# One-way ANOVA with Bimodal Error Terms


## Abstract

In this paper, we assume the error distribution of one-way ANOVA as alpha-skew normal distribution. Alpha-skew normal distribution gives us flexibility for modelling the data which has heavy tailness, skewness, bimodality and also symmetricity. We obtain the maximum likelihood estimator of the model parameters and the test statistics based on these estimators. Monte Carlo simulation study states that the maximum likelihood estimators of the parameters of interest are more efficient than the corresponding traditional estimators based on normality. Additionally, the test statistics based on maximum likelihood estimators is much more powerful than the test statistics based on traditional normal theory. At the end of this study, a real-life example is made just for illustration of the proposed methodology.



*Keywords: One-way ANOVA; Alpha skew-normal; Bimodal Data*

2010 Mathematics Subject Classification: 62E10; 62P99


## 1 Introduction

One-way analysis of variance is a statistical tool used to test whether there is a significant difference between the means of independent groups. Traditionally, normality assumption is made for the random error terms and the well-known least squares (LS) method is used for estimating the model parameters. However, in statistical literature, it is also pointed out that non-normal distributions are more common than the normal distribution, see . It is also known that, when the normality assumption is not satisfied, the LS estimators of the model parameters and the $F$ statistics based on these estimators lose their efficiency. ANOVA models are reasonably robust against certain types of departures from the model and Type I error is not much different than that for a normal distribution and the power of the $F$ test is considerably lower than that for a normal distribution, see, Geary (1947), Tiku (1971), Donaldson (1968), Spjotvoll and Aastveit (1980), Tan and Tiku (1999) and Celik (2022).

There are some studies based on nonnormal error term in ANOVA models. Senoglu and Tiku (2001) assume the distribution of error terms in one and two-way ANOVA as Weibull and generalized logistic distribution and obtain the modified maximum likelihood estimators and the test statistics

based on these statistics. Celik et al (2015) study the one-way ANOVA with skew normal error terms and Celik and Senoglu (2019) obtain the estimators of unknown model parameters in one-way ANOVA with Type II censored skew normal error terms. In this study, we assume that the distribution of error terms in one-way ANOVA model is Alpha-skew normal (ASN) proposed by Elal-Olivereo, (2010). The reason that we chose the ASN distribution as an error distribution one-way ANOVA that it is flexible enough to support both unimodal and bimodal shape.

The probability density function (pdf) and the cumulative distribution function (cdf) of the ASN distribution are

$$f(x) = \frac{(1-\lambda x)^2 + 1}{2+\lambda^2}\phi(x) \tag{1.1}$$

and

$$F(x) = \Phi(x) + \lambda\Big(\frac{2-\lambda x}{2+\lambda^2}\Big)\phi(x) \tag{1.2}$$

respectively, where $\lambda$ is shape parameter and $\phi(.)$ and $\Phi(.)$ are the pdf and the cdf of standard normal distribution.

It can be easily seen that if the shape parameter $\lambda$ goes to 0, the ASN distribution converges to standard normal distribution. Additionally, when the shape parameter converges to $\pm\infty$, the distribution becomes the bimodal normal distribution. The location-scale form of the ASN distribution can be obtained as

$$f(x;\mu,\sigma,\lambda) = \frac{[1-\lambda(\frac{x-\mu}{\sigma})]^2 + 1}{(2+\lambda^2)\sigma}\phi\Big(\frac{x-\mu}{\sigma}\Big). \tag{1.3}$$

The ASN distribution has at most two modes and the expected value and the variance of the ASN distribution can be derived as

$$E(X) = -\frac{2\lambda}{2+\lambda^2}, \text{ and } V(X) = \frac{2+3\lambda^2}{2+\lambda^2} - \Big(\frac{2\lambda}{2+\lambda^2}\Big)^2. \tag{1.4}$$

The rest of the paper is organized as follows. In Section 2, a brief information about one-way ANOVA is given and then, the estimators of parameters and the test statistics based on these estimators are obtained assuming the error distribution is the ASN. The Monte-Carlo simulation study is conducted in order to compare the efficiencies of the estimators and the powers of the proposed test statistics with the traditional ones in Section 3. A real-life example is analyzed in Section 4 to present the application of the proposed methodology. Conclusion is given at the end of the paper.

