ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Applications of Information Measures to Assess Convergence in the Central Limit Theorem

Ranjani Atukorala, Maxwell L. King and Sivagowry Srianantha kumar

December 2014

Working Paper 29/14

APPLICATIONS OF INFORMATION MEASURES TO ASSESS CONVERGENCE IN THE CENTRAL LIMIT THEOREM[1]

RANJANI ATUKORALA
Statistics & Reporting Unit
RMIT University,

MAXWELL L. KING
Department of Econometrics & Business Statistics
Monash University

SIVAGOWRY SRIANANTHAKUMAR*
School of Economics, Finance and Marketing
RMIT University,
GPO Box 2476 Melbourne Australia 3001
Phone: +6139925 1456 Fax: +6139925 5986,
Email: sivagowry.srianantha kumar@rmit.edu.au

Abstract

The Central Limit Theorem (CLT) is an important result in statistics and econometrics and econometricians often rely on the CLT for inference in practice. Even though different conditions apply to different kinds of data, the CLT results are believed to be generally available for a range of situations. This paper illustrates the use of the Kullback-Leibler Information (KLI) measure to assess how close an approximating distribution is to a true distribution in the context of investigating how different population distributions affect convergence in the CLT. For this purpose, three different non-parametric methods for estimating the KLI are proposed and investigated. The main findings of this paper are 1) the distribution of the sample means better approximates the normal distribution as the sample size increases, as expected, 2) for any fixed sample size, the distribution of means of samples from skewed distributions converges faster to the normal distribution as the kurtosis increases, 3) at least in the range of values of kurtosis considered, the distribution of means of small samples generated from symmetric distributions is well approximated by the normal distribution, and 4) among the nonparametric methods used, Vasicek's (1976) estimator seems to be the best for the purpose of assessing asymptotic approximations. Based on the results of this paper, recommendations on minimum sample sizes required for an accurate normal approximation of the true distribution of sample means are made.

Keywords: *Kullback-Leibler Information, Central Limit Theorem, skewness and kurtosis*

***JEL* codes:** C1, C2, C4, C5

*Corresponding author

[1] We would like to thank two anonymous referees and Professor Farshid Vahid for their helpful comments.

1. INTRODUCTION

A large part of asymptotic theory is based on the CLT. However, convergence in the CLT is not uniform in the underlying distribution. There are some distributions for which the normal approximation to the distribution can be very poor. We can improve on the normal approximation using higher order approximations but that does not always provide good results. When higher order terms in the expansion involve unknown parameters, the use of estimates for these parameters can sometimes worsen the approximation error rather than improve it (Rothenberg, 1984).

From time to time, researchers point out problems associated with the CLT. In contrast to textbook advice, the rate at which a sampling distribution of means converges to a normal distribution depends not only on sample size but also on the shape of the underlying population distribution. The CLT tends to work well when sampling from distributions with little skew, light tails and no outliers (Little, 2013; Wilcox, 2003; Wu, 2002). Wu (2002) in the psychological research context, discovered that sample sizes in excess of 260 can be necessary for a distribution of sample means to resemble a normal distribution when the population distribution is non-normal and samples are likely to contain outliers. Smith and Wells (2006) conducted a simulation study to generate sampling distributions of the mean from realistic non-normal parent distributions for a range of sample sizes in order to determine when the distribution of the sample mean is approximately normal. Their findings suggest that as the skewness and kurtosis of a distribution increase, the CLT will need sample sizes of up to 300 to provide accurate inference. Other studies revealed that standard tests such as z, t and F, can suffer from very inflated rates of Type 1 error when sampling from skewed distributions even when the sample sizes are as high as 100 (Bradley, 1980; Ott and Longnecker, 2010). Wilcox (2005) observed that the normal approximation's quality cannot be ensured for highly skewed distributions in the context of calculating confidence intervals using the normal quantiles even in very moderate sized samples (e.g. 30 or 50). Shilane et al. (2010) established that the normal confidence interval significantly under-covers the mean at moderate sample sizes and suggested alternative estimators based upon gamma and chi square approximations along with tail probability bounds such as Bernstein's inequality. Shilane and Bean (2013) proposed another method, namely the growth estimator, which provides improved confidence intervals for the mean of negative binomial random variables with unknown dispersion. They observed that their growth estimator produces intervals that are longer and more variable than the normal approximation. In the censored data context, Hong et al. (2008) pointed out that the normal approximation to confidence interval calculations can be poor when the sample size is not large or there is heavy censoring. In the context of approximation of the binomial distribution, Chang et al. (2008) made similar observations.

Econometric textbooks loosely define the CLT as the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution. The question is how 'large' the sample size should be for the normal distribution to provide a good approximation. Also which distribution from a class of distributions, causes the slowest convergence in the CLT. These are the important questions this paper seeks answers to using the KLI measure. In particular, the KLI is used to find which sample sizes are reasonably good for the normal distribution to be an accurate approximation to the true distribution of the sample mean. To do so, we use the KLI of the density functions for true distributions of means of a sequence of random samples with respect to the asymptotic normal distribution.

Using simulation methods, we generate random samples from a range of different underlying population distributions and calculate their sample means. In particular, Tukey's Lambda distribution is used for generating random numbers with known skewness and kurosis. We also find the maximum value of the KLI among a range of distributions in order to investigate the slowest convergence in the CLT. For convenience, we use the Lindeberg-Levy CLT which is the simplest and applies to independent and identically distributed random observations. Only one dimensional variables are considered for convenience. The estimated KLI numbers are used to study how 'large' the sample size should be to have an accurate normal approximation. We also try to find which distributions give poor normal approximations for a particular fixed sample size using this concept.

In summary, this paper investigates four important issues; 1) how large the sample size should be for the normal distribution to provide a good approximation to the distribution of the sample mean, 2) which distribution from a class of distributions, causes the slowest convergence in the CLT, 3) which distributions give poor normal approximations for particular fixed sample sizes and 4) of the nonparametric methods used, which seems to be the best for the purpose of assessing asymptotic approximations.

