ON THE COMPARISON OF THE WELCH TEST AND THE SINGLE-STAGE TEST: A SIMULATION STUDY

MELTEM EKIZ, HAMZA GAMGAM

ABSTRACT. It is known that using the F test for testing the equality of means in a one-way ANOVA is misleading when the assumption of equal population variances is violated. When the variances are unknown and unequal, Welch [1] developed the so-called Welch test and Chen and Chen [2] developed the single-stage test for ANOVA. In this paper, Welch and single-stage tests are compared in terms of their powers and the reject ratios of the hypothesis when the hypothesis is true. Simulation results indicate that the power of the single-stage test is better than the Welch test and that this test performs well in terms of the reject ratios for small number of populations and sample sizes.

1. INTRODUCTION

The procedure of testing the equality of means in the conventional one-way ANOVA is based on the assumptions of independence, normality and equal variances. When the populations have different variances, it is well-known that the results mislead to wrong conclusions. Studies have shown that the F-statistic is not robust under the violation of equal error variances especially in case of unequal sample sizes. In case of unequal variances Cochran suggested a test method where the reciprocals of sample variances are used as weights in the sum of squares explained and he provided a chi-squared test [3]. Weighting the terms in the sum of squares explained by using the reciprocals of the estimated variances of the respective sample means was proposed by James [4] and Welch [1]. Brown and Forsythe developed a test for one-way layout under heteroscedasticity [5]. The two-stage test suggested by Dudewicz and Dalal [6] was applied on variance analysis problems by Bishop and Dudewicz [7]. Chen and Chen compared the single-stage test with the two-stage test to test the null hypothesis about the equality of means [2].

In this study the commonly used Welch test and single-stage test are compared by means of their powers and the reject ratios of a true null hypothesis as the realizations of the specified α's. The comparisons are made by simulation. The

Welch test is introduced in Section 2 and the single-stage test in Section 3. In Section 4, these two tests are compared.

2. THE WELCH TEST

Let independent $Y'_{ij}s$ denote the j-th observation from the i-th population $N\left(\mu_i, \sigma_i^2\right)$, where $i = 1, ..., I$ and $j = 1, ..., n_i$. The sample mean and variance of the i-th sample are formulated, respectively, by

$$\overline{Y_i} = \sum_{j=1}^{n_j} Y_{ij}/n_i \qquad S_i^2 = \sum_{j=1}^{n_j} \left(Y_{ij} - \overline{Y_i}^2\right)/\left(n_i - 1\right).$$

The overall sample size and sample mean are shown by the following equations

$$n = \sum_{i=1}^{I} n_i \qquad \overline{Y} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}/n = \left(\sum_{i=1}^{I} n_i \overline{Y_i}\right)/n$$

[8]. When the population variances are equal, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_I^2$, the classical F-test with degrees of freedom I-1,n-I is appropriate for testing the null hypothesis about the equality of the means, $\mu_1 = \mu_2 = ... = \mu_I$. When the population variances are unequal, the Welch test can be used [1]. This test is defined as follows:

The weights,

$$w_i = n_i/\sigma_i^2 \qquad i = 1, ..., I$$

are used to define the weighted mean

$$\mu = \left(\sum_{i=1}^{I} w_i \mu_i\right)/\sum_{i=1}^{I} w_i.$$

Usually the weights are unknown and estimated by $\widehat{w}_i = n_i/S_i^2$. Then the weighted sample mean is obtained as

$$\overline{Y}_{\widehat{w}} = \left(\sum_{i=1}^{I} \widehat{w}_i \overline{Y}_i\right)/\sum_{i=1}^{I} \widehat{w}_i.$$

Explained weighted sum of squares, the variability of the weights from population to population, can be estimated by

$$q_{\widehat{w}} = \sum_{i=1}^{I} \widehat{w}_i \left(\overline{Y}_i - \overline{Y}_{\widehat{w}}\right)^2.$$