# 2 One-way ANOVA

Consider the following one-way ANOVA model,

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}; \;\; i = 1,2,...,a; j = 1,2,...,n_i \tag{2.1}$$

where, $y_{ij}$ are the responses corresponding to jth observation in the ith treatment, $\mu$ is the overall mean, $\alpha_i$ is the effect of ith treatment and $\epsilon_{ij}$ are the independent and identically distributed (iid) random error terms. Assume that

$$\epsilon_{ij} \sim ASN(0,\sigma,\lambda); \tag{2.2}$$

and the probability density function of $y_{ij}$ becomes

$$f(y;\mu,\sigma,\lambda) = \frac{(1-\lambda z_{ij})^2 + 1}{(2+\lambda^2)\sigma}\phi(z_{ij}) \tag{2.3}$$

where $z_{ij} = \frac{y_{ij}-\mu-\alpha_i}{\sigma}$ .

To obtain the maximum likelihood (ML) estimators of the model parameters, the following log-likelihood function is maximized with respect to parameters of interest.

$$lnL(\mu,\sigma,\lambda) = \sum_{i=1}^{a}\sum_{j=1}^{n} ln[(1-\lambda z_{ij})^2+1] - \sum_{i=1}^{a}\sum_{j=1}^{n} ln(\sigma(2+\lambda^2)) \\ -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{a}\sum_{j=1}^{n} z_{ij}^2 \tag{2.4}$$

Taking derivatives with respect to parameters and equating them to zero, we obtain the likelihood equations as:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu} &= -\frac{\lambda}{\sigma}\sum_{i=1}^{a}\sum_{j=1}^{n}\frac{2(1+\lambda z_{ij})}{(1-\lambda z_{ij})^2+1} + \frac{1}{\sigma}\sum_{i=1}^{a}\sum_{j=1}^{n} z_{ij} = 0 \\
\frac{\partial \ln L}{\partial \alpha_i} &= -\frac{\lambda}{\sigma}\sum_{j=1}^{n}\frac{2(1+\lambda z_{ij})}{(1-\lambda z_{ij})^2+1} + \frac{1}{\sigma}\sum_{j=1}^{n} z_{ij} = 0, i = 1,2,...,a \\
\frac{\partial \ln L}{\partial \sigma} &= -\frac{\lambda}{\sigma}\sum_{i=1}^{a}\sum_{j=1}^{n}\frac{2(1+\lambda z_{ij})z_{ij}}{(1-\lambda z_{ij})^2+1} - \frac{N}{2\sigma} - \frac{1}{\sigma}\sum_{i=1}^{a}\sum_{j=1}^{n} z_{ij}^2 = 0
\end{aligned} \tag{2.5}$$

and

$$\frac{\partial \ln L}{\partial \lambda} = -\sum_{i=1}^{a}\sum_{j=1}^{n}\frac{(1+\lambda z_{ij})z_{ij}}{(1-\lambda z_{ij})^2+1} - N\frac{\lambda}{2+\lambda^2} = 0.$$

The solutions of these likelihood equations are the ML estimators. It can be noticed that these equations have no explicit solutions and they can be solved by using iterative methods.

In one-way ANOVA, our aim is to compare the equality of treatment effects, in other words, to test the following hypothesis $H_0 : \alpha_i = 0, i = 1,2,..,a$ against the alternative hypothesis $H_1 : at\ least\ one\ \alpha_i \neq 0$. In order to compare the group means in one-way ANOVA, the traditional $F$ statistics is used. In this paper, we propose a new $F$ statistics based on maximum likelihood estimators of the ASN distribution obtained by solving equation (2.5) as;

$$F^* = \frac{n\sum_{i=1}^{n}\hat{\alpha}_i^*}{(a-1)\hat{\sigma}^{*2}} \tag{2.6}$$

where $\hat{\alpha}_i^*$ and $\hat{\sigma}^{*2}$ are the ML estimators of ASN distribution.

