Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 219 · 2015

Grażyna Trzpiot

Uniwersytet Ekonomiczny w Katowicach
Wydział Informatyki i Komunikacji
Katedra Demografii i Statystyki Ekonomicznej
trzpiot@ue.katowice.pl

SOME REMARKS OF TYPE III ERROR FOR DIRECTIONAL TWO-TAILED TEST

Summary: The major objective of this study was to investigate the effects of non-normality on Type III error rates for ANOVA F and its commonly recommended parametric counterparts namely Alexander-Govern test. The major objective of this study was to investigate the effects of non-normality on Type III error rates for ANOVA. Therefore these tests were compared in terms of Type III error rates across the variety of population distributions, mean difference (effect size), and sample sizes.

Key words: Type I error rates, power of test, Type III error rates.

Introduction

Many researchers from different fields of the social, biological and physical sciences are using null hypothesis significance testing. This procedures was suggested by Fisher [1926] and Neyman and E. Pearson [1928]. Researchers identify a null hypothesis of no relation or difference between one variable (i.e., the independent variable that a researcher manipulates) and another variable (i.e., the dependent variable that the researcher measures as a function of a change in the independent variable). This null hypothesis is tested against an alternative hypothesis that a statistically significant relation or difference is observed between the dependent and independent variables using an inferential statistical test, such as a *t* test or ANOVA. A relation or difference between variables is considered statistically significant if there is strong evidence that the observed relation or difference is unlikely to be due to chance. In the statistical test researchers reject the null hypothesis when the probability of incorrectly rejecting

a true null hypothesis falls beneath an established criterion level alpha (α). Alpha represents the maximum level that the researcher will accept for incorrectly rejecting the null hypothesis when the null is true, which by convention (in some fields, and by some researchers) is set at five percent. It is important to emphasize that alpha is a criterion level set by researchers, and should only be equated with the probability of making a type I error when the null hypothesis is true. When the null hypothesis is not true, there is no chance of a Type I error. Of course, the reason that we conduct a hypothesis test is because we do not know the truth of the null hypothesis. Once the researcher finds evidence for a significant relation or difference between an independent and dependent variable, the null hypothesis can be rejected, and the alternative hypothesis is assumed to be true. Null hypothesis significance testing is not without its critics and controversies over the practice and application of inferential statistics exist in many disciplines.

Table 1. The consequences of type I and II errors

Hypothesis		Reality	
		Null Hypothesis is True	Alternative Hypothesis is True
Findings	Null Hypothesis is True	$1 - \alpha$	Type II Error β
	Alternative Hypothesis is True	Type I Error α	$1 - \beta$

1. Estimation of Type III Error and Power for Directional Two-Tailed

In developing a test for deciding whether one of the k populations means is larger (smaller) than the rest, under the null hypothesis that all populations are continuous and identical, Mosteller [1948] identified three types of error for a statistical decision:

a) Type I error: the probability of rejecting the null hypothesis when it is true,
b) Type II error: the probability of failing to reject the null hypothesis when it is false,
c) Type III error: the probability of correctly rejecting the null hypothesis for the wrong reason (i.e., the risk that both the (rejected) null and (accepted) alternative hypotheses are false [Mosteller, 1948, p. 63]).

Type III error exists when the false null hypothesis is rejected but the sample having the largest (smallest) sample mean does not actually contain the largest (smallest) population mean. In other words, in testing the null hypothesis that the means discrepancy is zero, Type III error represents the risk of correctly rejecting the null hypothesis in supporting the "wrong direction" of the mean

difference (most often because the sample with the larger (smaller) sample mean does not come from the population having the larger (smaller) population mean). For a *t* test of any means difference among k population means, two immediate consequences of Mosteller's Type III error can be recognized:

1) It is encountered only when the null hypothesis is false but the predicted direction based on the alternative hypothesis does not represent the sign (i.e., direction) of the true population means difference,
2) If Type III error is possible, or encountered, then the conventional definition of statistical power for the test should be modified. For example, the conventional definition of power as "the probability that we reject the null hypothesis, say, because the rightmost population yields a sample with too many large observations" [Mosteller, 1949, p. 61] can be revised to be "the probability of both correct rejection and correct of rightmost population, when it exists" [Mosteller, 1948, p. 63].