Let $f_i = n_i - 1$ be the degrees of freedom for the i-th sample and

$$\widehat{A} = \sum_{i=1}^{I}\left[1 - \left(\widehat{w}_i/\sum_{i=1}^{I}\widehat{w}_i\right)\right]^2/f_i.$$

Then the Welch test statistic W, which has an approximate F with degrees of freedom $v_1 = I - 1$, $\widehat{v}_2 = \left(I^2 - 1\right)/3\widehat{A}$ is defined by

$$W = \frac{q_{\widehat{w}}}{(I-1)} \left\{1 + \frac{2(I-2)\widehat{A}}{(I^2-1)}\right\}^{-1}.$$

3. THE SINGLE-STAGE TEST

The single-stage test is an alternative for the Welch test. Let independent Y_{ij}'s, which is given heretofore, denote the j-th observation from the i-th population $N\left(\mu_i, \sigma_i^2\right)$, where $i = 1, ..., I$ and $j = 1, ..., n_i$. Employ the first (or randomly chosen) $(n_i - 1)$ observations to define the sample mean and sample variance, respectively, by

$$\overline{Y}_{i.} = \sum_{j=1}^{n_i-1} Y_{ij}/(n_i - 1)$$

and

$$S_i^2 = \sum_{j=1}^{n_i-1} \left(Y_{ij} - \overline{Y_i}^2\right)/(n_i - 2).$$

The weighted sample mean is

$$\overline{Y}_{i.} = \sum_{j=1}^{n_j} \widehat{w}_{ij} Y_{ij},$$

where the weights are random variables and defined as

$$\widehat{w}_{ij} = \begin{cases} U_i, & \text{for } 1 \leq j \leq n_i - 1 \\ V_i, & \text{for } j = n_i \end{cases}. \tag{3.1}$$

In literature the two different weights U_i and V_i, given in Eq.(3.1), are formulated as

$$U_i = \frac{1}{n_i} + \frac{1}{n_i}\sqrt{\frac{1}{n_i-1}\left\{\left(S_{[k]}^2/S_i^2\right) - 1\right\}}$$

and

$$V_i = \frac{1}{n_i} - \frac{1}{n_i}\sqrt{\frac{1}{n_i-1}\left\{\left(S_{[k]}^2/S_i^2\right) - 1\right\}}$$

where $S_{[k]}^2$ is the maximum of $S_1^2, ..., S_I^2$. Also U_i and V_i satisfy the following conditions:

$$(n_i - 1)U_i + V_i = 1$$

$$(n_i - 1) U_i^2 + V_i^2 = S_{[k]}^2/n_i S_i^2. \tag{3.2}$$

When the sample variances are given $\left(S_i^2 = s_i^2\right)$, the weighted sample mean $\overline{Y}_{i.}$ has a conditional normal distribution with mean μ_i and variance $\sum_{j=1}^{n_i} \widehat{w}_{ij}^2 \sigma_i^2$ [2]. Furthermore, for given value of S_i^2 $\left(S_i^2 = s_i^2\right)$ the statistic t_i defined as

$$t_i = \frac{\overline{Y}_{i.} - \mu_i}{\sqrt{s_i^2 \sum_{j=1}^{n_i} \widehat{w}_{ij}^2}}$$

has a conditional normal distribution with mean zero and variance σ_i^2/s_i^2. It was proved that the conditional normal distributions of t_i are unconditional and independent Student's t distributions with $n_i - 2$ degrees of freedom [2]. In other words, the joint p.d.f. of $t_1, ..., t_I$, given $S_i^2 = s_i^2$, was written as the p.d.f. of Student $t_i's$ with $n_i - 2$ degrees of freedom. Using equation (3.2), t_i can be written as

$$t_i = \frac{\overline{Y}_{i.} - \mu_i}{s_{[k]}/\sqrt{n_i}}.$$

These statistics are distributed as independent Student's t with degrees of freedom $n_i - 2$. When the sample sizes are equal, i.e. $n_1 = ... = n_I = n$, the overall weighted mean is formulated by