The rest of the paper is planned as follows. Section 2 outlines the theory and the details of estimating the KLI. The design of the Monte Carlo experiments including the data generation process is discussed in Section 3. Section 4 reports the Monte Carlo results. Some concluding remarks are made in Section 5.

2. THE THEORY

2.1 Generating observations from Tukey's Lambda (λ) distribution

Our simulation experiments used random drawings from a generalisation of Tukey's λ distribution proposed by Ramberg and Schmeiser (1972, 1974). The distribution is defined by the percentile function (the inverse of the distribution function)

$$R(p) = \lambda_1 + \frac{\left[p^{\lambda_3} - (1-p)^{\lambda_4} \right]}{\lambda_2}, \quad 0 \leq p \leq 1, \tag{1}$$

where p is the percentile value, λ_1 is a location parameter, λ_2 is a scale parameter and λ_3 and λ_4 are shape parameters. It has the advantage that random drawings from this distribution can be made using (1), where p is now a random drawing from the uniform distribution on the unit interval. The density function corresponding to (1) is given by

$$f(z) = f\left[R(p)\right] = \lambda_2 \left[\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1} \right]^{-1} \tag{2}$$

and can be plotted by substituting values of p in (1) to get $z = R(p)$ and then substituting the same values of p in (2) to get the corresponding $f(z)$ values. Ramberg et al. (1979) discuss this distribution and its potential use in some detail. They also give tables that allow one to choose λ_1, λ_2, λ_3 and λ_4 values that correspond to particular skewness and kurtosis values when the mean is zero and the variance is one[2]. Therefore by an appropriate choice of skewness and kurtosis values, a number of distributions can be approximated by a distribution that has the same first four moments. These include the uniform, normal, Weibull, beta, gamma, log-normal and Student's t distributions. For examples of the use of this distribution in econometric simulation studies see Evans (1992), Brooks and King (1994) and King and Harris (1995).

2.2 Estimation of KLI

In order to evaluate the quality of an approximating distribution, we need a convenient way to measure divergence between distributions. One such tool is the KLI measure, introduced by Kullback and Leibler

[2] The simultaneous equations (for any mean, variance, skewness and kurtosis values) which can be solved to obtain the corresponding λ values are also given by Ramberg et al. (1979).

(1951). Let $g(x)$ be the true density function of a $q\times 1$ random vector x and $f_a(x)$ be an approximating density for x. The KLI measure is defined as:

$$I(g; f_a) = \mathrm{E}\left[\log\{g(x)/ f_a(x)\}\right] = \int_{R^q} \log\left\{g(x)/ f_a(x)\right\}g(x)dx$$

$$= \int_R g(x)\log\left\{g(x)\right\}dx - \int_R g(x)\log\left\{ f_a(x)\right\}dx. \qquad (3)$$

Its usefulness as a measure of the quality of approximation comes from the following properties

1 $I(g; f_a) \geq 0$ for all g and f_a.

2 $I(g; f_a) = 0$ if and only if $g(x) = f_a(x)$ almost everywhere.

As observed by Renyi (1961, 1970), the KLI measure can be interpreted as the surprise experienced on average when we believe $f_a(x)$ is the true underlying distribution and we are told it is in fact $g(x)$. The smaller the value of $I(g; f_a)$ the less the surprise, and the closer we consider the approximating distribution $f_a(x)$ to be to the true distribution $g(x)$. Also note that $I(g; f_a)$ is the expected value of the log of the likelihood ratio which, according to the Neyman-Pearson Lemma, provides the best test of $\mathrm{H}_0: x \sim g(x)$ against $\mathrm{H}_1: x \sim f_a(x)$.

Let $x_1, x_2, ..., x_m$ be a simulated iid random sample in which x_i, $i = 1,..,m$, is an $n\times 1$ vector from either H_0 or H_1, then the most powerful test can be based on rejecting H_0 for small values of

$$\frac{1}{m}\sum_{i=1}^{m}\log\left\{g(x_i)/ f_a(x_i)\right\} \qquad (4)$$

which is the standard estimate of

$$I(g; f_a) = \mathrm{E}\left\{\log((g(x)/ f_a(x))\right\} \qquad (5)$$

from a simple random sample of size m. In this sense we feel confident in using $I(g; f_a)$ as a measure of distance between $g(x)$ and $f_a(x)$. For further discussion of the KLI measure, see Kullback (1959), Renyi (1961, 1970), Vuong (1989), Maasoumi (1993) and White (1982, 1994).

Our aim is to estimate (3) where $g(x)$ not known but a simple random sample of observations from g can be taken. The negative value of the first term of (3),

$$H(g) = -\int_{-\infty}^{\infty} g(x)\log\{g(x)\}dx \qquad (6)$$

is the continuous version of the entropy of the probability density function $g(x)$[3]. When the distribution $g(x)$ is known, it is obvious that the KLI measure can be easily estimated via the estimation of the entropy for the known distribution. But when the true distribution of $g(x)$ is unknown, nonparametric estimation methods are needed to estimate the unknown true distribution or the entropy of the unknown true distribution. A number of nonparametric techniques are available for estimating the entropy of the true distribution, however, we use the Vasicek's (1976) estimator because of its reliability (Atukorala, 1999; Guo et al., 2010), the kernel estimator because of its popularity and simplicity and the Maximum Entrophy (ME) principle because of its popularity.

