Journal of Engineering and Applied Sciences 13 (Special Issue 9): 6939-6943, 2018
ISSN: 1816-949X

Estimators of the Degrees of Freedom of a Chi-Square Distribution Obtained from the Method of Moments Procedure

[1]Abdul Rahman Othman, [1]Chin Ee Laine and [2]Teh Sin Yin
[1]School of Distance Education,
[2]School of Management, Universiti Sains Malaysia, Penang, Malaysia

Abstract: When skewed data is generated, the Chi-square distribution is often used. Skewness in the Chi-square distribution is represented by the degrees of freedom. As the degrees of freedom increase, the skewness becomes less apparent and the Chi-square distribution will approach normality. This study shows the estimation of the degrees of freedom of the Chi-square distribution using the method of moments. The procedure shows $U = \frac{1}{n}\sum_{i=1}^{n} X_i$ as the first sample moment and $V = \frac{1}{2(n-1)}\sum_{i=1}^{n}(X_i - \overline{X})^2$ as the second central moment. The mean square error and consistency for both U and V were evaluated. From this, the first moment estimator U was concluded to be the better of the two estimators.

Key words: Skewed data, degrees of freedom, method of moments, less apparent, distribution, Malaysia

INTRODUCTION

The Chi-square random variable is theoretically defined as the sum of squares of independently distributed standard normal random variables. This distribution has a parameter called the degrees of freedom, which is the sum of the number of standard normal random variables.

Consequently, a Chi-square distribution with f degrees of freedom will have a mean value f and a variance 2f where f is an integer. The Chi-square distribution is positively skewed and has a lower bound of zero. As the degrees of freedom increase, the skewness of the distribution becomes less obvious and the Ch-square distribution will approach normality.

The degrees of freedom of the Chi-square distribution are estimated using a point estimation method known as the method of moments. The methods of moments procedure will produce multiple estimators of the degrees of freedom. However, the best estimator is obtained or chosen using the standard criteria of evaluating estimators.

Literature review: The Chi-square distribution is often used in the generation of skewed data for the study of robustness of statistical tests, especially, the testing of grouped means and grouped variances. For example, Keselman *et al*. (2002) and Hogg *et al*. (2012) generated pseudo-random variates of the Chi-square distribution with three degrees of freedom as part of their study on how the Type 1 error of 56 test statistics were controlled under conditions of extreme heterogeneity and nonnormality.

Following that, Keselman *et al*. (2004, 2007) also, used the Chi-square distribution with three degrees of freedom as skewed data. The earlier work was on the power characteristics of seven adaptive test statistics comparing group means. Six were modified from the Welch-James statistic and one modified from the one-step M-estimator. The latter research examined the effect of different methods of trimming on the robustness of the Welch-James statistic modified with the trimmed means and Winsorized variances from those trimming methods.

As skewness is represented by the degrees of freedom and the degrees of freedom itself is a parameter of the Chi-square it can therefore be estimated. However, this was not considered in any statistical test that involves the Chi-square distribution like the basic F-test in Analysis of Variance (ANOVA) (Montgomery, 2008).

In general, the process of estimating a single parameter value is called point estimation. According to Lehmann and Casella (2003) there are several methods of obtaining point estimators such as the Minimum-Variance Mean-Unbiased Estimator (MVUE), Minimum Mean Squared Error (MMSE), Maximum Likelihood Estimator (MLE) and the Method of Moments estimator (MOM).

The two common methods are the MLE and the MOM, the latter will be the focus in this study. The

Corresponding Author: Chin Ee Laine, School of Distance Education, Universiti Sains Malaysia, Penang, Malaysia

MOM is chosen because of its less-complicated computational steps than that of what the MLE incurs.

In estimation it is important that, the estimator corresponds to the parameter of interest and the way to determine this is to evaluate the properties of the estimator. Three key properties required of a good estimator are unbiasedness, consistency and efficiency.

An estimator is unbiased when the expected value of the estimator is equal to the parameter that is being estimated. It should also be estimating the parameter of interest and not any other parameter. Lastly, it should possess minimum variance, smaller than any other estimators of that parameter and at most equal to the theoretical Rao-Cramer Lower Bound (RCLB) (Casella and Berger, 2002).

MATERIALS AND METHODS

The approach taken for this study is to first obtain the estimators of the degrees of freedom, analytically. Upon obtaining the two estimators, the three important qualities of estimators are worked out. These analytical results are then validated using a simulation of Monte Carlo runs.