This concept of Type III error renders important implications for statistical decisions. First, in acknowledging the possibility of Type III error, we should prefer a (directional) two-tailed test over the one tailed alternative. Given the same data, the one-tailed test tends to yield higher power than the two-tailed test if the assumed direction is correct. However, if the supported one-tailed alternative is false, its power is misleading because, instead of power, it may represent the probability of Type III error. Secondly, it is important to realize that Type III error could exist in all statistical tests (e.g., test of means, variances, correlations, proportions, etc.) as well as tests with any set of k samples (k > 1).

1.1. Type III Error and Power for t-tests

In the conventional two-tailed test, one is restricted to the choice of two hypotheses: null (H_0): $\delta = 0$ versus alternative (H_1): $\Delta \neq 0$, where Δ represents differences in one of the following parameters: means ($\Delta = \mu_X - \mu_Y$) of variables X and Y, proportions ($\Delta = \pi_A - \pi_B$), or correlation coefficients ($\Delta = \rho_A - \rho_B$) of groups A and B. For a given level of Type I error (α), the conventional test power is defined as

$$\psi_C(\alpha) = 1 - \beta,$$

where β denotes Type II error (which is defined as "the probability of failing to reject a false null hypothesis"). Kaiser [1960] proposed a test that involves three (mutually exclusive) true states of nature (H_1, H_0 and H_2) for which the three corresponding hypotheses are specified as:

Left tailed alternative	$H_1: \Delta < 0$
Null alternative	$H_0: \Delta = 0$
Right tailed alternative	$H_2: \Delta > 0$

where Δ represents one of the differences in population parameters mentioned above. This approach is often called the *directional two-tailed* test.

Recently, Leventhal and Huynh [1996], Leventhal [1999], Jones and Tukey [2000] have reviewed interest in the directional two-tailed t test, knowing the null hypothesis is false implies that one of the alternative hypothesis is true, but not which true. The directional two-tailed test makes its contribution by telling us which directional alternative to accept [Leventhal, 1999, p. 6], Mendes [2002] had a simulation study for Type III error rates of some variance homogeneity tests.

The power of a test is traditionally defined as the probability of rejecting a false null hypothesis. But, this definition is not always appropriate, Leventhal and Huynh [1996] suggested that power can be defined as the probability of correctly rejecting a false null hypothesis.
Its test power is defined as:

$$\psi_K(\alpha) = 1 - \beta - \gamma,$$

where γ represents the Type III error.
The maximum value of Type III error for *directional two-tailed* test is equal to $\alpha/2$ [Kaiser, 1960, p. 164).

This is also equal to the "power" of one-tailed test evaluated at $\alpha/2$ but the direction specified in the alternative hypothesis is "wrong." Note that the conventional and Kaiserian test powers share the same values of α and β. However, the latter is smaller by the presence of Type III error (γ).

Kaiser [1960] and Shaffer [1972] showed that the *directional two-tailed* test at a predetermined α level is equivalent to the testing of two simultaneous one-tailed hypotheses, each evaluated at $\alpha/2$. Therefore, by conducting two one-tailed tests, each at the size $\alpha/2$, ψ_K is equal to the power of the one-tailed test for which the null hypothesis is rejected, and γ is equal to the power of the other test for which the null hypothesis is retained.

There are two equivalent ways to calculate Type III error and power for the *directional two-tailed* test.

1.2. Calculation of Type III Error and Power for t-tests

Type III error, test power and planned sample size for prospective design of *directional two-tailed* tests can be obtained as follows:

(a) Estimating prospective power and Type III error for *t* tests for a given sample size (n):
   (1) Choosing an alpha level (α), say $\alpha = 0{,}01$,
   (2) Evaluating the right-tailed test at $\alpha^* = \alpha/2$, i.e., $\alpha^* = 0{,}005$, by using procedure for statistical significance evaluation (function in parametric test in RExcel or procedure PROC TTEST in SAS) and procedure for sample size determination (function in parametric test in RExcel or procedure PROC POWER in SAS),
   (3) Repeat (2) for the left-tailed test,
   (4) In (2) and (3), if in the t-test of PROC TTEST, it is found that $P(|T| < T_\alpha) < \alpha/2$ then $\psi_K(\alpha)$ value was obtained in function in parametric test in RExcel or procedure PROC POWER. On the other hand, if $P(|T| < T_\alpha) > \alpha/2$, then γ is Type III error (the power value obtained in PROC POWER).