2.2.1 The use of kernel density estimation (hereafter referred to as M1)

The kernel estimator is the most commonly used density estimator. Even though this method is not the best to use in all circumstances, it is widely used particularly in the univariate case. We use this method in estimating the true density function g in equation (3). A nonparametric estimator of the Shannon entropy defined as in (6) for an absolutely continuous distribution g, is given by

$$\hat{H}_k(g) = -\frac{1}{m}\sum_{i=1}^{m}\log\{\hat{g}(x_i)\}, \qquad (7)$$

where $x_1, x_2, \ldots, x_m$ is a random sample generated from g and $\hat{g}(x)$ is the kernel estimate of g (Rosenblatt, 1956; Parzen, 1962; Ahmad and Lin, 1976; Rao, 1973). Accordingly, an estimator for the first term in (3) is,

$$\hat{I}_T(g) = \frac{1}{m}\sum_{i=1}^{m}\log\hat{g}(x_i). \qquad (8)$$

in which $\hat{g}(x)$ can be calculated as

$$\hat{g}(x) = \frac{1}{mh}\sum_{j=1}^{m}k\left[\frac{(x-x_j)}{h}\right]. \qquad (9)$$

[3] The entropy measure is nonparametric since it needs not assume the probability distribution is in any parametric form.

Thus the estimation amounts to drawing a simple random sample, estimating $\hat{g}(x)$ using this sample and then taking a second sample to calculate $\hat{I}_T(g)$. The kernel density function $k(\cdot)$ and the smoothing parameter h have to be chosen appropriately. The choice of the kernel does not seem very important to the statistical performance of the estimation method. That is, the shape of the kernel does not significantly influence the final shape of the estimated density because it just determines the local behaviour (Bolance et al., 2012). Therefore, in our study, we use the standard Gaussian density for $k(\cdot)$. For the normal kernel, our best choice of the smoothing parameter is[4]

$$h = 1.06\hat{\sigma}m^{-1/5}, \tag{10}$$

where $\hat{\sigma}$ is the standard error of the observed data and m is the number of observations in the data set. Then the KLI can be estimated as

$$\hat{I}_k = \frac{1}{m}\sum_{i=1}^{m}\log \hat{g}(x_i) - \frac{1}{m}\sum_{i=1}^{m}\log f_N(x_i) \tag{11}$$

In (11), $\hat{g}$ is the estimated density function of the true distribution of means,

$$x_i = \frac{1}{n}\sum_{j=1}^{n} z_j, \quad i = 1, 2, \ldots, m, \tag{12}$$

where n is the size of the samples generated from Tukey's λ distribution, for calculating means as explained in Section 2.1 and f_N is the normal density function with zero mean and variance $\frac{1}{n}$.

We also calculated the standard errors of estimated KLI using the square root of the statistic,

$$\text{var}(\hat{I}_k) = \frac{1}{m}\sum_{i=1}^{m}\left[\log\left\{\frac{\hat{g}(x_i)}{f_N(x_i)}\right\} - \hat{I}_k\right]^2. \tag{13}$$

2.2.2 The use of the Maximum Entropy (ME) distribution (hereafter referred to as M2)

Suppose we have a simple random sample of observations from an unknown continuous distribution with range $(-\infty, \infty)$; say $x_1, x_2, \ldots, x_m$. In the ME approach, the objective is to exploit the knowledge that the parent distribution is continuous in constructing an estimated density function, written $h(.)$. This

[4] See Silverman (1978, 1986).

function is derived by maximising its entropy subject to certain constraints. Those constraints reflect the knowledge of the parent distribution provided by the sample.

Calculating the univariate ME distribution amounts to ordering the sample observations

$$x^1 < x^2 < \ldots < x^m .$$

As given by Theil and Fiebig (1984), the two constraints called (i) the mass-preserving constraint and (ii) the mean-preserving constraint have to be imposed in order to calculate the univariate ME distribution. Then, the intermediate points between successive order statistics need to be defined as,

$$\xi_i = \xi(x^i, x^{i+1}),\ i = 1, \ldots, m-1, \tag{14}$$

where $\xi(\cdot)$ is a symmetric differentiable function of its two arguments whose values are not outside the range defined by these arguments. The ME density function (Theil and Fiebig, 1984) is as follows:

$$f_i(x) = \frac{2}{x^{i+1} - x^{i-1}}, \qquad \text{for } \xi_{i-1} < x \le \xi_i, \tag{15}$$

$$f_m(x) = \frac{4}{x^m - x^{m-1}} \exp\left\{ -\frac{x - \frac{1}{2}\left(x^{m-1} + x^m\right)}{\frac{1}{4}\left(x^m - x^{m-1}\right)} \right\}, \qquad \text{for } x \ge \xi_{m-1}, \tag{16}$$

$$f_1(x) = \frac{4}{x^2 - x^1} \exp\left\{ \frac{x - \frac{1}{2}\left(x^1 + x^2\right)}{\frac{1}{4}\left(x^2 - x^1\right)} \right\}, \qquad \text{for } x \le \xi_1 . \tag{17}$$

The ME distribution is obtained by maximising the entropy and the value of that maximum is called the maximum entropy. The value of the maximum entropy is

$$H_{ME} = \frac{2}{m}(1-\log 2) + \frac{1}{m}\sum_{i=1}^{m}\log\left\{\frac{m}{2}\left(x^{i+1}-x^{i-1}\right)\right\}. \qquad (18)$$

The first term, $\frac{2}{m}(1-\log 2) \approx \frac{0.6137}{m}$, which is called an end-term correction, results from the exponential tails. In this paper, we use (18) to estimate the entropy of the true density function involved in (3)[5]. This amounts to

$$\hat{I}_{ME} = \frac{-1}{m}\sum_{i=1}^{m}\left[\log\left\{\frac{m}{2}\left(x^{i+1}-x^{i-1}\right)\right\}+\log f_N(x^i)\right] - \frac{2(1-\log 2)}{m}, \qquad (19)$$

where f_N is the normal density function with mean zero and variance $\frac{1}{n}$.