Maximum-likelihood estimation: As mentioned in this study, the two most common methods for estimation are the MLE and the MOM. These two methods are used to estimate the parameter of the Chi-square distribution which is the degrees of freedom. While the MLE is the most popular method for estimating a distribution, the analytical calculation for the estimators following this method involves the differentiation of the parameter interest in factorial form. This is not a close form function and solving for the estimator is not possible. Therefore, the MOM estimator is used.

Method of moments estimation: If X is Chi-square distributed with degrees of freedom then the probability density function (pdf) of the Chi-square distribution is:

$$f(x)=\frac{x^{\frac{v-2}{2}}\exp\left(-\frac{x}{2}\right)}{\iint 2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)},\ 0\le x<\infty,\ v\in N^{*} \tag{1}$$

where variable $x\ge 0$, parameter v is a positive integer and $\Gamma(v/2)$ denotes the Gamma function. The expected value of X, E(X) is v and the variance, Var(X) is 2v (Forbes *et al*., 2011).

Let, $X_1, X_2, \ldots, X_n$ be a random sample of size n obtained f rom a population distributed as Chi-square with degrees of freedom. The population mean of X is E(X) = v and the sample mean or the first moment is given by:

$$M_1=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{2}$$

Hence, the first MOM estimator of can be set as:

$$U=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{3}$$

The population variance of X is Var (X) = 2v and the second theoretical moment about the mean or better known as the variance is:

$$M_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{4}$$

Therefore, the second MOM estimator v of can be set as:

$$V=\frac{1}{2(n-1)}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{5}$$

Properties of estimators: Upon obtaining the estimators, U and V it is important to know if they have the qualities of a good estimator.

Unbiasedness: For this property, a good estimator is unbiased if the expected value of the estimator is equal to the degrees of freedom. Therefore, the expected value of U is:

$$\begin{aligned}E(U)&=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\\&=\frac{1}{n}\sum_{i=1}^{n}E(X_i)\\&=\frac{1}{n}(nv)\\&=v\end{aligned} \tag{6}$$

As a result, the first moment estimator, U is an unbiased estimator of the degrees of freedom, v. Following that, the expected value of V is given by:

$$\begin{aligned}E(V)&=E\left[\frac{1}{2(n-1)}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]\\&=\frac{1}{2(n-1)}E\left[\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right]\\&=\frac{1}{2(n-1)}\left\{\sum_{i=1}^{n}E\left[X_i^2\right]-nE[\overline{X}^2]\right\}\end{aligned} \tag{7}$$

The formula to evaluate variance in terms of expected values is given by:

$$\mathrm{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 \quad (8)$$

Hence:

$$E(X_i^2) = \mathrm{Var}(X_i) + [E(X_i)]^2 = 2v + v^2 \quad (9)$$

Now to obtain the expected value for mean square, since $X \sim \chi^2_v$, therefore, $\sum_{i=1}^{n} X_i \sim \chi^2_{nv}$:

$$\begin{aligned} \mathrm{Var}(n\bar{X}) &= 2nv \\ n^2\left\{E(\bar{X}^2) - [E(\bar{X})]^2\right\} &= 2nv \\ n^2\left\{E(\bar{X}^2) - v^2\right\} &= 2nv \\ E(\bar{X}^2) &= \frac{2nv}{n^2} + v^2 \end{aligned} \quad (10)$$

Substituting Equ. 9 and 10 into Eq. 7:

$$\begin{aligned} E(V) &= \frac{1}{2(n-1)}\left\{n(2v+v^2) - n\left(\frac{2nv}{n^2} + v^2\right)\right\} \\ &= \frac{1}{2(n-1)}\left\{2nv + nv^2 - 2v - nv^2\right\} \\ &= \frac{1}{2(n-1)}\left\{2nv - 2v\right\} \\ &= \frac{2v(n-1)}{2(n-1)} \\ &= v \end{aligned} \quad (11)$$

Hence, the second moment estimator, V is also, an unbiased estimator of the degrees of freedom, v.

Consistency: Another desirable property of estimators is consistency. An estimator $\hat{v}$ is called consistent if $\lim_{n\to\infty} MSE(\hat{v}) = 0$ which means that as the number of observations increase, the Mean Squared Error (MSE) tends to 0. Therefore, the mean squared error for estimator U is:

$$MSE(U) = E\left[(U-v)^2\right] = \mathrm{Var}(U) + (E(U)-v)^2 \quad (12)$$

Since, U is unbiased, the bias term of is equal to zero hence, MSE(U) = Var(U). Now, consider $n\bar{X} \sim \chi^2_{nv}$:

$$\begin{aligned} \mathrm{Var}(n\bar{X}) &= 2nv \\ n^2\mathrm{Var}(\bar{X}) &= 2nv \\ \mathrm{Var}(\bar{X}) &= \frac{2nv}{n^2} \end{aligned} \quad (13)$$

Hence, Var(U) = 2V/n. Then $\lim_{n\to\infty} MSE(U) = \lim_{n\to\infty}\left(\frac{2v}{n}\right) = 0$.

