▼ Journal of Research (Science), Bahauddin Zakariya University, Multan, Pakistan.
Vol.12, No.2, December 2001, pp. 180-188 ISSN 1021-1012

EFFECT OF DEPARTURES FROM STANDARD ASSUMPTIONS USED IN ANALYSIS OF VARIANCE

Hayat M. Awan
Institute of Management Sciences, Bahauddin Zakariya University, Multan.

Abstract: In this paper the effect of non-normality and non-homogeneity on the Analysis of Variance is investigated. Three non-normal distributions have been selected (Poisson, Exponential and Lognormal) for the assessment of such effect. The empirical distributions of MST, MSE and F are obtained when the samples are generated from these non-normal distributions and then these are compared with corresponding results obtained under the normal distribution assumption. The variances of the empirical distributions of MST, MSE and covariance of MST and MSE have been also computed under the assumption of these non-normal distributions and comparison was made with normal case of equal variances and independence. These results show that non- normality and non-homogeneity have very little effect on the Analysis of Variance and correlation between MST and MSE tends to zero with the increase in the sample size per group. However the Analysis of Variance test is conservative and one has to provide more protection for non-normal distribution of larger peakedness, like Lognormal for the small size per group.

Keywords: Analysis of variance, empirical distribution, F, MSE, MST, non-homogeneous, non-normal.

INTRODUCTION

Since most of the statistical procedures are derived under the set of specific assumptions about generating such methods and statistical tests. It has been the subject of several investigations to determine how sensitive the conclusion drawn from the data are to departure from the undergoing assumptions. The F-test is used for testing the equity of means in the analysis of variance. The effect of using the F-test when assumptions are violated has been topic of interest and research for sometime. A review and summary of much of earlier work is given in Schefee [1959] and Donaldson [1966]. The general conclusion however is that F-test in case of non-normality has little effect on inferences about the equality of means. The same is true concerning the inequality of error variances, when samples are of equal sizes. The insensitivity of statistical procedures to underlying assumptions is referred to as robustness. However, there are some serious effects of violation of these assumptions on the inference made. The degree of non-normality is more serious because the sample means and variances for non-symmetric are correlated and unequal variances within means are expected when null hypothesis is false.

It should be mentioned that Baker was the first to derive the distribution of "t" for a sample of two items from a composition of two normal functions with different centers. Subrahmanium *et al*. [1975], and Lee and Gurland

[1977] have studied the behavior of some test of significance when sampling is from mixture of two univariate normal populations. The effect of mixture of two multivariate normals on Hotelling's T^2, simple, multiple and partial correlation co-efficient have been extensively studied [Srivastava and Awan 1982 and 1984, Awan 1989].

Awan [1983 and 1989] investigated the effect of contaminated normal and Inverse Guassian distributions on the inference robustness studies of location parameters. Some other studies [Gayen 1950, David and Johnson 1951, Box and Tiao 1964, Carter *et al.* 1991] were made to see the effect of non-normality and homogeneity on "t" and F tests.

Considering this aspect, the three non-normal distributions; Exponential, Lognormal (Continuous) and Poisson (Discrete) are taken for this study. These distributions were used as they are very common in practice and also they allow one to determine the effects of extreme non-normality on F-test.

The lognormal distribution arises from a theory of elementary errors combined by a multiplicative process, just as the normal distribution arises form a theory of elementary errors combined by addition. There are, as Aitchison and Brown [1976], and Galton [1979] have pointed out, many situations in nature where it is more reasonable to suggest that the process underlying change or growth is multiplicative rather than additive.

Many examples of lognormal distribution have been noted in the nature from a variety of fields, particularly in economics many phenomena like income distributions, measure of concentration of income and consumer demand can nicely be represented by the lognormal distribution [Roy 1950, Champernowne 1953, Aitchison and Brown 1954, Brown 1955].

In this study we try to assess the effect of the non-Normality on F-test used for testing the differences among sample means by varying the degree of Non-Normal Populations we have selected. The study was limited to the single classification analysis of variance in which the F-test is used to test for differences among the means of k cells. Various researchers have discussed the effect of such non-normality on equality of means.

