as a partial differential equation, which he solved by a series approach. Although his series approach converges to the *t*-quantile for the special case of one sample, it does not seem to converge in the two-sample case when the variances of the sample means are rather equal, and it is believable throughout that the convergence fails for arbitrarily high ν_i. This non-convergence resolves the contradiction with Linnik [11], who states in Theorem 8.3.1 that an exact continuous solution for the rejection region of the Behrens-Fisher test cannot exist unless one sample size is less than four.

Nevertheless, for 60 years there has not been developed a better test than the Welch-Aspin test. It is the most accurate solution of the Behrens-Fisher problem. It performs especially well for high ν_i; for very small ν_i only the first two or three orders should be expanded, so that also small *P*-values can be computed.

Although Welch's [22] series approach does not converge, the good performance of its finite order development recommends Welch's method to other test problems. It has been used by James [8] for ANOVA under heterogeneous variances and by Bachmaier [2] for equivalence tests.

Currently, the present author also applies this method to the role-reversal of the latter test, which is the two-sided test for relevant difference [3].

References

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

[2] M. Bachmaier, Testing for equivalence of means under heteroskedasticity by an approximate solution of a partial differential equation of infinite order, Scand. J. Stat. 38 (2011), 147-168.

[3] M. Bachmaier, Extending Welch's differential approach of the Behrens-Fisher problem to testing for relevant difference, Far East J. Theo. Stat. 39(1) (2012), 1-29

[4] S. K. Banerjee, Approximate confidence interval for linear functions of means of *k* populations when the population variances are not equal, Sankhyā 35 (1960), 88-96.

[5] W. U. Behrens, Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen, Landwirtschaftliche Jahrbücher 68 (1929), 807-837.

[6] R. A. Fisher, The Fiducial Argument in Statistical Inference, Ann. Eugenics 6 (1935), 391-398.

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

[8] G. S. James, The comparison of several groups of observations when the ratios of the population variances are unknown, Biometrika 38 (1951), 324-329.

[9] A. F. S. Lee and J. Gurland, Size and power of tests for equality of means of two normal populations with unequal variances, J. Amer. Statist. Assoc. 70 (1975), 933-941.

[10] E. L. Lehmann, Testing Statistical Hypotheses, John Wiley & Sons, New York, 1986.

[11] Yu. V. Linnik, Statistical problems with nuisance parameters, Translations of mathematical monographs; American Mathematical Society, Providence, Rhode Island, 1968.

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

[13] J. Neyman, Fiducial argument and the theory of confidence intervals, Biometrika 32 (1941), 128-150.

[14] V. I. Pagurova, Tests of comparison of mean values based on two normal samples, Reports on Computational Mathematics, No. 5, Computing Center of the Akademy of Sciences of the U.S.S.R. Moscow, 1968 (in Russian).

[15] J. Pfanzagl, On the Behrens-Fisher problem, Biometrika 61 (1974), 39-47.

[16] I. L. Romanovskaja, On the Wald test with unequal samples, Proc. Steklov Inst. Math. 79 (1966), 69-76 (in Russian).

[17] O. V. Šalaevskiĭ, On the non-existence of regularly varying tests for the Behrens-Fisher problem, Sovjet Mathematics Dokl. 4 (1963), 1043-1045.

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

[19] H. Scheffé, A note on the Behrens-Fisher problem, Ann. Math. Statist. 15 (1944), 430-432.

[20] A. Wald, Testing the difference between the means of two normal populations with unknown standard deviations, Selected Papers in Statistics and Probability, McGraw-Hill, New York, 1955.

[21] B. L. Welch, The significance of the difference between two means when the population variances are unequal, Biometrika 29 (1938), 350-362.

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

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