Therefore, U is a consistent estimator of the degrees of freedom, v. Since V is also, unbiased, therefore, MSE(V) = Var(V). To obtain the variance of V:

$$\begin{aligned} \mathrm{Var}(V) &= \frac{1}{4(n-1)^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] \\ &= \frac{1}{4(n-1)^2}\left\{\begin{array}{l} \mathrm{Var}\left[\sum_{i=1}^{n} X_i^2\right] + n^2\mathrm{Var}\left[\bar{X}^2\right] \\ -2\mathrm{Cov}\left[\sum_{i=1}^{n} X_i^2,\ n\bar{X}^2\right] \end{array}\right\} \end{aligned} \quad (14)$$

Upon further calculations, Var(V) does not approach zero as the sample size increases, hence, V is not a consistent estimator.

Efficiency: Lastly, the third essential quality of an estimator is its efficiency. An unbiased estimator is preferable as it is able to constantly produce estimates which have high precision. On the contrary an estimator that produces widely varying estimates on different occasions is probably one to avoid. With these said, the variance of an estimator should be as small as possible to be considered being efficient and this is achieved if and only if the variance of the unbiased estimator attains the RCLB. An estimator whose variance is equal to the RCLB is called a most efficient estimator (Hogg *et al*., 2012).

To define the RCLB for this research, let $X_1, X_2, \ldots, X_n$ be independent and identically distributed Chi-square with v degrees of freedom random variables and with probability density function f(x, y) for $v \in N^*$ as previously shown. The two estimators U and V are both unbiased estimators of the population parameter, V . In order for them to be efficient estimators their variances should be equal to the RCLB given by 1/nI(v) where I(V) is the Fisher information defined as Hogg *et al.* (2012):

$$I(v) = E\left[\left(\frac{\partial \log f(x; v)}{\partial v}\right)^2\right] \quad (15)$$

However, the computations of the density function for the Chi-square with v degrees of freedom distribution

involved differentiation of v factorial form. This is not a closed form and therefore, the RCLB cannot be computed. Although, it cannot be determined whether both estimators are efficient it can be deduced which is the more efficient estimator relatively. In this particular case, based upon Eq. 13 and 14, U is more efficient than V.

Validation: The theoretical findings in this study were validated by using the Monte Carlo simulation techniques reviewed and elaborated by Othman. The steps for this validation were reproduced from the Monte Carlo study in estimation.

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from the Chi-square distribution with four degrees of freedom. The sample size of was chosen. Since, the degrees of freedom is the known population parameter then all of the values of U and V generated in the validation stage should be close to this value. This becomes the target value to check against.

The steps involved in this validation are as follows: (steps 1-7 are realized in a SAS/IML program):

- Generate a sample size of 25 from the Chi-square distribution with four degrees of freedom
- Calculate U and V using the data from step 1
- Repeat steps 1 and 2 for 1000 runs
- From 1000 estimates of U and V, the mean values of the 1000 estimates for both U and V are obtained
- The mean values in step 4 are checked for bias by comparing them against the target of v = 4 or rather the differences between each of them against the target value are obtained
- The standard deviation values of the 1000 U and V values are then obtained
- From steps 5 and 6, the mean squared error values of U and V are then calculated

RESULTS AND DISCUSSION

The data shown in Table 1 are the properties of the 1000 Monte Carlo estimates. The first MOM estimator, U, has an estimate of 4.006 whereas the second estimator, V, has an estimate of 3.808. With the target value of hence, the mean biases of the 1000 estimates for U and V are 0.006 and -0.192, respectively for the two estimators. Therefore, U is the better unbiased estimator compared to V. This validates the analytical results.

Both estimators are theoretically unbiased, hence, the bias term is zero. Therefore, the MSE of U and V are equal to their variances. The MSE of 1000 estimates of U is 0.318 and for V, the value is 2.894. Theoretically, U converges to zero when the sample size increases whereas V does not. This is validated here with the MSE of U being less than that of V. Therefore, U is a more consistent estimator compared to V.