SAMPLING PROCEDURE

In a simulation study a random sample of n observations from distributions specified in each cell of single classification layout has been selected. The type of parent distribution and its mean and variance for each of k cells was specified. An F-test was computed for k random samples. The computer listed the distribution of MST, MSE and F. MST is defined as mean square between treatments (k-cells) whereas MSE as mean square error (The unbiased estimate of σ^2). The F-test for null Hypothesis (H_0) of no difference among population means is F=MST/MSE Where F has (k-1), k (n-1) degrees of freedom (hereafter called as d.f.). That is under H_0: MST has a chi-square distribution with k-1 d.f. Whereas

MSE has chi-square with k (n-1) d.f and two chi-squares are independent. These means squares are computed from between and within cells variation in X_{ij} and are

$$MST = \frac{n}{k-1}\sum_{j=1}^{k}(\overline{X}j - \overline{X})^2$$

and

$$MSE = \frac{1}{k(n-1)}\sum_{j=1}^{k}\sum_{i=1}^{n}(Xij - \overline{X}j)^2$$

Where

$$Xj = \frac{\sum Xij}{n} \text{ and } X = \frac{\sum\sum Xij}{n.k}$$

The process of selection of sample is repeated 10,000 times for these chosen distributions. Thus when null hypothesis was true (when means of parent distributions were equal, normality and homogeneity held) then approximately 10,000.α of obtained F's exceeded F_α, where F_α is the value of F tabulated for specific values of n and k. The tabulated F had d.f. given by (k-1) and k(n-1). For example k=4, n=16, $F_{0.05,\ 3,60}$ = 2.76. Observed Type I error is obtained by counting the number of F's that actually exceeded 2.76 and dividing by total number of replications.

Under null assumptions, we used different combinations of the parameters in the selected non- normal distributions to see the effect of non-normality and heterogeneity. However, for the sake of brevity the tabular values are given for the values specified for each non-normal distribution and corresponding coefficients of skewness and kutosis, γ_1 and γ_2 are given below:

For exponential distribution $f(x) = \beta\, e^{-}\beta x$, $x>0$, $\beta=10$

Therefore, $\gamma_{1=}\, 2$ and $\gamma_{2\,=}\, 6$ for all values of β

For lognormal distribution $f(x) = \frac{1}{x(2\pi.\sigma)^{1/2}} .e^{-1/2[(\log x-\mu)/\sigma]^2}$

Where $y\sim N(\mu, \sigma)$ and $x=e^y$

then, $\gamma_1 = \eta^3+3.\eta$ and $\gamma_2 = \eta^8{}_+ 6\eta^6{}_+ 15\eta^4{}_+ 16\eta^2$ and $\eta = e^{\sigma^2} -1$

and $\sigma^2=100$ and $\mu=10$ generate the values of. γ_1 and γ_2 as 6 and 38 respectively.

For Poisson distribution $P(x) = \beta^x\, e^{-\beta} / x!$, x = 0,1,2 ---- and β = 4, and the values of coefficients of skewness and kutososis are $\gamma_1 = 1/\sqrt{\beta}$ and $\gamma_2 = 1/\beta$.

RESULTS

The effect of non-normality when H_0 is true (α, the Type I error) will be presented in tabular form when the within cell variances are equal. The effect of non-normality on F as a function of correlation between numerator and denominator will be discussed & some empirical and analytical results presented.

The observed Type I error (ά) for each of three distributions is shown in Table 3.1 for three non-normal distributions (where α is true type I error for F-test, when assumptions underlying the test are met).