2.2.3 The use of the Vasicek's entropy (hereafter referred to as M3)

When our sample observations are rearranged in the form of order statistics as given by $x^1 < x^2 < \ldots < x^m$, the entropy estimate introduced by Vasicek (1976) can be written as

$$H_v(m_1,m) = \frac{1}{m}\sum_{i=1}^{m}\log\left\{\frac{m}{2m_1}\left(x^{i+m_1}-x^{i-m_1}\right)\right\}, \qquad (20)$$

where m_1 is a positive integer smaller than $m/2$. If the variance of the underlying distribution is finite, $H_v(m_1,m)$ converges to $H(g)$ in (6) as $m \to \infty$, $m_1 \to \infty$ and $m_1/m \to 0$.

When we use Vasicek's method for estimating KLI, we replace the first integral of (3) with minus the estimate $H_v(m_1,m)$ given by (20). Then the estimate for the KLI can be given as

$$\hat{I}_v = -\frac{1}{m}\sum_{i=1}^{m}\left[\log\left\{\frac{m}{2m_1}\left(x^{i+m_1}-x^{i-m_1}\right)\right\}\right]-\frac{1}{m}\sum_{i=1}^{m}\log f_N(x^i) \qquad (21)$$

[5] Note that the second term of (18) is identical to Vasicek's (1976) sample entropy which is used to test for normality.

In (21), an appropriate value for m_1 has to be chosen. Our approach to choosing m_1 is explained in the next section.

Because the theoretical variance for these estimators given in (19) and (21) are complicated and difficult to derive, we use the nonparametric bootstrap method for estimating the standard errors for these cases (Efron, 1979)[6]. 250 bootstrap samples were used in our experiments.

3. MONTE CARLO EXPERIMENT

As explained in Section 2.1, data is generated from a generalisation of Tukey's Lambda distribution with μ = 0 and σ^2 = 1. The following grid points for the skewness, kurtosis and n values were used in the Monte Carlo experiments.

Skewness: 0, 0.25, 0.5, 0.75, 1.0, 1.2, 1.5, 1.7 and 2.0.
Kurtosis: 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 and 15.
Sample size (n): 3, 4, 5, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, and 34.

In cases of skewness values of 0, 0.25, 0.75, 1.2, 1.5, 1.7 and 2.0 with kurtosis values of 6, 7, 8, 9, 10 and 11, sample sizes of 3, 5, 10, 15, 20, 25, 30, 35 and 40 were used.

The grid of skewness and kurtosis values given above, is sufficient for our purposes because those combinations of skewness, kurtosis and sample sizes cover a wide range of values.

By generating random numbers for each combination of these parameters, we calculate m means of each sample. To determine the value of m, we estimated our final estimator given in (11) for a range of different values for m until our estimated values are stable. The values of m range from 2,000 to 22,000 in steps of 2,000 in this case. In the case of M1, the density functions for true distributions of means were estimated using these m values under the different combinations of parameters given above.

Estimates of KLI decline as the value of m increases for all methods used in the experiments. For different m values between 18,000 and 22,000, there was not much difference between the KLI

[6] Researchers find that bootstrap standard errors perform better than the conventional asymptotic standard errors in the linear regression context (Goncalves and White, 2005).

estimates. For the M1 estimator, even from the point of view of density estimation, this range seems to be reasonably good for m to use because it gives smooth density functions for most of the parameter combinations. We found m = 20,000 is a reasonably good number to use in Monte Carlo experiments by considering both the density functions and the KLI estimates. For M3, in addition to the selection of m, an appropriate value for m_1 in equation (21) has to be chosen.

We know that the true distributions of means of independent random samples taken from the standard normal distribution is normal, so there is no approximation error and the true value of KLI is zero. Therefore, estimates of KLI in this case with respect to the approximate normal distribution should be near zero, typically insignificantly different from zero. Thus, the case of generating sample means using a distribution with the same first four moments as the standard normal distribution as the underlying population distribution can be considered as a benchmark for comparing results and choosing an optimal m_1 value. Table 1 lists some of the results for estimated KLI for different values of m_1 when sample sizes are 3, 5 and 10. Most of the estimates for high m_1 values such as 70 to 100 are within two standard errors of zero. Based on these results, we selected m_1 = 85 as the best value to use in our experiments for estimating KLI.

4. MONTE CARLO RESULTS

Selected estimated results for the KLI obtained using the three methods, namely, M1, M2 and M3, for different sample sizes (n) and different skewness and kurtosis values are given in Tables 2 to 5. First, we shall consider the results in the case of generating random observations from Tukey's Lambda distribution with the same first four moments as the normal distribution as they provide a bench mark for the comparison of Monte Carlo results for different methods. The corresponding results are given in the first column of Table 2. The KLI estimates obtained using M1 and M3, are very small and for all the sample sizes are not significantly different from zero. Compared to M2, they also have low standard errors. This implies that the normal distribution approximates the true distribution of the sample means extremely well, as expected.

M2 gives estimates much higher and much more variable than the other two methods. Almost all these estimated values are significantly different from zero even in the case of the underlying population distribution's skewness being zero and kurtosis being three which is the case for data generation from the normal distribution (see Table 2). These results imply that divergence between true distributions of sample means and the asymptotic distribution is high, even when the underlying population distribution is symmetric with the same fourth moment as the standard normal distribution. Even at the highest sample size for these moment values, 30 in our experiments, the results behave in a similar manner. It seems that

M2 clearly produces biased and very variable estimates of KLI. Thus, it is clearly not appropriate to use it for assessing asymptotic approximations in our settings. One reason for getting these biased and variable estimates of KLI is that the maximum entropy principle provides an extreme entropy estimate. It seems that the formula for the value of the maximum entropy given in (18) should not be used as an estimate for the entropy of an unknown distribution in our case.