Table 1: Properties of monte carlo estimates of U and V

Variables	U	V
Estimate	4.0064510	3.8080288
Bias	0.0064510	-0.1919710
SD	0.5638797	1.6902075
Mean-square-error	0.3180020	2.8936545

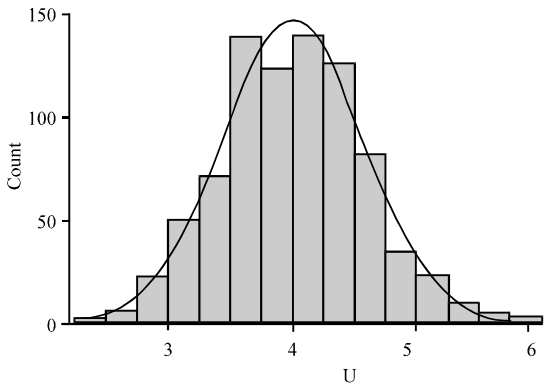

Fig. 1: Histogram for 1000 U estimates

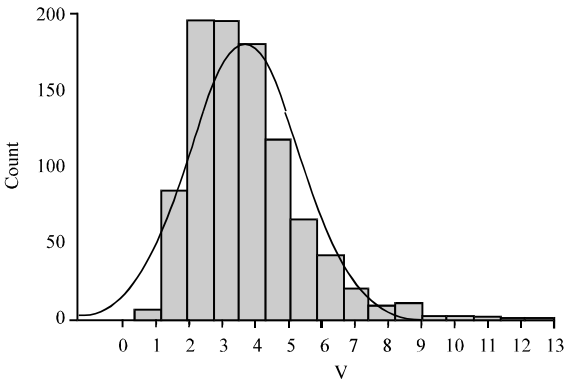

Fig. 2: Histogram for 1000 V estimates

The standard deviation of the 1000 estimates of U is 0.564 and for V, the value is 1.690 and this is visualized in Fig. 1 and 2. Although, the efficiency of both U and V estimators cannot be determined theoretically as it involves the steps to find the RCLB it can be deduced that U is relatively more efficient than V, theoretically. This is validated here with U having a smaller standard deviation compared to V.

CONCLUSION

This study shows that the degrees of freedom of the Chi-square distribution can be estimated using a known point estimation method which is the method of moments. However, the method of moments produces multiple estimators. As the order of moments increases, the

function of the random sample becomes complex with polynomials. This results in the complexity of computing the estimates as well as evaluating the properties of the estimators.

In this study, the research done was stopped at the second order of moments. From the acquired results, it can be said that statistic U, a function of the first moment is the better estimator of the degrees of freedom, than V which is a function of the second moment. Although, both estimators are unbiased, U is consistent while V is not. The efficiency of both estimators are not known but relatively U is more efficient than V.

ACKNOWLEDGEMENT

This research is supported by the Universiti Sains Malaysia Fundamental Research Grant Scheme (FRGS), No. 203/PMGT/6711345.

REFERENCES

Casella, G. and R.L. Berger, 2002. Statistical Inference. 2nd Edn., Cengage Learning, Boston, Massachusetts, ISBN:9780534243128, Pages: 660.

Forbes, C., M. Evans, N. Hastings and B. Peacock, 2011. Statistical Distributions. 4th Edn., John Wiley and Sons, Hoboken, New Jersey, ISBN:978-0-470-39063-4,.

Hogg, R.V., J. McKean and A.T. Craig, 2012. Introduction to Mathematical Statistics. 7th Edn., Pearson Education, Upper Saddle River, New Jersey, ISBN: 9780321795434, Pages: 694.

Keselman, H.J., R.R. Wilcox, J. Algina and A.R. Othman, 2004. A power comparison of robust test statistics based on adaptive estimators . J. Mod. Appl. Stat. Methods, 3: 27-38.

Keselman, H.J., R.R. Wilcox, L.M. Lix, J. Algina and K. Fradette, 2007. Adaptive robust estimation and testing. Br. J . Math. Stat. Psychol., 60: 267-293.

Keselman, H.J., R.R. Wilcox, P. Othman, M.A. Rahman and K. Fradette, 2002. Trimming, transforming statistics and bootstrapping: Circumventing the biasing effects of heterescedasticity and nonnormality. J. Mod. Appl. Stat. Methods, 1: 288-309.

Lehmann, E.L. and G. Casella, 2003. Theory of Point Estimation. 2nd Edn., Springer, New York, USA., ISBN:9780387985022, Pages: 590.

Montgomery, D.C., 2008. Design and Analysis of Experiments. John Wiley and Sons, Hoboken, New Jersey, ISBN:978-0-470-12866-4, Pages: 487.