(b) Determining the required sample size and Type III error for a desired level of power:
   (1) Choosing an alpha level (α), say $\alpha = 0{,}01$ and a desired value of power (ψ_K), say $\psi_K = 0{,}90$,
   (2) Evaluating the *directional two-tailed* test at $\alpha^* = \alpha$, i.e., $\alpha = 0{,}01$ by using function in parametric test in RExcel or PROC POWER in SAS with a specified value for power to be $\psi_K = 0{,}90$. The resulting value of Ntotal is the required sample size (n) for the Kaiserian test,
   (3) Evaluating the one-tailed test at $\alpha^* = \alpha/2 = 0{,}005$, by using PROC POWER with a specified value for Ntotal = n, is the required sample size above. Then, ψ_C is the resulting value of power. Finally, Type III error = $\gamma = \psi_C - \psi_K$.

Table 2. Selected results for two-sample *t* test with equal variances

Obs	NTotal	ψ_C	ψ_K	Type III error
1	50	0,19138	0,19021	0,001170645
2	72	0,25765	0,25711	0,000541829
3	94	0,32272	0,32245	0,000270861
4	116	0,38556	0,38542	0,000142378
5	138	0,44543	0,44535	0,000077589
6	160	0,50181	0,50177	0,000043461
7	182	0,55439	0,55437	0,000024883
8	204	0,60301	0,60300	0,000014504
9	226	0,64763	0,64762	0,000008582

Table 2 cont.

10	248	0,68831	0,68830	0,000005144
11	270	0,72518	0,72517	0,000003117
12	292	0,75841	0,75841	0,000001908
13	314	0,78822	0,78822	0,000001178
14	336	0,81485	0,81485	0,000000733

Table 3. Selected results for two-sample *t* test with unequal variances

Obs	NTotal	ψ_C	ψ_K	Type3
1	50	0,18982	0,18865	0,001177782
2	73	0,25894	0,25841	0,000530647
3	96	0,32660	0,32635	0,000259131
4	119	0,39175	0,39162	0,000133357
5	142	0,45359	0,45352	0,000071258
6	165	0,51160	0,51156	0,000039180
7	188	0,56544	0,56542	0,000022037
8	211	0,61498	0,61497	0,000012626
9	234	0,66021	0,66020	0,000007348
10	257	0,70121	0,70121	0,000004333
11	280	0,73816	0,73816	0,000002585
12	303	0,77126	0,77126	0,000001557
13	326	0,80077	0,80077	0,000000947

Table 4. Selected results for paired *t* test with dependent means

Obs	Ntotal	ψ_C	ψ_K	Type3
1	5	0,33123	0,33107	0,000163312
2	6	0,42025	0,42018	0,000069486
3	7	0,50333	0,50330	0,000030868
4	8	0,57873	0,57871	0,000014185
5	9	0,64575	0,64574	0,000006700
6	10	0,70437	0,70437	0,000003236
7	11	0,75497	0,75496	0,000001593
8	12	0,79815	0,79815	0,000000797

2. The effects of non-normality on type III error for comparing independent means

Type III error (γ) refers to correctly rejecting the null hypothesis, but incorrectly inferring the direction of the effect. Directional decisions on non-directional tests will overestimate power, underestimate sample size, and ignore the risk of Type III error under the definition of Leventhal and Huyhn [1996]. By studying the Type III error rates for tests, one can evaluate, empirically, relative merits of using the statistical tests to analyze data. Correction of the power value adjusted to the Type III error rate is much lower than the power value classically calculated, especially in small samples [Muller and Lavange, 1992; Sansgiry and Akman, 2000].

For instance, if true mean differences exist between population A and population B, or among population A, population B, and population C on some measures of interest (e.g., for two populations $\mu_A > \mu_B$, and for three populations $\mu_A > \mu_B$ and $\mu_A > \mu_c$), it would be possible for a researcher to commit two types of errors:

1) Type II error, which is the acceptance of a false null hypothesis with the conditional probability β,
2) Type III error, which is the rejection of a false null hypothesis with the conditional probability of γ and concluding a mean difference in the wrong direction (e,g,, for two populations $\mu_A < \mu_B$, and for three populations $\mu_A < \mu_B$ and $\mu_A < \mu_c$).