The results obtained using the other two methods, (M1 and M3), show that the distributions of the means of random samples taken from symmetric distributions with zero skewness and kurtosis of three give KLI values very close to zero (see Table 2). This indicates that the normal distribution better approximates the true distribution of sample means taken from such symmetric distributions. This is not surprising because the mean of random samples taken from (symmetric) normal distributions have a normal distribution. If we look at the small sample results, when M3 is used and the kurtosis of the underlying population distribution increases, the KLI estimates become significantly different from zero at the 5% level of significance. However M1 does not produce results with a similar pattern. Only the estimates for kurtosis values of 8, 9 and 10 when skewness is equal to 0.5, are significantly different from zero at the 5% level (see Table 3). In small sample sizes such as 3 and 4, we observed that as kurtosis of less skewed underlying population distributions increases, the KLI estimates increase.

When sampling is done from distributions with lower skewness values such as 0, 0.5, and 1.0, some of the estimated values were small negative numbers near zero (see Tables 2 & 3). One of the reasons for this could be sampling errors because all these negative values are insignificantly different from zero. Thus, these values can be considered as negligible positive values because KLI cannot be negative by definition.

Overall, the standard errors show that M3 gives estimates with much less variation than those of M1 for all the different parameter combinations, with a few exceptions. Thus M3 seems better than M1 for estimating KLI for our purpose. Consequently, we shall now interpret the results based on M3. According to the Monte Carlo results, the estimated KLI values range between 0 and 0.124 when considering insignificant negative values (only 3 numbers) as zero. Among the KLI estimates which are significantly different from zero, the lowest value is 0.0042 whereas the highest value is 0.124. The lowest value occurs for a sample size of 5 when skewness is zero. If we can categorise this range to sub-ranges such as higher or moderate values of 0.0042 to 0.124, and small values such as values less than 0.0042, we can discuss how well the true distribution converges to the normal distribution based on these low and high limits of estimates. The asymptotic normal approximation seems to be very reasonable for the distributions which have very small KLI values close to zero (values less than 0.0042). The reverse occurs when the KLI values are very high. In order to illustrate how well the true distributions are

approximated by the normal distribution using our estimates of KLI, we can choose a reasonably appropriate value within the range of significant KLI estimates, as a threshold. The next question is, what should the threshold value be?

We observed that as the sample size increases, the values of the estimated KLI decline. Thus the lowest sample size, which is 3, gives the highest KLI for all kurtosis values. According to the results, as the kurtosis of the underlying population distribution increases, values of estimated KLI increase implying that when the underlying population distribution is away from the normal distribution, KLI increases. For the sample size of 3, the KLI estimates for kurtosis values of 3, 4, 5, 6, 7, 8 and 9 are -0.0001, 0.0008, 0.0047, 0.0097, 0.0145, 0.0189 and 0.0229, respectively. Among these, only the KLI estimates for kurtosis values of 5, 6, 7, 8 and 9 are significantly different from zero. Among the significant values, 0.0145 is the one with the mid value of kurtosis which give significant KLI estimates. The average of significant KLI estimates is also nearly 0.0145. Therefore 0.0145 seems to be a good choice for the threshold value for KLI estimates. Thus we can use the following rule concerning the distributions of sample means:

- KLI estimates $< 0.0042 \Rightarrow$ well approximated by the normal distribution.
- $0.0042 <$ KLI estimates $\leq 0.0145 \Rightarrow$ reasonably approximated by the normal distribution.
- KLI $> 0.0145 \Rightarrow$ poorly approximated by the normal distribution.

We find there are small KLI estimates which are less than 0.0145, for sample sizes 30 and above for kurtosis value of 9 and 10 when data is generated from highly skewed distributions (see Table 4). But when data is generated from distributions with kurtosis of 11 and 12, the minimum sample sizes required for having low KLI estimates are 26 and 24, respectively. When skewness is 2, and as the kurtosis of the underlying population distribution increases, the minimum sample size required for a reasonably normal approximation seems to decrease. For example, for kurtosis in the range 9 – 10, minimum sample size required seems equal to 30 whereas for kurtosis in the range 12 – 15, this becomes 24 (see Tables 4 & 5). Based on our results, sample sizes greater than 30 can be recommended for use of the asymptotic normal approximation in the CLT when sampling from skewed and leptokurtic or medium tailed distributions (see Table 4). However, sample sizes less than that also give a relatively good normal approximation when the population distribution's skewness is less than or equal to one. But as the skewness increases, the possibility of getting a good normal approximation for a small sample diminishes[7].

[7] For brevity these and the following results are not reported. They are available from the corresponding author.

If we look at the behaviour of estimates with changes of kurtosis and skewness values, for leptokurtic distributions with small skewness values, even sample sizes of 3 – 10 can be used for a reasonably good normal approximation. However, the sample mean of random samples taken from highly positive skewed distributions (for example, skewness of 2) does not have a good normal approximation compared to the others. Thus, for sample sizes such as 3 – 20, the normal approximation cannot be recommended when sampling from such distributions because the divergence between the true distribution and the approximating normal distribution is comparatively high. When samples are taken from skewed distributions (for example, skewness of 1.5), sample sizes less than 10 might give poor normal approximations to the distributions of sample means. When sampling is done from asymmetric distributions[8], we clearly see that the KLI values of the true distribution of the sample means with respect to the normal distribution, decreases and converges to zero as sample sizes increase. The results are justifiable due to the CLT. When a threshold value such as 0.0145 is chosen, then sample sizes higher than or equal to 14, give KLI estimates less than 0.0145. Therefore at least 18 observations should be used for the true distribution to be better approximated by the normal distribution, when sampling from an underlying population distribution with skewness of 1.5. When skewness is 2, a similar pattern in KLI estimates can be observed but the minimum sample sizes required for a better normal approximation is higher. For sample sizes greater than 6-8, almost all the KLI estimates are less than 0.0145 in the case of generating data from distributions with skewness of 1. Therefore, it seems that these distributions are reasonably approximated by the normal distribution for sample sizes greater than 8 for all the kurtosis values used in the experiments.