# 3 Simulation Study

In this section, we compare the ML estimator of the model parameters with respect to traditional LS estimators under the assumption of the ASN distribution. The comparison is made by using relative efficiencies and simulation studies are based on [100,000/n] Monte Carlo runs. We use $a = 3$ and $n = 5, 10$ and $15$. Without loss of generality, we choose the following setting in our simulation: $\mu_i(\mu + \alpha_i = 0(i = 1,2,...,a)$ and $\sigma = 1$.

Table 1 shows that the relative efficiencies of the ML estimators of the $\mu_i = \mu + \alpha_i$ and $\sigma$ with different shape parameter values. It can be seen from the Table 1, under the assumption of the ASN distribution, the ML estimators are more efficient than the corresponding LS estimators as expected. When the sample size and the value of the shape parameter gets higher, the ML estimators become more efficient. As the sake of brevity, we only produce the table for the positive values of the shape parameters. The table remains same for the negative values of $\lambda$.

Table 1: Relative efficiencies of ML estimators.

| | n | $\lambda = 0$ | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 5$ |
|---|---|---|---|---|---|
| $\mu_i$ | 5 | 1.00 | 0.96 | 0.91 | 0. 89 |
| | 10 | 1.00 | 0.93 | 0.85 | 0.78 |
| | 15 | 1.00 | 0.91 | 0.82 | 0.74 |
| $\sigma$ | 5 | 1.00 | 0.98 | 0.96 | 0. 93 |
| | 10 | 1.00 | 0.97 | 0.96 | 0.92 |
| | 15 | 1.00 | 0.95 | 0.93 | 0.91 |

Now, we compare the power values of the proposed test statistics given in (10) and traditional $F$ test statistics under different $\lambda$ values.

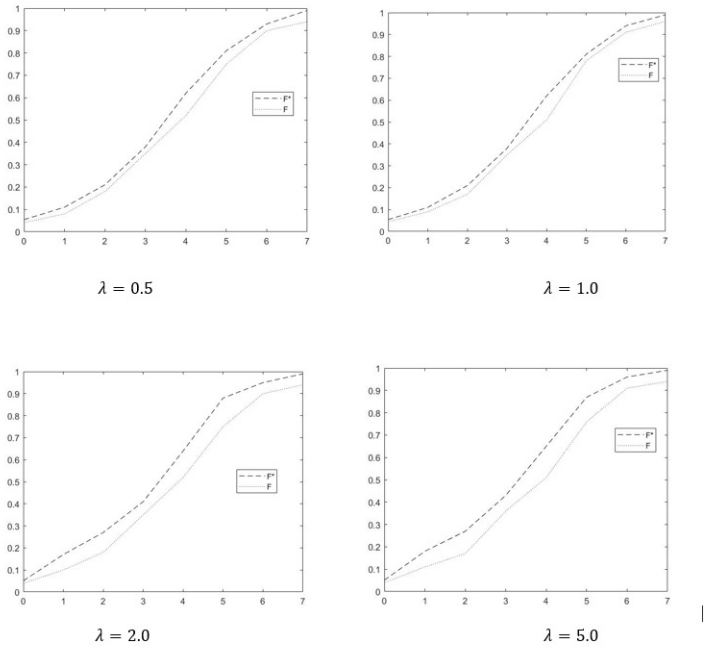


Figure 1: The power values of $F*$ and $F$.

As can be seen from the Figure 1, the proposed test statistics is much more powerful than corresponding traditional test statistics based on LS estimators.

# 4 Numerical Example

In numerical example, we use the Environmental Performance Index (EPI) report available in https://sedac.ciesin.columbia.edu/
The 2020 Environmental Performance Index (EPI) provides a data-driven summary of the state of sustainability around the world. Using 40 performance indicators across 11 issue categories, the EPI

ranks 180 countries on their progress toward improving environmental health, protecting ecosystem vitality, and mitigating climate change. (EPI 2022 report). We take 9 countries in Asia, 10 countries from Europe and 6 countries from America, therefore, we want to learn that whether there is a significant difference between the continents according to EPI points. Figure 1 shows the histogram of the EPI data.