Table 3.1: Observed values of type I errors (ά) corresponding to α =0.10

		n			
K	Distribution	4	8	16	32
	Lognormal	.074	.085	.094	.096
2	Exponential	.087	.098	.099	.099
	Poisson	.093	.099	.099	.0
	Lognormal	.072	.079	.087	.093
4	Exponential	.082	.093	.096	.099
	Poisson	.091	.096	.099	.099
	Lognormal	.070	.076	.083	.092
8	Exponential	.080	.091	.095	.095
	Poisson	.089	.094	.098	.099

It is observed in this table that non-normal distributions lead to conservative type I error i.e. observed values are always smaller than theoretical level. Thus if a test is designed with α protection against a type I error under assumption of normal distribution even more protection against a type I error exists if distribution is of non-normal type specified here. It may be noted that as sample size increases the difference between the ά and α computed from normal and non-normal distribution decreases. Further the size of (α – ά) increases as size of skewness and kurtosis increases.

DISTRIBUTION OF MST AND MSE

The effect of underlying distribution on F-test originates from the combined influences of numerator (MST) and denominator (MSE). When the underlying distribution is normal, MST and MSE are distributed as chi-square and they are independent. If the underlying distributions are non-normal then MST and MSE are not distributed as chi-square, nor they are independent. Thus the effect of non-normality on F may be due to deviation in distribution of MST and MSE from that of chi-square and to the correlation between MST and MSE.

Cumulative empirical distribution of MST and MSE under H_0 are obtained in case of the samples are taken from the normal distribution and the other three non-normal distributions already specified.

In normal case.

$$\chi_v^2 = v.s^2/\sigma^2$$

where s^2 is the sample variance and

$$MST_{(1-\alpha)} = \frac{\sigma^2 X^2_{(1-\alpha),(k-1)}}{k-1}$$

Results indicate that effect of non-normality on MST is slight compared to its effect on MSE and further as n increases, the effect on MST

decreases, while it gets worse for MSE, which is shown in Tables 3.2 and 3.3.

Table 3.2: Distribution of MST

n=4 k=2 Relative Cumulative frequencies					n=32, k=2 Relative Cumulative frequencies				
Class Interval	Normal	Poisson	Exponential	Lognormal	Class Interval	Normal	Poisson	Exponential	Lognormal
00-3.2	0.6280	0.6480	0.6615	0.6841	00-4.8	0.7280	0.7320	0.7415	0.7485
3.2-6.4	0.7920	0.8160	0.8310	0.8532	4.8-9.6	0.9000	0.9140	0.9180	0.9213
6.4-9.6	0.8760	0.8980	0.9100	0.9312	9.6-14.4	0.9440	0.9560	0.9612	0.9710
9.612.8	0.9240	0.9360	0.9480	0.9535	14.4-19.2	0.9730	0.9720	0.9725	0.9785
12.816.0	0.9500	0.9600	0.9689	0.9745	19.2-24.0	0.9860	0.9840	0.9886	0.989
16.0-19.2	0.7680	0.9720	0.9810	0.9915	24.0-28.8	0.9960	0.9920	0.9925	0.9935
19.2-22.4	0.9880	1.00	1.00	1.00	28.8-33.6	1.000	0.9960	0.9062	0.9972
					33.6-38.4	1.000	0.9960	0.9964	0.9975
					38.6+	1.000	1.000	1.000	0.9964

Table 3.3: Distribution of MSE

N=4, k=2 Relative Cumulative frequencies					n=32, k=2 Relative Cumulative frequencies				
Class Interval	Normal	Poisson	Exponential	Lognormal	Class Interval	Normal	Poisson	Exponential	Lognormal
00-1.2	0.06	0.0760	0.0825	0.0875	0-.774	0.0000	0.0040	0.0050	0.0060
1.2-2.4	0.2720	0.2800	0.2845	0.2915	0.774-1.548	0.0000	0.0040	0.0065	0.0078
2.4-3.6	0.5080	0.5200	0.5280	0.5312	1.548-2.322	0.0040	0.0040	0.0073	0.0082
3.6-4.8	0.7000	0.7720	0.7167	0.7226	2.322-3.096	0.0480	0.0520	0.0612	0.0654
4.8-6.0	0.8280	0.8400	0.8510	0.8582	3.096-3.870	0.4520	0.4526	0.4575	0.4591
6.0-7.2	0.9080	0.9160	0.9190	0.9208	3.870-4.844	0.8200	0.8263	0.8305	0.8372
7.2-8.4	0.9520	0.9560	0.9572	0.9576	4.844-5.418	0.9620	0.9685	0.9697	0.9712
8.4-.96	0.9800	0.9880	0.9894	0.9900	5.418-6.192	0.9960	0.9968	0.9972	0.9980
9.6+	1.000	1.000	1.000	1.000	6.192+	1.0000	1.000	1.000	1.000

The calculation of MST depends only upon averages, thus by central limit theorem, as n increases, MST becomes less sensitive to non-normality. Further as n or k increases, the effect of non-normality on variance of MST rapidly reduces.