$$\overline{Y}_{..} = \sum_{i=1}^{I} \overline{Y}_{i.}/I$$

In order to test the null hypothesis the statistic $\widetilde{F}^1$,

$$\widetilde{F}^1 = \sum_{i=1}^{I} \left(\frac{\overline{Y}_{i.} - \overline{Y}_{..}}{s_{[k]}/\sqrt{n_i}}\right)^2,$$

is recommended [2]. It is the sum of squares of independent Student's t variables. Under the null hypothesis, the distribution of statistic $\widetilde{F}^1$ was obtained by simulations and the critical values were tabulated for the chosen sample sizes n-2=2,3,4,5,6,8 and number of populations I=3,4,5,6,8. For the implementation of the single-stage test $\widetilde{F}^1$ is calculated then its value is compared with the critical value $\widetilde{F}_{\alpha,I,n-2}$. If

$$\widetilde{F}^1 > \widetilde{F}_{\alpha,I,n-2},$$

the null hypothesis that the population means are all equal is rejected.

4. Comparison of the Welch Test and the Single-Stage Test

In this study, we compared the two tests in two respects. First we compared them in terms of the powers, in other words the reject ratios of the null hypothesis under unequal means. To do this, for given values of μ_i and σ_i^2 (as can be seen in Tablo 1), the data sets are generated from the distribution $Y_{ij} \sim N\left(\mu_i, \sigma_i^2\right)$, when $i = 1, \ldots I$, $I = 3, 4, 5, 6, 8$, $j = 1, ..., n_i$, $n_1 = ... = n_I = n = 4, 5, 6, 7, 8, 10$ and $\alpha = 0.25, 0.10, 0.05, 0.025, 0.01$. Then the power of the tests for each statistic are estimated after 10000 simulation runs. The average values of the powers for each test are summarized in Table 1. From Table 1 we see that the power of the single-stage test is better than the Welch test for each sample size as well as levels, when the number of populations are 3, 4. In this case the single-stage test may be a suitable choice. On the other hand the Welch test becomes more powerful as the number of populations and the sample sizes increase.

Secondly, these tests are compared with respect to the reject ratios of a true null hypothesis as the realization of the specified α's, under equal means. For this, the data sets are generated from the distributions $Y_{ij} \sim N\left(2, \sigma_i^2\right)$ after 10000 simulation runs for the same number of populations, sample sizes and variances given above. Then average values of the reject ratios are calculated and tabulated in Table 2 for level of significance $\alpha = 0.25, 0.10, 0.05, 0.025, 0.01$. From the simulation results shown in Table 2 we see that the reject ratios of the single-stage test are more close to the level of significance than those of Welch test, when the number of populations are only 3 and 4 and the sample size is 4. Furthermore in Welch test, the realization of the specified α's increase when the number of populations and the sample sizes increase. On the other hand, the estimated values in Table 2 overestimate the level α's. However in the single-stage test, the estimated values usually overestimate the α's only when I=3.

Table 1. The powers of the single-stage test and the Welch test.

		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
3	4	0.7967*	0.5384	0.3576	0.2112	0.1003
		0.5890**	0.3463	0.2184	0.1219	0.0590
	5	0.8968	0.7384	0.5729	0.4186	0.2491
		0.6696	0.4309	0.2748	0.1716	0.0923
	6	0.9344	0.8354	0.7140	0.5845	0.3991
		0.7345	0.5109	0.3554	0.2351	0.1258
	7	0.9544	0.8822	0.7979	0.6919	0.5270
		0.7910	0.5742	0.4238	0.2952	0.1674
	8	0.9652	0.9061	0.8402	0.7600	0.6266
		0.8327	0.6368	0.4977	0.3519	0.2122
	10	0.9727	0.9368	0.8904	0.8245	0.7210
		0.8917	0.7509	0.6084	0.4839	0.3228

