Colloquium Biometricum, 37
2007, str. 125–135

POWER COMPARISON OF FOUR TESTS IN BEHRENS-FISHER PROBLEM

Joanna Tarasińska

Department of Applied Mathematics
Agricultural University of Lublin
Akademicka 13, 20-950 Lublin
e-mail: joanna.tarasinska@.ar.lublin.pl

Summary

The problem of testing the equality of means from two independent normal populations with unknown different variances is known as Behrens-Fisher problem. In the paper the powers of four tests are compared: Student's *t* test, Welch-Satterthwaite test, Saxena-Srivastava test and the test combined of *F* test for equality of variances and Student's *t* or Welch- Satterthwaite according to its result.

Key words and phrases: Behrens-Fisher problem, testing hypothesis, testing equality of means, power comparison, unequal variances

Classification AMS 2000: 62F03

1. Introduction

Let $X_1,\ldots,X_{n_1}$ and $Y_1,\ldots,Y_{n_2}$ be two independent samples, where

$$X_i \sim N\left(\mu_X,\sigma_X^2\right) \quad iid \quad i=1,\ldots,n_1,$$

$$Y_j \sim N\left(\mu_Y,\sigma_Y^2\right) \quad iid \quad j=1,\ldots,n_2$$

The difference of sample means $\overline{X}-\overline{Y}$ has got normal distribution with expected value $\mu_X-\mu_Y$ and variance $\frac{\sigma_X^2}{n_1}+\frac{\sigma_Y^2}{n_2}$. When variances σ_X^2 and σ_Y^2 are known then testing hypothesis $H_0:\mu_X=\mu_Y$ is based on the statistic

$$t=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_1}+\frac{\sigma_Y^2}{n_2}}}\ , \tag{1.1}$$

which, under H_0, has got standard normal distribution. When σ_X^2 and σ_Y^2 are unknown but the same, the pooled unbiased estimate of common σ^2 i.e.

$$S^2=\frac{(n_1-1)S_X^2+(n_2-1)S_Y^2}{n_1+n_2-2}=\frac{\sum_{i=1}^{n_1}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{n_2}\left(Y_i-\overline{Y}\right)^2}{n_1+n_2-2}$$

is put into (1.1) and the uniformly most powerful Student's *t* test based on

$$t=\frac{\overline{X}-\overline{Y}}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \tag{1.2}$$

is obtained. Under H_0 the statistic (1.2) has got Student's *t* distribution with n_1+n_2-2 degrees of freedom. It is known that for unequal sample sizes this test is not robust under violation of the assumption $\sigma_X^2=\sigma_Y^2$. The type I error can differ considerably from assumed significance level (Hsu, 1938).

The hypothesis $\sigma_X^2=\sigma_Y^2$ should be tested by means of *F* test. If $\sigma_X^2=\sigma_Y^2$ is rejected, the problem of testing the hypothesis H_0 is known in literature as Behrens-Fisher problem . Putting estimates S_X^2 and S_Y^2 instead of σ_X^2 and σ_Y^2 into (1.1) and denoting $S_{\overline{X}}^2=\frac{S_X^2}{n_1}$ and $S_{\overline{Y}}^2=\frac{S_Y^2}{n_2}$ we have statistic

$$\tilde{t} = \frac{\overline{X} - \overline{Y}}{\sqrt{S_{\overline{X}}^2 + S_{\overline{Y}}^2}} \tag{1.3}$$

considered by Behrens (1929) and Fisher (1935). Its distribution depends on $\nu_1 = n_1 - 1,\ \nu_2 = n_2 - 1$ and $c = \frac{S_{\overline{X}}^2}{S_{\overline{X}}^2 + S_{\overline{Y}}^2}$ and is the mixture of two *t* distributions. Tables for critical values $t(\nu_1, \nu_2, c, \alpha)$ such that $P(\tilde{t} > t(\nu_1, \nu_2, c, \alpha)) = \alpha$ were given by Aspin (1949), (see also Elandt, 1964; Domański, 1990; Zieliński and Zieliński, 1990**).**

The distribution of (1.3) has been investigated by many statisticians for last seven decades (for references see for example Dudewicz and Ahmed, 1998; or Singh, Saxena and Srivastava, 2002). Several approximations were derived to avoid using special tables. Some of them can be found also in polish popular handbooks , among others Elandt (1964), Oktaba (1984), Wagner and Błażczak (1986), Krysicki et al. (1986), Domański (1990), Klonecki (1999), Chudzik et al.(2006). Bootstrap approach is given for example in Domański, Pruska and Wagner (1998). There are also papers considering multivariate Behrens-Fisher problem (for references see Krzyśko, 2000).