As $\text{Var (MST)} = \frac{2\sigma^4}{k-1}\{1+ \tfrac{1}{2}.\gamma_2 . (k-1)/nk\}$

The actual rate at which var(MST) converges to its normal theory for distribution used in this study is shown in Table 3.4 as a function of n, k.

If $\gamma_2 > 0$, then from following formula it is apparent that variance of MSE in non-normal case is larger than in Normal case, further as n increases the variance relative to normal decreases.

As $\text{Var (MSE)} = \frac{2\sigma^4}{k(n-1)}\{1+ \tfrac{1}{2}\gamma_2 (n-1)/n\}$

Table 3.4: Variance (MST) for non-normal relative to Normal Distribution

N	POISSON DISTRIBUTION k			EXPONENTIAL DISTRIBUTIONS k		
	2	4	8	2	4	8
4	1.0156	1.0078	1.0039	1.3750	1.5625	1.6563
8	1.0078	1.0079	1.0019	1.3750	1.5625	1.6563
16	1.0039	1.0019	1.0009	1.0938	1.1406	1.1641
32	1.0019	1.0009	1.0004	1.0469	1.0703	1.0820

N	LOGNORMAL DISTRIBUTION k		
	2	4	8
4	3.3750	4.5625	5.1563
8	2.1875	2.7813	3.0781
16	1.5938	1.8906	2.0391
32	1.0742	1.1113	1.1299

Table 3.5: Distribution of F

n-4, k=2

Relative Cumulative frequencies

Class Interval	Normal	Poisson	Exponential	Log normal
0.0-1.60	0.7491	0.7280	0.7012	0.6813
1.60-3.12	0.8780	0.8440	0.8142	0.7989
3.12-4.80	0.9309	0.8840	0.8671	0.8412
4.80-5.40	0.9572	0.9320	0.9215	0.9220
5.40-8.00	0.9719	0.9520	0.9412	0.9319
8.00-9.60	0.9807	0.9600	0.9501	0.9488
9.60-11.20	0.9864	0.9760	0.9681	0.9532
11.20-12.80	0.9902	0.9880	0.9767	0.9301
12.80-14.40	0.9938	0.9880	0.9812	0.9718
14.40-15.00	0.9947	0.9960	0.9932	0.9812
15.00-17.60	0.9961	0.9960	0.9941	0.9912
17.60-19.20	0.9972	1.0000	1.000	1.000
19.60- +	1.0000	1.0000	1.000	1.000

n=32, k=2

Relative Cumulative frequencies

Class Interval	Normal	Poisson	Exponential	Log normal
00-.80	0.6251	0.6120	0.6081	0.6062
0.80-1.60	0.7890	0.7600	0.7582	0.7561
1.60-2.40	0.8732	0.8480	0.8423	0.8403
2.40-3.30	0.9211	0.9040	0.8914	0.8801
3.30-4.00	0.9497	0.9440	0.9312	0.9285
4.00-4.80	0.9673	0.9680	0.9592	0.9501
4.80-5.60	0.9784	0.9840	0.9716	0.9700
5.60-6.40	0.9855	0.9880	0.9786	0.9718
6.40-7.20	0.9901	0.9920	0.9862	0.9800
7.20-8.00	0.9951	0.9960	0.9900	0.9840
8.00-9.60	0.9965	1.0000	0.9918	0.9882
9.60-10.40	0.9974	1.0000	0.9968	0.9912
10.40- +	1.0000	1.0000	1.000	1.000