Note that we are only considering the case where one mean μ_A differs from the rest as opposed to general departure from equality when there are more than two groups. These two types of errors directly affect the power of a test. Under this definition of power, the probability of making a Type III error must be eliminated for calculations of power and sample size. If the direction of an effect is known, results will be more informative.

Another way to understand the directional two-tailed test is to view it as a single test evaluating three statistical hypotheses: H_0, H_1, and H_2. When testing the difference between two sample means, the hypotheses are

H_0: $\mu_1 = \mu_2$,
H_1: $\mu_1 < \mu_2$ and H_2: $\mu_1 > \mu_2$,

where H_0 is the null hypothesis, H_1, and H_2 are the alternative hypotheses.

Table 5. Relationship of the "truth" and the decision about null hypothesis

Decision	Nature			
		H_1 true	H_0 true	H_2 true
Decision about Nature	H_1 accept	Correct decision	Type I error (α)	Type III error (γ)
	H_0 accept	Type II error (β)	Correct decision	Type II error (β)
	H_2 accept	Type III error (γ)	Type I error (α)	Correct decision

Therefore, Type III error (γ) is only possible only when H_1 or H_2 is true. Two cells, accept H_2 when H_1 is true and accept H_1 when H_2 is true make different type of this error. There is no Type III error if null hypothesis is accepts. It can be seen that the non-directional two tailed test does not provide for a directional decision and, hence cannot make a Type III error, Schaffer [1972] notified that a one-tailed test could make a Type III error by accepting directional alternative when the truth falls in the opposite direction.

Therefore, in power studies, accordingly, with the revisited definition, the three-choice test's power is $\psi = 1 - \gamma$ for a given state of nature. In the simplest case, two groups with equal variance; the Type III error rate can be analytically derived from the non-central *t* distribution. The difference in means $\overline{X}_A - \overline{X}_B$ has standard error $2S^2/n$ for two samples of size *n*.

2.1. Definition of Statistical Tests

Let X_{ik} be the i^{th} observation in the k^{th} group, where $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$; let $\Sigma n_k = N$. The random variables X_{ik} are assumed to be independent and normally distributed with expected values μ_k and variances σ_k^2, The best linear unbiased estimates of μ_k and σ_k^2 are respectively:

$$\overline{X}_{\cdot k} = \frac{\sum X_{ik}}{n_k} \text{ and } S_k^2 = \frac{\sum \left(X_{ik} - \overline{X}_{\cdot k}\right)^2}{\left(n_k - 1\right)}.$$

The populations were standardized because they have different means and variances. Shape of distributions was not changed while the means were changed to 0 and the standard deviations were changed to 1. The effect sizes (standardized mean differences (δ) of 0,8 and more standard deviation) to represent large effect sizes.

We apply 0,25 standard deviation to represent small effect size, 0,75 – standard deviation to represent medium effect size. To make a difference between the population means in which generated samples were taken from, specific constant numbers in standard deviation form (δ = 0,25, 0,75) were added to the random numbers of the first population.

We have done computations for chosen distribution and each given set of parameter values and frequencies of samples for the rejection regions were counted for the ANOVA F test and the Alexander-Govern test.

ANOVA-F test (F) and Alexander-Govern (AG) test statistics were calculated (for the F test we compute F and count the frequency satisfying $F > F_{(k-1, N-k-1)}$ degree of freedom, and for Alexander-Govern test we compute AG and count the frequency satisfying $AG > \chi 2_{(k-1)}$ and a check was made to see if the hypothesis which is actually true was rejected and which is actually false was rejected at α = 0,05. The experiment was repeated. This proportion estimation is test power if the means from the populations do differ ($\mu_1 \neq \mu_2$).

Anova F test

The test statistic is done by equation:

$$F = \frac{\sum_k n_k \left(\overline{X}_{.k} - \overline{X}_{..}\right)^2 / (K-1)}{\sum_i \sum_k \left(X_{i.k} - X_{.k}\right)^2 / (N-K)},$$

where $X_{..} = \frac{\sum_k n_k \overline{X}_{.k}}{N}$ when population variances are equal, F is distributed as a central F variable with (K-1) and (N-K) degree of freedom.