Among many approximated solutions to Behrens-Fisher problem, based on statistic (1.3), perhaps the most popular in practice is the one proposed by Welch (1937,1947) and Satterthwaite (1946), called in this paper after Sinhg et al. (2002) as Welch-Satterthwaite test. According to it the distribution of (1.3) under H_0 is approximated by Student's *t* distribution with ν degrees of freedom, where ν is rounded down to the nearest integer of following $\tilde{\nu}$:

$$\tilde{\nu} = \frac{\left(S_{\overline{X}}^2 + S_{\overline{Y}}^2\right)^2}{\frac{\left(S_{\overline{X}}^2\right)^2}{n_1 - 1} + \frac{\left(S_{\overline{Y}}^2\right)^2}{n_2 - 1}}.$$

So, the critical value of Welch-Satterthwaite test varies from one sample to another.

It is worth noticing that there exists generalization of this test for the case of several populations, i.e. approximated F test in one-way ANOVA (Wagner and Błażczak, 1986).

One of the more recent tests for Behrens-Fisher problem was proposed by Saxena and Srivastava (1986). It is not based on statistic (1.3) but on statistic

$$t_{SS} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_2} + \frac{S_Y^2}{n_1}}}, \tag{1.4}$$

which, under H_0, has got approximated $\frac{1}{\sqrt{\lambda}} t_{\hat{\nu}}$ distribution where $\lambda = \frac{\frac{S_X^2}{n_2} + \frac{S_Y^2}{n_1}}{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}$ and $\hat{\nu} = \frac{\left(\frac{S_X^2}{n_2} + \frac{S_Y^2}{n_1}\right)^2}{\frac{\left(\frac{S_X^2}{n_2}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_Y^2}{n_1}\right)^2}{n_2 - 1}}$, rounded down to the nearest integer, (Singh et al.(2002)).

The test statistic (1.4) was obtained by putting into (1.2) the jackknife estimator of σ^2 instead of pooled sample variance S^2. The critical value in this test also varies from one sample to another and hypothesis H_0 should be rejected against $H_1 : \mu_X > \mu_Y$ if $\sqrt{\lambda}\, t_{SS} > t_{1-\alpha,\hat{\nu}}$. Saxena and Srivastava proposed also another approximation of distribution for statistic (1.4) but we don't consider it here as it is computationally troublesome.

Let us notice that for equal sample sizes $n_1 = n_2$ Saxena-Srivastava test is identical to Welch-Satterthwaite test.

Singh, Saxena and Srivastava (2002) made comparison of Student's *t*-test, Saxena-Srivastava test, Welch-Satterthwaite test and Cochran-Cox test with regard to probability of type I errors and to the powers. They concluded that their test kept probability of type I error very stable and close to the nominal value of 0.05. The power of their test turned out to be better than Cochran-Cox test and comparable to Welch-Satterthwaite test.

In the present paper as the competitor to Student's t, Welch-Satterthwaite and Saxena-Srivastava tests we consider the intuitive test based on choice between Student's *t* or Welch-Satterthwaite test according to results of previously performed *F* test for equality of variances. So this "combined" test can be de-

scribed as follows: test $H_0^{(0)}: \sigma_1^2 = \sigma_2^2$ against $H_1^{(0)}: \sigma_1^2 \neq \sigma_2^2$ on significance level α. If $H_0^{(0)}$ is not rejected then for testing hypothesis $H_0: \mu_X = \mu_Y$ use Student's *t*-test. If $H_0^{(0)}$ is rejected then use Welch-Satterthwaite test.

The aim of this paper is especially to compare this "ombined" test to three others.