The important point is the speed with which, F based on non-normal distribution, approaches its normal theory values as n increases. The

results of this study indicate, the convergence is very rapid and only small error can be expected at a sample size of even 32 given in Table 3.5.
Conservative feature of F based on non-normal distributions may also be explained with the help of correlation between MST and MSE which is of substantial size for larger values γ_2 , as Cov(MST, MSE) = $\gamma_2 \sigma^4 / nk$

$$\text{Var(MST)} = \frac{2\sigma^4}{k-1} \{1 + \tfrac{1}{2}\gamma_2 (k-1)/nk\}$$

$$\text{Var(MSE)} = \frac{2\sigma^4}{k(n-1)} \{1 + \tfrac{1}{2}\gamma_2 (n-1)/n\}$$

$$\text{Therefore } \rho = \gamma_2 / [\{\frac{4n^2.k}{(k-1)(n-1)} + \frac{2n(nk-1)}{(k-1)(n-1)} \gamma_2 + \gamma^2_2\}^{1/2}]$$

It is apparent that as n increases then $\rho \rightarrow 0$
The values of ρ as a function of n and k are shown in Table 3.6.

Table 3.6: Correlation Between MST and MSE

N	Exponential			Lognormal			Poisson		
	k			k			k		
	2	4	8	2	4	8	2	4	8
4	0.435	0.599	0.524	0.811	0.854	0.868	0.0363	0.0443	0.0477
8	0.338	0.399	0.423	0.716	0.777	0.798	0.0276	0.0337	0.0364
16	0.252	0.302	0.322	0.594	0.668	0.685	0.0202	0.0247	0.0266
32	0.183	0.221	0.238	0.466	0.541	0.570	0.0145	0.0177	0.0191

CONCLUSIONS

The simulated results indicates that the non-normal distributions lead to conservative type I error i.e. the observed values of $\alpha - \acute{\alpha}$ (level of signification) are always smaller then the assumption of normality in the analysis of variance. The difference of this observed level of significance with tabular value ($\alpha - \acute{\alpha}$) increases those non-normal distributions whose parchedness increases. The effect of this non-normality is the highest in case of lognormal and is negligible for the case of Poisson and moderate for the exponential distribution. The explanation of this phenomenon is that value of peakedness is the highest for lognormal and is quite low for Poisson distribution. We have also simulated these results for other values of parameters of these non-normal distributions (the detail is given by Awan [2001]) and found that

i.) The effect of changes in the parameters of exponential distribution is almost negligible on observed level of significance, empirical distributions of MST, MSE and F. Also variances of MST, MSE and covariance of MST and MSE remain stable for different values of mean of exponential distribution. It is also seen that the effect of has non-normality reduces with the increase in the sample size per group; however, with the increase in number of groups with same sample size, the level of significance reduces and the test is a little bit becomes more conservative. Hence, one have to need more protection but can safely apply analysis of variance on the data which follow the exponential distribution.

ii.) The results of lognormal indicate that effect is very pronounced on all the simulated results tabulated for level of significance, distributions of MST, MSE, and F and variances of MST, MSE and covariance of MST and MSE. Therefore, it is suggested that the analysis of variance for the equality of means should be used very consciously when data is being generated from the lognormal population. We used different combinations values of parameters μ and σ^2 in the lognormal distribution in these simulation study and it is found that the effect is highest for the combination $\mu = \sigma^2$. The other combinations also lead to the conservative values of the observed level significant, but these are low as compared to the combination already mentioned. It shows that the peaked ness of lognormal normal is largest at $\mu = \sigma^2$

iii.) The larger values of mean of Poisson distribution leads to smaller effect on the observed level of significance, distributions of MST, MSE, F and variances of MST, MSE and covariance of MST and MSE. The peakedness of Poisson distribution is the function of mean and as it increases, the parchedness of Poisson distribution reduces to normal distribution value. It is also observed that the effect of this non-normality on the analysis of variance is almost negligible.

iv.) The effect of non-normality decreases with the increase in sample size per group, however this effect increases with the increase in the number of groups.

v.) The coefficient of variation is approximately constant in the sample considered from the lognormal distribution. F-test is significantly affected in case of sample is drawn from the lognormal distribution for n small. It justifies for the logarithmic transformation of data. Satisfactory results may however be obtained from analysis of variance solely because a logarithmic transformation achieves stabilization. This approach has been studied by Curtiss [1943] who establishes the results under fairly general conditions; and Cochran [1938] advocates the transformation even when the untransformed data seem to indicate constant variance.