Alexander-Govern Test

The test statistic is:

$$AG = \sum_{k=1}^{K} Z_k^2 ,$$

where

$$Z_k = c + \frac{\left(c^3 + 3c\right)}{b} - \frac{\left(4c^4 + 33c^5 + 240c^3 + 855\right)}{\left(10b^2 + 8bc^4 + 1000b\right)},$$

$$a = v_k - 0{,}5, \quad b = 48a^2, \qquad \text{c}=\sqrt{a \cdot ln\left(1 + \frac{t_k^2}{v_k}\right)}, \qquad t_k = \frac{\overline{X}_k - X^+}{S_{\overline{X}_k}},$$

$X^+ = \sum_{k=1}^{K} W_k \overline{X}_k$ and $v_k = n_k - 1$.

AG statistic is approximately distributed as a chi-square distribution with (K-1) degrees of freedom [Alexander and Govern, 1994].

2.2. Simulation study

A computer simulation program was used and Monte Carlo techniques to investigate the effects of non-normality on Type III error rates. The error rates of tests were evaluated under six different population shapes: Normal N(0, 1), *t* distribution with 5 df (*t*(5)), χ^2 distribution with 3 df ($\chi^2(3)$), and sample-size pairings (n_1, n_2, n_3) of (5, 5, 5), (10, 10, 10), (20, 20, 20), (30, 30, 30) and (10, 20, 30).

Distributions were generated using random number generators. The effects of Type III error on test power were more obvious, especially when sample sizes were small. The populations were standardized because they have different means and variances.

The results are presented in Tables 6-8, which contains the Type III error rates of tests when distributions were normal. That result demonstrated that the alternative tests were more robust than the F test at controlling the probability of Type III error rates. On the other hand, it can be said that AG test is more robust than the others at controlling the probability of Type III error. Probability of a rejection in the wrong direction decreased as sample size and population mean differences increased. It was also seen that the effects of small sample sizes on Type III error is more pronounced.

When samples were drawn from three $t(5)$ distributions, Type III error was higher for F test than that for AG test (Table 7). And, this was more obvious in small sample sizes and effect size (0,25). The Type III error rate was affected by total sample sizes rather than inequality in sample sizes. Under this distribution, AG test is still better.

Therefore, it can be said that the effects of $t(5)$ and $\chi^2(3)$ distributions on Type III error rates for all tests were similar. At he same time, the effect of Type III error were similar too. The superiority of the AG test can be seen for all distributions and sample sizes, because, across the all distributions, sample sizes and population mean differences, the AG test obtained higher estimates for power, lower estimates of Type III error (γ). Therefore, revisited version of test power of the AG test, $\psi = 1 - \beta - \gamma$, will be higher than the others. Power of F test is smaller than the alternatives in general,

Because, Type III error rates for F test were higher AG test in general. On the other hand, simulation results suggested that Type III error rates for tests were not affected from distribution shape.

Table 6. Type III error (%) for F and AG tests ($\alpha = 0,05$)

	N(0,1)			
	$\check{e}_1 : \check{e}_2 : \check{e}_3 = 0 : 0 : 0,25$		$\check{e}_1 : \check{e}_2 : \check{e}_3 = 0 : 0 : 0,75$	
$n_1, n_2, n_3,$	F	AG	F	AG
5,5,5	2,20	0,48	1,12	0,22
10,10,10	1,88	0,64	0,66	0,26
20,20,20	1,51	0,60	0,27	0,12
30,30,30	1,31	0,58	0,11	0,05
10,20,30	1,42	0,36	0,28	0,11

Table 7. Type III error (%) for F and AG tests (α = 0,05)

	t(5)			
	$\check{e}_1 : \check{e}_2 : \check{e}_3 = 0 : 0 : 0{,}25$		$\check{e}_1 : \check{e}_2 : \check{e}_3 = 0 : 0 : 0{,}75$	
$n_1, n_2, n_3,$	F	AG	F	AG
5,5,5	1,96	0,31	1,12	0,17
10,10,10	1,79	0,55	0,95	0,33
20,20,20	1,61	0,50	0,55	0,25
30,30,30	1,47	0,45	0,35	0,17
10,20,30	1,56	0,34	0,55	0,14