$\mu_i = 10, 12, 14; \ \sigma_i^2 = 1, 3, 6$

$\mu_i = 10, 12, 14, 16; \ \sigma_i^2 = 1, 3, 6, 9$

		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
4	4	0.7497	0.4816	0.3118	0.1701	0.0824
		0.6560	0.3938	0.2696	0.1686	0.0889
	5	0.8899	0.6949	0.5354	0.3651	0.2068
		0.7374	0.4986	0.3486	0.2336	0.1232
	6	0.9401	0.8213	0.6900	0.5472	0.3650
		0.8099	0.5969	0.4429	0.2970	0.1767
	7	0.9610	0.8810	0.7953	0.6744	0.4953
		0.8547	0.6743	0.5194	0.3818	0.2355
	8	0.9762	0.9136	0.8346	0.7443	0.5880
		0.8987	0.7500	0.6085	0.4723	0.2952
	10	0.9927	0.9678	0.9327	0.8709	0.7693
		0.9457	0.8497	0.7455	0.6251	0.4546

$\mu_i = 10, 12, 14, 16, 18; \sigma_i^2 = 1, 3, 6, 9, 3$					
		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
5	4	0.8210	0.5286	0.3301	0.1857	0.0842
		0.9904	0.9365	0.8409	0.7020	0.5128
	5	0.9414	0.7848	0.6058	0.4261	0.2302
		0.9988	0.9876	0.9533	0.8920	0.7551
	6	0.9770	0.9030	0.7996	0.6494	0.4407
		0.9999	0.9980	0.9907	0.9718	0.9109
	7	0.9896	0.9538	0.8874	0.7873	0.5988
		1.0000	0.9998	0.9985	0.9931	0.9730
	8	0.9953	0.9834	0.9395	0.8780	0.7613
		1.0000	1.0000	0.9999	0.9989	0.9936
	10	0.9975	0.9904	0.9726	0.9458	0.8706
		1.0000	1.0000	1.0000	1.0000	0.9997
$\mu_i = 10, 12, 14, 16, 18, 12; \sigma_i^2 = 1, 3, 6, 9, 3, 1$					
		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
6	4	0.9189	0.6928	0.4618	0.2804	0.1301
		0.9939	0.9466	0.8633	0.7475	0.5662
	5	0.9883	0.9270	0.8103	0.6385	0.3893
		0.9997	0.9914	0.9667	0.9101	0.8136
	6	0.9971	0.9847	0.9506	0.8796	0.6998
		1.0000	0.9986	0.9915	0.9789	0.9324
	7	0.9991	0.9950	0.9826	0.9599	0.8790
		1.0000	0.9990	0.9947	0.9806	0.9172
	8	0.9996	0.9993	0.9945	0.9829	0.9505
		1.0000	1.0000	1.0000	0.9995	0.9969
	10	1.0000	0.9996	0.9987	0.9971	0.9895
		1.0000	1.0000	1.0000	0.9999	0.9999

$\mu_i = 10, 12, 14, 16, 18, 12, 16, 20; \sigma_i^2 = 1, 3, 6, 9, 3, 1, 6, 12$						
		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
8	4	0.8879	0.6289	0.4049	0.2470	0.1117
		0.9936	0.9601	0.8900	0.8004	0.6373
	5	0.9850	0.9087	0.7742	0.5905	0.3482
		0.9990	0.9926	0.9744	0.9390	0.8539
	6	0.9979	0.9790	0.9333	0.8496	0.6507
		1.0000	0.9990	0.9958	0.9845	0.9572
	7	0.9987	0.9957	0.9806	0.9461	0.8517
		1.0000	1.0000	0.9993	0.9974	0.9906
	8	0.9998	0.9972	0.9937	0.9775	0.9449
		1.0000	1.0000	1.0000	0.9997	0.9982
	10	1.0000	0.9996	0.9992	0.9957	0.9886
		1.0000	1.0000	1.0000	1.0000	1.0000

*: Average value for the power of the single-stage test. **: Average value for the power of the Welch test.

Table 2. The reject ratios of the single-stage test and the Welch test.