Based on the estimated KLI values, Table 6 summarises the minimum sample size needed for the true distribution of the sample mean to be reasonably approximated by the normal distribution, for particular choices of skewness and kurtosis values. It should be noted that these recommendations are made on the basis of the distributions used in this study. One should not assume that they extend to all distributions with these particular values of skewness and kurtosis. Obviously the shape of the underlying population distribution influences the rate at which a sampling distribution of means converges to a normal distribution.

5. CONCLUSION

This paper considers three nonparametric estimators (kernel, maximum entropy principle and Vasicek's entropy) of the KLI measure to investigate how well the true distribution of means of independent random samples are approximated by the approximating normal distribution in the context of the CLT. For this study, a range of sample sizes were used and the samples were generated from Tukey's lambda distribution with different skewness and kurtosis values. Overall, the Vasicek's entropy performs better

[8] The skewness values 1.5 – 2 used in this paper can be considered as such asymmetric cases.

than the other methods in terms of estimating KLI for assessing asymptotic approximations. Based on this best method, we investigate how distributions affect convergence in the CLT and find which type of distributions give poor asymptotic approximations. As expected, the results suggest that the distribution of the sample mean better approximates the normal distribution as the sample size increases. We have also made some recommendations on minimum sample sizes required for an accurate normal approximation of the true distribution of the sample mean.

Our results indicate that the true distribution of the sample mean when the sample is taken from a highly skewed distribution better approximates the normal distribution as the thickness of the tail of the population distribution increases. In the range of kurtosis values considered, means of small samples generated from symmetric distributions are well approximated by the normal distribution.

REFERENCES

Ahmad, P. and I. Lin (1976). A nonparametric estimation of the entropy for absolute continuous distributions, IEEE Transactions on Information Theory 22, 372-375.

Atukorala, R. (1999). The use of an information criterion for assessing asymptotic approximations in econometrics, PhD Thesis, Monash University, Melbourne.

Bolance, C., Guillen, M., Gustafsson, J. and J.P. Nielsen (2012). Quantitative Operational Risk Models, Taylor & Francis Group, LLC.

Bradley, J.V. (1980). Nonrobustness in z, t and F tests at large sample sizes, Bulletin of the Phychonomic Society 16, 333-336.

Brooks, R.D. and M.L. King (1994). Testing Hildreth-Houck against return to normalcy random regression coefficients, Journal of Quantitative Economics 10, 33-52.

Chang, C.H., Lin, J.J., Pal, N. and M.C. Chiang (2008). A Note on Improved Approximation of the Binomial Distribution by the Skew-Normal Distribution, The American Statistician 62, 167-170.

Evans, M.A. (1992). Robustness of size of tests of autocorrelation and heteroscedasticity to nonnormality, Journal of Econometrics 51, 7-24.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, Annals of Statistics 7, 1-26.

Goncalves, S. and H. White (2005). Bootstrap standard error estimates for linear regression, Journal of the American Statistical Association 100, 970-979.

Guo, J., Alemayehu, D., and Y. Shao (2010). Tests for normality based on entropy divergence, Biopharmaceutical Research 2, 408-418.

Hong, Y., Meeker, W.Q. and L.A. Escobar (2008). Avoiding problems with normal approximation confidence intervals for probabilities, Technometrics 50, 64-68.

King, M.L. and D.C. Harris (1995). The application of the Durbin-Watson test to the dynamic regression model under normal and non-normal errors, Econometric Reviews 14, 487-510.

Kullback, S. (1959). Information Theory and Statistics. New York: John Wiley and Sons.

Kullback, S. and R.A. Leibler (1951). On information and sufficiency, Annals of Mathematical Statistics 22, 79-86.

Little, T.D. (2013). The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis, Oxford University Press.

Maasoumi, E. (1993). A compendium to information theory in economics and econometrics, Econometric Reviews 12, 137-81.

Ott, R. and M. Longnecker (2010). An Introduction to Statistical Methods and Data Analysis, Cengage Learning, USA.

Parzen, E. (1962). On estimation of a probability density and mode, Annals of Mathematical Statistics 33, 1065 – 1076.

Ramberg, J.S., Dudewicz, E.J., Tadikamalla, P.R. and E.F. Mykytka (1979). A probability distribution and its uses in fitting data, Technometrics 21, 201-214.

Ramberg, J.S. and B.W. Schmeiser (1972). An approximate method for generating symmetric random variables, Communication of the Association for Computing Machinery 15, 987 – 990.

Ramberg, J.S. and B.W. Schmeiser (1974). An approximate method for generating asymmetric random variables, Communication of the Association for Computing Machinery 17, 78-87.

Rao, C. (1973). Linear Statistical Inference and its Applications, Wiley, New York.

Renyi, A. (1961). On measures of entropy and information, Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics, University of California Press.

Renyi, A. (1970). Probability Theory, Amsterdam: North-Holland.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates for a probability density function, Annals of Mathematical Statistics 27, 832-837.

Rothenberg, T.J. (1984). Approximating the distributions of econometric estimators and test statistics, Handbook of Econometrics 2, Z. Griliches and M.D. Intriligator (eds.), North-Holland, Amsterdam, 881 – 935.

Shilane, D. and D. Bean (2013). Growth estimators and confidence intervals for the mean of negative binomial random variables with unknown dispersion, Journal of Probability and Statistics, Volume 2013, Article ID 602940.

Shilane, D., Evans, S.N. and A. Hubbard (2010). Confidence intervals for negative binomial random variables of high dispersion, The International Journal of Biostatistics 6, 1-9 .

Silverman, B.W. (1978). Choosing the window width when estimating a density, Biometrika 65, 1-11.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

Smith, Z.R. and C.S. Wells (2006, October 18 – 20). Central Limit Theorem and Sample Size, Retrieved from The Annual Meeting of the Northeastern Educational Research Association: http://www.umass.edu/remp/Papers/Smith&Wells_NERA06.pdf.