2. Empirical study of power and type I error rate

It is easy to see that powers of all four considered tests depend on n_1, n_2, $k = \frac{\sigma_X}{\sigma_Y}$ and $\frac{\mu_X - \mu_Y}{\sigma_Y} = \frac{\Delta}{\sigma_Y}$. In order to make comparisons of the considered tests, 5000 random samples of size n_1 and n_2 were generated from $N(\Delta, k^2)$ and $N(0,1)$, respectively, for $\Delta = 0,1,2,3,4,5$ and $k = 1,2,3,4$. The different combinations of n_1 and n_2 were $n_1 = 5$ and $n_2 = 5$, $n_1 = 5$ and $n_2 = 10$, $n_1 = 10$ and $n_2 = 5$, $n_1 = 10$ and $n_2 = 10$, $n_1 = 15$ and $n_2 = 5$, $n_1 = 5$ and $n_2 = 15$. The one-sided alternative was considered: $H_1: \mu_X > \mu_Y$. The estimate of power is taken as the relative frequency with which test statistics exceed their critical values on significance level $\alpha = 0.05$. For $\Delta = 0$ the same frequency is just the probability of type I error.

Table 1 presents probability of Type I Errors. In tables 2-7 there are powers of the tests. All values are rounded to the third decimal point. Following notations are used in tables: t – Student's *t* test, t^* – "combined" test, t_{WS} – Welch-Satterthwaite test, t_{SS} – Saxena-Srivastava test. If any row for specified above Δ's is omitted it means that in such a case power is 1.

Additionally in tables 2-7 the power of *F* test (the frequency of choice t_{WS} in "combined" test) is given.

Table 1. Probability of Type I Errors

	n_1=5 , n_2=5				n_1=10 , n_2=10			
	k=1	k=2	k=3	k=4	k=1	k=2	k=3	k=4
t	.051	.055	.061	.061	.048	.052	.053	.055
t*	.050	.052	.054	.053	.047	.051	.049	.049
t_{WS}	.044	.047	.048	.049	.046	.050	.049	.049
t_{SS}	.044	.047	.048	.049	.046	.050	.049	.049
	n_1=5 , n_2=10				n_1=10 , n_2=5			
t	.044	.086	.106	.116	.055	.026	.019	.016
t*	.045	.067	.059	.047	.055	.036	.037	.044
t_{WS}	.042	.044	.042	.041	.053	.053	.050	.051
t_{SS}	.046	.054	.051	.048	.059	.051	.048	.049
	n_1=5 , n_2=15				n_1=15 , n_2=5			
t	.050	.120	.155	.171	.047	.015	.008	.005
t*	.050	.081	.061	.052	.049	.029	.033	.042
t_{WS}	.046	.047	.045	.045	.052	.047	.046	.047
t_{SS}	.056	.063	.061	.058	.061	.049	.046	.046

Table 2. Powers for n_1=5, n_2=5

	Δ	t	t*	t_{WS}	t_{SS}	F
k=1	1	.424	.420	.401	.401	.053
	2	.895	.892	.877	.877	
	3	.996	.996	.996	.996	
	4	1	1	1	1	
k=2	1	.244	.232	.212	.212	.219
	2	.578	.559	.528	.528	
	3	.859	.842	.823	.823	
	4	.976	.963	.957	.957	
	5	.998	.996	.996	.996	
k=3	1	.175	.156	.142	.142	.492
	2	.382	.342	.322	.322	
	3	.630	.581	.557	.557	
	4	.822	.777	.763	.763	
	5	.933	.905	.899	.899	
k=4	1	.143	.121	.112	.112	.691
	2	.280	.238	.227	.227	
	3	.454	.397	.385	.385	
	4	.645	.574	.563	.563	
	5	.791	.731	.723	.723	

Table 3. Powers for n_1=5, n_2=10

	Δ	t	t*	t_{WS}	t_{SS}	F
k=1	1	.534	.530	.494	.523	.049
	2	.965	.961	.943	.958	
	3	.999	.999	.999	.999	
	4	1	1	1	1	
k=2	1	.374	.288	.219	.256	.383
	2	.753	.619	.550	.611	
	3	.954	.871	.840	.874	
	4	.996	.972	.967	.976	
	5	1	.997	.996	.998	
k=3	1	.305	.174	.143	.164	.715
	2	.573	.362	.325	.366	
	3	.805	.584	.562	.599	
	4	.937	.780	.771	.795	
	5	.988	.909	.908	.919	
k=4	1	.263	.121	.108	.122	.867
	2	.454	.238	.223	.246	
	3	.661	.402	.389	.412	
	4	.823	.575	.570	.586	
	5	.926	.733	.731	.745	