$\sigma_i^2 = 1, 3, 6$						
		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
3	4	0.2640*	0.1022	0.0489	0.0209	0.0078
		0.2572**	0.1089	0.0592	0.0314	0.0160
	5	0.3000	0.1212	0.0571	0.0264	0.0093
		0.2626	0.1130	0.0588	0.0288	0.0108
	6	0.3301	0.1293	0.0616	0.0269	0.0119
		0.2589	0.1051	0.0543	0.0281	0.0129
	7	0.3456	0.1429	0.0685	0.0310	0.0115
		0.2559	0.1062	0.0553	0.0253	0.0100
	8	0.3528	0.1407	0.0750	0.0310	0.0108
		0.2544	0.1052	0.0577	0.0277	0.0100
	10	0.3768	0.1609	0.0736	0.0354	0.0162
		0.2504	0.1082	0.0488	0.0258	0.0132

$\sigma_i^2 = 1, 3, 6, 9$						
		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
4	4	0.2230	0.0829	0.0389	0.0203	0.0071
		0.2708	0.1247	0.0726	0.0410	0.0209
	5	0.2406	0.0820	0.0387	0.0182	0.0077
		0.2643	0.1145	0.0649	0.0330	0.0172
	6	0.2702	0.0969	0.0412	0.0173	0.0067
		0.2569	0.1103	0.0615	0.0341	0.0136
	7	0.2723	0.1012	0.0444	0.0234	0.0068
		0.2565	0.1075	0.0573	0.0299	0.0161
	8	0.2822	0.1073	0.0469	0.0195	0.0086
		0.2596	0.1059	0.0521	0.0276	0.0128
	10	0.2623	0.0880	0.0346	0.0155	0.0055
		0.2606	0.1070	0.0502	0.0282	0.0123
$\sigma_i^2 = 1, 3, 6, 9, 3$						
		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
5	4	0.1967	0.0754	0.0344	0.0173	0.0066
		0.2861	0.1354	0.7726	0.0442	0.0217
	5	0.2013	0.0695	0.0297	0.0137	0.0067
		0.2722	0.1219	0.0643	0.0384	0.0187
	6	0.2097	0.0723	0.0330	0.0138	0.0054
		0.2686	0.1192	0.0631	0.0382	0.0163
	7	0.2223	0.0716	0.0330	0.0132	0.0054
		0.2740	0.1123	0.0598	0.0336	0.0161
	8	0.2274	0.0749	0.0354	0.0150	0.0057
		0.2625	0.1099	0.0515	0.0292	0.0141
	10	0.2180	0.0753	0.0329	0.0166	0.0046
		0.2510	0.1046	0.0568	0.0275	0.0122

		Level of significance α				

$\sigma_i^2 = 1, 3, 6, 9, 3, 1$

		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
6	4	0.2073	0.0836	0.0410	0.0206	0.0068
		0.2962	0.1402	0.0829	0.0528	0.0263
	5	0.2201	0.0776	0.0410	0.0166	0.0080
		0.2874	0.1253	0.0771	0.0408	0.0199
	6	0.2302	0.0789	0.0385	0.0215	0.0059
		0.2714	0.1182	0.0655	0.0346	0.0187
	7	0.2391	0.0877	0.0372	0.0172	0.0063
		0.2571	0.1095	0.0610	0.0350	0.0162
	8	0.2494	0.0844	0.0433	0.0182	0.0070
		0.2649	0.1050	0.0604	0.0292	0.0142
	10	0.2641	0.0948	0.0428	0.0180	0.0065
		0.2592	0.1072	0.0544	0.0273	0.0154