Theil, H. and D.G. Fiebig (1984). Exploiting Continuity: Maximum Entropy Estimation of Continuous Distributions, Ballinger Publishing Company, Cambridge.

Vasicek, O. (1976). A test for normality based on sample entropy, Journal of the Royal Statistical Society B 38, 54-59.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 50, 1-26.

White, H. (1982). Maximum likelihood estimation of misspecified models, Econometrica 57, 307-33.

White, H. (1994). Estimation, Inference and Specification Analysis, Cambridge University Press, USA.

Wilcox, R.R. (2003). Applying Contemporary Statistical Techniques, San Diego, CA: Academic Press.

Wilcox, R.R. (2005). Robust estimation and Hypothesis Testing, Elsevier Academic Press, Burlington, MA.

Wu, P.C. (2002). The central limit theorem and comparing means, trimmed means, one step m-estimators and modified one step m-estimators under non-normality, Unpublished monograph, University of Southern California.

Table 1: Selected KLI estimates for different values of m_1 and associated standard errors when the underlying population distribution has the same first four moments as the standard normal

m_1	Sample size (n)					
	3		5		10	
	KLI	s.e	KLI	s.e	KLI	s.e
1	0.271	0.049	0.268	0.049	0.264	0.050
5	0.052	0.003	0.052	0.002	0.051	0.003
10	0.026	0.002	0.026	0.002	0.026	0.002
15	0.017	0.001	0.018	0.001	0.017	0.002
25	0.010	0.001	0.010	0.001	0.010	0.001
40	0.005	0.001	0.005	0.001	0.005	0.001
50	0.004	0.001	0.004	0.001	0.003	0.001
65	0.002	0.001	0.002	0.001	0.002	0.001
75	0.001	0.001	0.001	0.001	0.001	0.001
85	0.000	0.001	0.000	0.001	0.000	0.001
90	0.000	0.001	0.000	0.001	0.000	0.001
100	-0.001	0.001	-0.002	0.001	-0.001	0.001

Table 2: Selected estimates of KLI and associated standard errors for different methods (M1, M2 and M3) and sample sizes (*n*) when skewness = 0 and kurtosis = 3 - 6

		kurtosis							
n		3		4		5		6	
		KLI	s.e	KLI	s.e	KLI	s.e	KLI	s.e
3	M1	-0.006	0.007	-0.006	0.008	-0.004	0.008	0.001	0.008
	M2	0.267	0.038	0.272	0.034	0.279	0.039	0.281	0.038
	M3	0.000	0.001	0.001	0.002	0.005	0.002	0.010	0.002
8	M1	-0.006	0.007	-0.006	0.007	-0.006	0.007	-0.005	0.008
	M2	0.270	0.033	0.268	0.033	0.274	0.035	0.276	0.037
	M3	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001
12	M1	-0.004	0.007	-0.002	0.007	-0.001	0.007	0.000	0.007
	M2	0.274	0.034	0.272	0.037	0.267	0.034	0.273	0.034
	M3	0.002	0.001	0.002	0.001	0.003	0.002	0.002	0.001
16	M1	0.003	0.007	0.008	0.007	-0.004	0.007	0.009	0.009
	M2	0.273	0.034	0.273	0.273	0.272	0.034	0.270	0.034
	M3	0.001	0.001	0.000	0.001	0.001	0.001	0.001	0.001
22	M1	-0.001	0.007	0.000	0.007	-0.002	0.008	-0.001	0.008
	M2	0.267	0.034	0.274	0.038	0.267	0.034	0.270	0.033
	M3	0.001	0.001	0.000	0.001	0.002	0.001	0.002	0.001
26	M1	-0.002	0.007	-0.003	0.007	-0.004	0.007	-0.005	0.007
	M2	0.275	0.035	0.273	0.034	0.268	0.032	0.277	0.032
	M3	0.000	0.001	0.001	0.001	0.000	0.001	0.001	0.001
28	M1	-0.006	0.007	0.002	0.007	0.006	0.007	0.005	0.007
	M2	0.265	0.032	0.270	0.034	0.265	0.033	0.270	0.035
	M3	0.001	0.001	0.002	0.001	0.002	0.001	0.001	0.001
30	M1	0.001	0.007	0.006	0.007	0.005	0.007	0.005	0.007
	M2	0.267	0.037	0.270	0.033	0.270	0.032	0.278	0.033
	M3	0.000	0.001	0.000	0.001	0.000	0.001	0.001	0.001

Table 3: Selected estimates of KLI and associated standard errors for different methods (M1, M2 and M3) and sample sizes (*n*) when skewness = 0.5 and kurtosis = 8-10

		kurtosis					
n		8		9		10	
		KLI	s.e	KLI	s.e	KLI	s.e
3	M1	0.019	0.009	0.022	0.009	0.023	0.010
	M2	0.294	0.036	0.291	0.038	0.302	0.037
	M3	0.022	0.005	0.027	0.005	0.029	0.006
8	M1	0.006	0.008	0.008	0.008	-0.008	0.008
	M2	0.272	0.033	0.280	0.035	0.268	0.033
	M3	0.004	0.002	0.005	0.002	0.005	0.002
12	M1	-0.029	0.028	0.033	-0.034	0.037	-0.037
	M2	0.273	0.038	0.277	0.035	0.273	0.036
	M3	0.005	0.002	0.005	0.002	0.005	0.002
16	M1	-0.001	0.007	0.000	0.008	0.000	0.008
	M2	0.268	0.037	0.267	0.034	0.276	0.033
	M3	0.002	0.001	0.003	0.001	0.003	0.001
22	M1	0.005	0.005	0.005	0.005	0.006	0.006
	M2	0.270	0.032	0.274	0.034	0.266	0.035
	M3	0.004	0.001	0.004	0.001	0.004	0.001
26	M1	0.005	0.007	0.005	0.007	0.005	0.008
	M2	0.267	0.035	0.271	0.035	0.272	0.034
	M3	0.002	0.001	0.002	0.001	0.002	0.001
28	M1	-0.001	0.007	0.001	0.007	0.000	0.007
	M2	0.276	0.035	0.273	0.035	0.270	0.034
	M3	0.002	0.001	0.003	0.001	0.003	0.001
30	M1	0.000	0.007	-0.002	0.007	-0.001	0.007
	M2	0.271	0.033	0.271	0.035	0.274	0.033
	M3	0.002	0.001	0.003	0.001	0.003	0.001