Table 4. Powers for n_1=10, n_2=5

	Δ	t	t*	t_{WS}	t_{SS}	F
k=1	1	.532	.527	.500	.533	.050
	2	.967	.964	.949	.958	
	3	.999	.999	.999	.999	
	4	1	1	1	1	
k=2	1	.221	.259	.332	.333	.230
	2	.649	.698	.776	.777	
	3	.944	.959	.979	.979	
	4	.997	.997	.999	.999	
	5	1	1	1	1	
k=3	1	.120	.184	.224	.219	.548
	2	.364	.483	.546	.537	
	3	.690	.796	.846	.840	
	4	.911	.960	.976	.974	
	5	.985	.994	.997	.996	
k=4	1	.077	.152	.174	.169	.781
	2	.222	.368	.400	.390	
	3	.454	.632	.667	.662	
	4	.703	.848	.870	.867	
	5	.882	.961	.968	.967	

Table 5. Powers for n_1=10, n_2=10

	Δ	t	t^*	t_{WS}	t_{SS}	F
k=1	1	.688	.688	.683	.683	.050
	2	.995	.995	.995	.995	
	3	1	1	1	1	
k=2	1	.388	.375	.371	.371	.484
	2	.855	.847	.843	.843	
	3	.990	.989	.989	.989	
	4	1	1	1	1	
k=3	1	.254	.235	.234	.234	.874
	2	.609	.585	.585	.585	
	3	.890	.876	.876	.876	
	4	.984	.980	.980	.980	
		.999	.998	.998	.999	
k=4	1	.192	.173	.173	.173	.977
	2	.435	.402	.401	.401	
	3	.710	.684	.684	.684	
	4	.899	.886	.886	.886	
	5	.977	.971	.971	.971	

Table 6. Powers for n_1=5, n_2=15

	Δ	t	t^*	t_{WS}	t_{SS}	F
k=1	1	.584	.579	.531	.591	.047
	2	.979	.974	.953	.972	
	3	1	1	.999	1	
k=2	1	.443	.303	.228	.285	.461
	2	.811	.625	.556	.631	
	3	.972	.869	.844	.892	
	4	.997	.968	.966	.977	
	5	1	.995	.995	.997	
k=3	1	.371	.176	.145	.181	.782
	2	.643	.355	.328	.383	
	3	.862	.585	.567	.619	
	4	.961	.779	.773	.814	
	5	.993	.913	.911	.927	
k=4	1	.333	.118	.109	.136	.910
	2	.537	.240	.231	.259	
	3	.734	.397	.389	.425	
	4	.880	.574	.570	.600	
	5	.955	.729	.727	.753	

Table 7. Powers for n_1=15, n_2=5

	Δ	*t*	*t**	t_{WS}	t_{SS}	*F*
k=1	1	.600	.595	.537	.596	.057
	2	.983	.980	.958	.976	
	3	1	1	1	1	
k=2	1	.205	.281	.399	.408	.240
	2	.708	.772	.873	.879	
	3	.971	.978	.993	.995	
	4	1	1	1	1	
k=3	1	.080	.208	.276	.275	.581
	2	.356	.576	.687	.687	
	3	.739	.883	.944	.944	
	4	.951	.985	.996	.995	
	5	.995	.999	1	1	
k=4	1	.043	.183	.210	.207	.815
	2	.183	.476	.522	.519	
	3	.462	.770	.817	.812	
	4	.753	.940	.960	.959	
	5	.930	.992	.996	.996	

3. Conclusions

It is very well known that Student's *t* test does not preserve significance level in presence of heterogeneity of variances for unequal sample sizes. The type I error is too small when the larger sample size is associated with the larger variance and it is too large in opposite case. Table 1 confirms this knowledge. The paper of Singh et al. (2002) also does. The next three tests perform well in controlling significance level. Nevertheless the "combined" test is a little worse on that score than t_{WS} and t_{SS}.

For equal sample sizes the most powerful is Student's *t* test, even if $\sigma_X = 4\sigma_Y$. For unequal sample sizes it shouldn't be taken into considerations as its type I error differs considerably from nominal 0.05. In such a case t_{SS} should be preferred. It is a little more powerful than *t** and t_{WS} especially when the smaller sample size is associated with the larger variance. Even if $\sigma_X = \sigma_Y$ the power of t_{SS} is almost equal to the power of Student's *t*.