References

Aitchison, J. and Brown, J.A.C, (**1954**) “On the criterion for description of income distribution” *Metro economic*, 6, 88.

Aitchison, J. and Brown, J.A.C. (**1976**) “The Lognormal Distribution with special reference to its uses in Economics”, Cambridge University Press, Cambridge.

Awan, H. M. (**1983**) “Inference robustness studies of location parameters of a class of contaminated normal distribution” *Journal of Pure and Applied Sciences*, 2(2), 63-75.

Awan, H. M. (**1989**) “Inference Robustness Studies in the contaminated Inverse Guassian distribution”; *Karachi Univ. Journal of Science*, 17(18) 2, 101-110.

Awan, H.M. (**1989**) "On the robustness of partial and multiple correlation coefficients in sampling from a mixture of two normals." *Journal of Engineering and Applied Sciences*, 8(2), 129-137.

Awan, H.M. (**2001**) "Effects of departures from standard assumptions in analysis of variance", Working paper No.2, Institute of Management Sciences, Bahauddin Zakariya University, Multan.

Baker, G.A., "Distribution of means divided by the student deviation of samples from non-homogeneous populations", *Ann. Math. Statistic* 3, 1-9

Box, G.E.P. and Tiao, G.C. (**1964**) "A note on criticism robustness and inference robustness", *Biometrika*, 51, 169-73.

Brown, J.A.C. (**1955**) "Economics of nutrition and family budget." *Proc. Nutr. Soc.*14, 63.

Carter, E.M., Khatri G.G. and Srivastava, M.S. (**1991**) "The effect of inequality of variances on t-test", Sankhya B.

Champernowne, D.G. (**1953**) "A model of income distribution." *Econ. J.* 63, 318.

Cochran, W.G. (**1938**) "Some difficulties in statistical analysis of replicated experiments." *Emp. J. Exp. Agric*. 6, 157.

Curtiss, J.H. (**1943**) "On transformation used in Analysis of variance", *Ann. Math. Statist*. 14, 107.

David, F.M. and Johnson, M.L. (**1951**) "The effect of non-normality on the power function of F-test in the analysis of variance" *Biometrika*, 38, 43.

Donaldson, T.S. (**1966**) "Power of the F-test for Non-normal distribution and unequal variances", Memorandum RM-5072 PR.

Galton, F. (**1879**) "The geometric mean in vital and social statistics", *Proc.Roy. Soc.* 29, 365.

Gayen, A.K. (**1950**) "The distribution of variance ratio in random samples of any size drawn from non-normal universes" *Biometrika*, 37, 236

Lee, A.F.S. and Gurland, J. (**1977**) "One sample t-test when sampling from a mixture of normal distribution", *Ann. Statist*. 5(8), 3-7.

Roy, A.D. (**1950**) "The distribution of earnings and individual output" *Econ. J*. 60, 489.

Scheffe, S.H. (**1959**) "The analysis of variance", John Wiley.

Srivastava, M.S. and Awan, H. M. (**1984**) "On the robustness of the correlation coefficient in sampling from a mixture of two Bivariate normals" *Comm. Statist. Theor.Meth*. 13(3), 371-382.

Srivastava, M.S. and Awan, H.M. (**1982**) "On the robustness of Hotelling's T^2 test and the distribution of linear and quadratic forms in sampling from a mixture of two multivariate normal populations", *Comm. in statistics*, 11(1), 81-107

Subrahmaniam, K., Subrahmaniam, K. and Messri, J.Y. (**1975**) "On robustness of some tests of significance in sampling from a compound normal population", *J. Amer. Statist. Asso*. 70, 435-38.