Table 8. Type III error (%) for F and AG tests (α = 0,05)

	chi(3)			
	$\check{e}_1 : \check{e}_2 : \check{e}_3 = 0 : 0 : 0{,}25$		$\check{e}_1 : \check{e}_2 : \check{e}_3 = 0 : 0 : 0{,}75$	
$n_1, n_2, n_3,$	F	AG	F	AG
5,5,5	1,88	0,26	1,22	0,31
10,10,10	1,66	0,62	0,80	0,49
20,20,20	1,70	0,87	0,41	0,35
30,30,30	1,47	0,45	0,35	0,17
10,20,30	1,40	0,47	0,37	0,22

Conclusion

The results of the present simulation of the Type III error rates of the ANOVA F and its three commonly recommended parametric alternatives indicate that the AG test provides a considerable advantage over the F test in all experimental conditions. Because, in almost every experimental situation, the Type III error rates were lower for the AG test and the power of the AG test was higher than the others in many cases.

References

Alexander R.A., Govern D.M. (1994), *A new and simpler approximation for ANOVA under variance heterogeneity*, "Journal of Educational Statistics", No. 19.

Fisher R.A. (1926), *The arrangement of field experiments*, "Journal of the Ministry of Agriculture of Great Britain", No. 33.

Jones L.V., Tukey J.W. (2000), *A sensible formulation of the significance test*, "Psychological Methods", No. 5.

Kaiser H.F. (1960), *Directional statistical decision*, "Psychological Review", No. 67.

Knapp I.R. (1999), *Letter to the Editor*, "Nursing Research", No. 48.

Leventhal L., Huynh C.L. (1996), *Directional decisions for two-tailed tests: Power, error rates, and sample size*, "Psychological Methods", No. 1(3).

Leventhal L. (1999), *Updating the debate on one- versus two-tailed test with the directional two-tailed test*, "Psychological Reports", No. 84.

Leventhal L. (1999), *Answering two criticisms of hypothesis testing*, "Psychological Reports", No. 85.

Meeks S.L., D'Agostino R.B. (1983), *A note on the use of confidence limits following rejection of a null hypothesis*, "The American Statistician", No. 57(2).

Mendes M., (2002), *The comparison of some parametric alternative test to one-way Analysis of Variance in terms of Type I error rates and power of test under non-normality and heterogeneity of variance*, PhD. Thesis, Ankara University Graduates School of Natural and Applied Sciences Department of Animal Science (unpublished).

Mosteller F. (1948), *A k-sample slippage test for an extreme population*, "Annals of Mathematical Statistics", No. 19.

Muller K.E., Lavange L.M. (1992), *Power calculations for general linear multivariate models including repeated measures applications*, "Journal of the American Statistical Association", No. 87.

Neyman J., Pearson E. (1928), *On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I*, "Biometrika", Vol. 20A, No. 1-2.

Sansgiry P., Akman O. (2000), *Transformations of the lognormal distribution as a selection model*, "The American Statistician", No. 54 (4).

Shaffer J.P. (1972), *Directional statistical hypotheses and comparisons among means*, "Psychological Bulletin", No. 77.

Zumbo B.D., Hubley A.M. (1998), *A note on misconceptions concerning prospective and retrospective power*, "The Statistician", No. 47, Part 2.

UWAGI O BŁĘDZIE III RODZAJU DLA KIERUNKOWYCH TESTÓW O DWUSTRONNYM OBSZARZE KRYTYCZNYM

Streszczenie: Główny cel tej pracy to zbadanie konsekwencji braku normalności rozkładu dla błędu III rodzaju w teście ANOVA F oraz jego parametrycznym odpowiedniku, mianowicie w teście znanym jako Alexander-Govern. Testy zostały porównane pod względem poziomu błędów III rodzaju przez wybór różnych rozkładów zmiennych losowych, o różnych średnich oraz wariancjach (wielkość efektu) oraz różnych (małych) wielkościach próby.

Słowa kluczowe: błąd I rodzaju, moc testu, błąd III rodzaju.