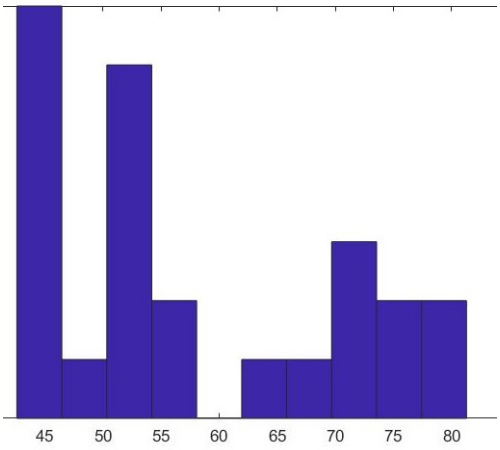


Figure 2: The histogram for EPI data.

It can be realized from the histogram; the data is distributed bimodal as shown in Figure 2. We fit the data to the ASN distribution and obtained the estimator of $\lambda$ as 2.436. Table 2 shows the results of one-way ANOVA for the ASN distribution and traditional normal distribution.

Table 2: Results for one-way ANOVA Analysis.

| | $\mu$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\sigma$ | $F$ | p-value |
|---|---|---|---|---|---|---|---|
| ASN | 56.4 | -4.5 | 7.1 | -5.1 | 11.69 | 3.08 | 0.036 |
| N | 57.8 | -4.2 | 6.9 | -5.2 | 11.84 | 2.83 | 0.081 |

It can be seen from Table 2, we can conclude that there is no significance difference between the countries according to the EPI points if we use the traditional analysis with normal theory, on the other hand, we can conclude that, there is a significance difference between the countries if the proposed methodology is used. The result based on ASN distribution is more reliable with smaller standard deviation.

# 5 Conclusion

In this paper, we assume the distribution of the error terms in one-way ANOVA as ASN distribution. The ASN distribution gives us flexibility for modelling bimodal data. The estimators of the parameters of interest are derived by using ML methodology. The test statistics based on these estimators is also obtained in order to test the equality of the group means. Monte Carlo simulation study states that the ML estimators are much more efficient than the corresponding LS estimators in the assumption of ASN distribution. Additionally, the test statistics based on the ML estimators is more powerful than

the traditional test statistics based on normality. If the data is not bimodal, the ASN distribution may not be helpful, the other alternatives to normal distribution such as Skew normal or skew $t$ distribution.

# References


[1] Celik, N., Senoglu, B., and Arslan, O. (2015)."Estimation and testing in one-way ANOVA when the errors are skew-normal". *Revista Colombiana de Estadistica*, 38 (1), 75–91.

[2] Celik, N., and Senoglu, B. (2018). "Robust estimation and testing in one-way ANOVA for Type II censored samples: skew normal error terms". *Journal of Statistical Computation and Simulation*, 88 (7), 1382-1393.

[3] Celik, N. (2022). "Welch's ANOVA: Heteroskedastic skew-t error terms". *Communications in Statistics - Theory and Methods*, 51 (9), 3065-3071.

[4] Donaldson, T. (1968). "Robustness of the f-test to errors of both kinds and the correlation between the numerator and denominator of the F ratio". *Journal of American Statistical Association*, 63(322), 660– 667.

[5] Elal-Olivero, D., (2010). "Alpha-Skew-Normal Distribution". *Proyecciones Journal of Mathematics* , 29(3), 224-240.

[6] EPI Report., (2022), https://epi.yale.edu/downloads/epi2022report06062022.pdf

[7] Geary, R. (1947)." Testing for normality". *Biometrika*, 34, 209–242.

[8] Senoglu, B., and Tiku, M. L. (2001). "Analysis of variance in experimental design with nonnormal error distributions". *Communications in Statistics – Theory and Methods*, 30 (7):1335–52.

[9] Spjotvoll, E. and Aastveit, H. (1980), "Comparison of robust estimators on some data from field experiments", *Scandinavian Journal of Statistics*, 7, 1–13.

[10] Tan, W.Y., and Tiku, M.L. (1999). *Sampling distributions in terms of Laguerre Polynomials with applications*. New Age International (formerly, Wiley Eastern), New Delhi, tt