Table 4: Selected estimates of KLI and associated standard errors for different methods (M1, M2 and M3) and sample sizes (*n*) when skewness = 2 and kurtosis = 9-12

		kurtosis							
n		9		10		11		12	
		KLI	s.e	KLI	s.e	KLI	s.e	KLI	s.e
3	M1	0.112	0.009	0.095	0.009	0.084	0.010	0.077	0.010
	M2	0.396	0.037	0.370	0.037	0.363	0.037	0.345	0.035
	M3	0.124	0.005	0.105	0.006	0.097	0.007	0.091	0.008
8	M1	0.046	0.008	0.043	0.008	0.041	0.008	0.039	0.008
	M2	0.312	0.037	0.311	0.036	0.304	0.035	0.304	0.033
	M3	0.043	0.002	0.040	0.003	0.038	0.003	0.037	0.003
12	M1	0.047	0.007	0.045	0.008	0.044	0.008	0.032	0.008
	M2	0.298	0.034	0.291	0.034	0.294	0.032	0.291	0.035
	M3	0.030	0.002	0.029	0.002	0.028	0.002	0.027	0.002
22	M1	0.017	0.007	0.016	0.005	0.017	0.007	0.011	0.007
	M2	0.290	0.033	0.285	0.036	0.284	0.033	0.291	0.035
	M3	0.017	0.002	0.016	0.002	0.016	0.002	0.016	0.002
24	M1	0.017	0.007	0.017	0.007	0.012	0.007	0.033	0.017
	M2	0.282	0.035	0.282	0.034	0.284	0.034	0.285	0.036
	M3	0.016	0.002	0.015	0.001	0.016	0.002	0.013	0.002
26	M1	0.014	0.006	0.023	0.017	0.013	0.008	0.022	0.015
	M2	0.282	0.035	0.282	0.034	0.289	0.036	0.300	0.034
	M3	0.017	0.002	0.015	0.001	0.015	0.002	0.012	0.002
30	M1	0.001	0.007	0.008	0.007	0.007	0.007	0.007	0.007
	M2	0.285	0.035	0.280	0.036	0.284	0.034	0.285	0.036
	M3	0.013	0.001	0.013	0.002	0.012	0.001	0.011	0.001
34	M1	0.004	0.007	0.005	0.007	0.005	0.007	0.005	0.007
	M2	0.286	0.034	0.280	0.034	0.281	0.034	0.284	0.036
	M3	0.011	0.002	0.012	0.001	0.011	0.002	0.011	0.001

Table 5: Selected estimates of KLI and associated standard errors for different methods (M1, M2 and M3) and sample sizes (*n*) when skewness = 2 and kurtosis = 13-15

		kurtosis					
n		13		14		15	
		KLI	s.e	KLI	s.e	KLI	s.e
3	M1	0.072	0.010	0.069	0.010	0.066	0.011
	M2	0.345	0.037	0.343	0.037	0.337	0.038
	M3	0.088	0.009	0.086	0.009	0.085	0.010
8	M1	0.038	0.008	0.037	0.008	0.037	0.009
	M2	0.304	0.033	0.305	0.035	0.307	0.035
	M3	0.035	0.003	0.035	0.003	0.034	0.003
12	M1	0.042	0.008	0.041	0.008	0.041	0.008
	M2	0.294	0.036	0.295	0.033	0.298	0.034
	M3	0.027	0.003	0.026	0.003	0.027	0.003
16	M1	0.021	0.008	0.021	0.008	0.021	0.008
	M2	0.287	0.035	0.293	0.033	0.288	0.036
	M3	0.019	0.002	0.019	0.002	0.019	0.002
22	M1	0.010	0.008	0.010	0.008	0.010	0.008
	M2	0.278	0.036	0.285	0.033	0.280	0.034
	M3	0.016	0.002	0.015	0.002	0.015	0.002
24	M1	0.033	0.018	0.014	0.008	0.014	0.008
	M2	0.273	0.035	0.284	0.038	0.281	0.034
	M3	0.014	0.002	0.011	0.002	0.011	0.002
28	M1	0.016	0.007	0.015	0.008	0.015	0.008
	M2	0.280	0.032	0.279	0.033	0.280	0.035
	M3	0.014	0.002	0.013	0.002	0.012	0.002
34	M1	0.005	0.007	0.005	0.007	0.005	0.008
	M2	0.280	0.036	0.275	0.034	0.278	0.034
	M3	0.010	0.002	0.011	0.002	0.010	0.001

Table 6: Minimum sample size needed for the true distribution of the sample mean to be reasonably approximated by the normal distribution

Kurtosis	skewness = 0	skewness = 0.5	skewness = 1	skewness = 1.5	skewness = 2
3	3	.	.	.	.
4	3	.	8	.	.
5	3	3	6	.	.
6	3	3	8	14	.
7	3	4	6	14	.
8	4	4	6	14	.
9	4	5	8	14	30
10	.	6	8	.	30
11	.	.	.	.	26
12	.	.	.	.	24
13	.	.	.	.	24
14	.	.	.	.	24
15	.	.	.	.	24

Note: As noted in Section 3, not all combinations of skewness and kurtosis values are estimated, which explains the missing values.