When the smaller sample size is associated with the larger variance the "combined" test turns out to have a little greater power than t_{WS}. But we must admit that in such a case its type I error is a little too large as compared to

nominal 0.05. When the larger sample has larger variance the Welch-Satterthwaite test is better than *t**.

In any case the intuitive approach to Behrens-Fisher problem presented in "combined" test gives quite good results.

All obtained results can be generalize for hypothesis $H_0 : \mu_X - \mu_Y = \delta$. Only $\Delta = \mu_X - \mu_Y - \delta$ instead of $\Delta = \mu_X - \mu_Y$ should be considered.

References

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

Behrens W.V. (1929). Ein bei trag zur fehlerbere-Chung bei wenigen beobachtungen. *Landwirtschaftliches Jahrbuch* 68, 807-837.

Chudzik H., Kiełczewska H., Mejza I. (2006). *Statystyka matematyczna w przykładach i zadaniach*. Wydawnictwa AR w Poznaniu, Poznań.

Domański Cz. (1990) . *Testy statystyczne*. PWE Warszawa.

Domański Cz., Pruska K., Wagner R. (1998). *Wnioskowanie statystyczne przy nieklasycznych założeniach*. Wyd. Uniwersytetu Łódzkiego, Łódź.

Dudewicz E.J., Ahmed S.U. (1998). New exact and asymptotically optimal solution to the Behrens-Fisher problem, with tables. *American Journal of Mathematical and Management Sciences* 18, 359-426.

Dudewicz E.J., Ahmed S.U. (1999). New exact and asymptotically optimal heteroscedastic statistical procedures and tables. *American Journal of Mathematical and Management Sciences* 19, 157-180.

Elandt R. (1964). *Statystyka matematyczna w zastosowaniu do doświadczalnictwa rolniczego*. PWN, Warszawa.

Fisher R.A. (1935). Fiducial argument in statistical inference. *Annals of Eugenics* 6, 391-398.

Hsu P.L. (1938): Contributions to the theory of Student's t-test as applied to the problem of two samples. *Statistical Research Memoires* 2, 1-24.

Klonecki W. (1999). *Statystyka dla inżynierów*. PWN, Warszawa.

Krysicki W., Bartos J., Królikowska K., Wasilewski M. (1986). *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*. PWN, Warszawa.

Krzyśko M. (2000). *Wielowymiarowa analiza statystyczna*. UAM, Poznań.

Oktawa W. (1984). *Rachunek prawdopodobieństwa i statystyka matematyczna*. Wyd. AR w Lublinie.

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

Saxena K.K., Srivastava O.P. (1986). A new approximation to the critical point of t-distribution. *Statisztikai Szemle* (Hungary), 64, 1239-1244.

Singh P., Saxena K.K., Srivastava O.P. (2002). Power comparisons of solutions to the Behrens-Fisher problem. American *Journal of Mathematical and Menagement Sciences*, 22, 233-250.

Wagner R., Błażczak P. (1986). *Statystyka matematyczna z elementami doświadczalnictwa*. Wyd. AR w Poznaniu, Poznań.

Welch B.L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika* 29, 350-362.

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

Zieliński R., Zieliński W. (1990). *Tablice statystyczne*. PWN, Warszawa, 239-241.

PORÓWNANIE MOCY CZTERECH TESTÓW DLA PROBLEMU BEHRENSA-FISHERA

Streszczenie

Problem testowania równości średnich w dwóch populacjach normalnych z różnymi i nieznanymi wariancjami nosi w literaturze nazwę problemu Behrensa-Fishera. W pracy porównuje się moce czterech testów: *t*-Studenta, Welcha-Satterthwaite'a, Saxena-Srivastavy oraz testu składającego się z testu *F* dla równości wariancji oraz, w zależności od jego rezultatu, testu: *t*-Studenta lub Welcha-Satterthwaite'a.

Słowa kluczowe*:* problem Behrensa-Fishera, testowanie hipotez, testowanie równości średnich, porównanie mocy, nierówne wariancje

Klasyfikacja AMS 2000: 62F03