$\sigma_i^2 = 1, 3, 6, 9, 3, 1, 6, 12$

		Level of significance α				
I	n	0.25	0.10	0.05	0.025	0.01
8	4	0.1810	0.0654	0.0359	0.0137	0.0064
		0.3239	0.1668	0.0972	0.0670	0.0386
	5	0.1767	0.0635	0.0322	0.0147	0.0057
		0.2959	0.1465	0.0875	0.0520	0.0265
	6	0.1831	0.0648	0.0288	0.0145	0.0051
		0.2823	0.1277	0.0683	0.0468	0.0211
	7	0.1841	0.0635	0.0263	0.0128	0.0042
		0.2707	0.1179	0.0716	0.0413	0.0182
	8	0.1874	0.0616	0.0277	0.0112	0.0034
		0.2723	0.1203	0.0596	0.0363	0.0155
	10	0.1930	0.0604	0.0233	0.0113	0.0035
		0.2690	0.1138	0.0534	0.0328	0.0148

*: Average value for the reject ratios of the single-stage test. **: Average value for the reject ratios of the Welch test.

5. Discussion and Conclusions

The single-stage test and the Welch test both ignore the assumption of equal variances in one-way ANOVA. Therefore, in this study, these two tests are compared by means of the power of the tests and the reject ratios of true null hypothesis as the realization of the specified α's. For this, we used our computer programs (written in Matlab) based on 10000 simulation runs. Results are given in Table 1 and Table 2 for different number of groups, sample sizes and values of levels of significance. The power of the single stage test is better than that of the Welch test, when $I = 3, 4$. If I is 5 or larger, the Welch test gives better results. By comparing these two tests

in terms of the reject ratios, the single-stage test has given more close values to the level of significance α, when $I = 3, 4$ and sample size 4.

As a conclusion, the single-stage test has given better results in terms of both power and the reject ratios of true null hypothesis as the realization of the specified α's, when sample size is 4 and number of groups is 3 or 4. So, by using the single-stage test for small sample sizes and number of groups an experimenter can save time and money. Finally, the single-stage test may be his appropriate choise.

WELCH TESTİ İLE TEK AŞAMALI TESTİN KARŞILAŞTIRILMASI: BİR SİMÜLASYON ÇALIŞMASI

ÖZET:Grup varyanslarının eşitliği varsayımı sağlanmadığında tek-yönlü ANOVA'da ortalamaların eşitliği hipotezinin test edilmesinde kullanılan F testi yanıltıcı sonuçlara neden olur. Welch, ANOVA'da varyansların bilinmiyor ve farklı oldukları varsayımı altında Welch testini, Chen ve Chen ise tek-aşamalı testi önermişlerdir. Bu çalışmada Welch ve tek-aşamalı test, testin gücü ve doğru olan yokluk hipotezinin red edilmesi oranları bakımından karşılaştırılmıştır. Simülasyon sonuçlarına göre grup sayısı ve örnek çapı küçük iken, tek aşamalı test hem güç hem de red oranları bakımından Welch testinden daha iyidir.

References

[1] Welch, B.L., On the comparison of several mean values: an alternative approach, Biometrica, 38, 330-336, 1951.

[2] Chen, S.Y., Chen H.J., Single-stage analysis of variance under heteroscedasticity, Commun.Statist.-Simula, 27(3), 641-666, 1998.

[3] Cochran, W.G., Problems arising in the analysis of a series of similar experiments, Journal of the Royal Statistical Society, Suppl. 4, 102-118, 1937.

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

[5] Brown, M.B., Forsythe, A.B, The small-sample behavior of some statistics which test the equality of several means, Technometrics, 16, 129-132, 1974.

[6] Dudewicz, E.J. and Dalal, S.R., Allocations of observations in ranking and selection with unequal variances, Sankhya, 37B, 28-78, 1975.

[7] Bishop, T.A. and Dudewicz, E.J., Exact analysis of variance with unequal variances: test procedures and tables, Technometrics, 20, 419-430, 1978.

[8] Kulinskaya, E., Staudte, R.G. and Gao, H., Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances, Communications in Statistics-Theory and Methods, 32(12), 2353-2371, 2003.

Current address: Gazi University, Faculty of Arts and Science, Department of statistics, Ankara, Turkey

E-mail address: ozmeltem@gazi.edu.tr, hgamgam@gazi.edu.tr