Central Limit Theorem in a Skewed Leptokurtic Distribution

Teoram Had Memusat dalam Taburan Kepencongan Leptokurtic

Nor Aishah Ahad[1], Che Rohani Yaacob[2], Abdul Rahman Othman[2],
Ng Seou Ling[2] and Teoh Sian Hoon[3]
[1]Universiti Utara Malaysia,
[2]Universiti Sains Malaysia, [3]UiTM Pulau Pinang
[1]Corresponding author e-mail:

Abstract

According to the central limit theorem, the distribution of the sample mean is approximately normal if the sample size, *n*, is sufficiently large, regardless of original data distribution. However there seems to be a difference in opinion on how large *n* should be. Some books said that *n* = 25 is sufficient enough while some considered *n* = 30 to be sufficient. This paper investigates the size of *n* that would allow us to apply the central limit theorem when samples taken are from a $|^{3}_{2}$ population.

Keywords normal, $|^{3}_{2}$, central limit theorem

Abstrak

Menurut teorem had memusat, taburan bagi min sampel adalah menghampiri normal jika saiz sampel, *n*, adalah cukup besar, tanpa mengambil kira taburan data asal. Walau bagaimanapun terdapat perbezaan pendapat tentang nilai *n* yang sepatutnya digunakan. Terdapat beberapa buku mengatakan bahawa *n* = 25 adalah memadai sementara sesetengah menganggap *n* = 30 adalah mencukupi. Artikel ini mengkaji saiz *n* yang akan membenarkan kami untuk menggunakan teorem had memusat apabila sampel diambil daripada taburan $|^{3}_{2}$.

Kata kunci normal, $|^{3}_{2}$, teori had memusat

Introduction

To use a parametric test that deals with the sample mean, the samples taken must be from a normal distribution. Theoretically, if a population has a normal distribution, then for any sample size, the distribution of the sample mean will also be normally distributed.

If the parent distribution is non normal then we would have to use a nonparametric test. However, we can also use a parametric test on a non normal distribution provided that the sample size is large. But how large is considered large? To answer the question we refer to the central limit theorem (CLT).

The central limit theorem was first formulated in the early nineteenth century by Laplace and Gauss (Milton & Arnold, 2003). The theorem states that given a distribution with mean n and variance σ^2, the distribution of the sample mean approaches a normal distribution with mean n and variance $\frac{v^2}{n}$ as *n*, the sample size, increases. The central limit theorem from Chen (1995) states that when a random sample is drawn from a non-normal population with mean n and finite standard deviation σ, the sampling distribution of $\overline{X}$ is approximately normal for large n, with mean n and standard deviation $\frac{v}{\sqrt{n}}$, This means that no matter what the shape of the original distribution is, the distribution of the sample mean is normal. However, there are some distributions for which CLT does not hold, for example the extremely dispersed Cauchy distribution (Lovric, 2007).

It is often suggested that a sample size of 30 will produce an approximately normal sampling distribution for the sample mean from a non-normal parent distribution. There is little to no documented evidence to support that a sample size of 30 is the magic number for non-normal distributions (Smith & Wells, 2006). Bluman (2007), Mann (2005) and Freund (1995) said that in general, the central limit theorem holds when *n* is at least 30. Smith (1998) stated that the central limit theorem can be applied when the sample size *n* exceeds 25. Newbold, Carlson and Thorne (2007) used random Monte Carlo sample simulations to demonstrate the central limit theorem. They choose 1,000 random samples of size 25 from symmetric and skewed distributions. They concluded that for symmetric and skewed distribution, the distribution of samples means is approximately normal when the sample size is 25 or more.

Hence, when the CLT holds, what sample size that can be deemed large enough? The central limit theorem, however, does not clearly state what is meant by “a sufficiently large sample”. Arsham (2005) as cited in Smith and Wells (2006) claims that it is not even feasible to state when the central limit theorem works or what sample size is large enough for good approximation, but the only thing most statisticians agree on is that, when the parent distributions is symmetric and relatively short tailed, then the sample mean reaches approximately normality for smaller samples than if the parent population is skewed or long-tailed.

According to Anderson and Slove (1974), how large *n* must be in order to have a ‘good’ approximation depends upon the shape of the parent population. If the shape of the parent population is symmetric, the sample means will be normally distributed at smaller sample sizes then when the parent population is skewed. On the other hand, Johnson and Kuby (2007) showed that the sample means from a *J*-shaped and *U*-shaped distributions have some resemblance of a normal curve when *n* is only 5.

In this study, however, we will not look into the shapes of the parent distribution but rather the sample size issue. We select the $|^3_2$ distribution, which is skewed leptokurtic and randomly choosed samples of various sizes. We then note the smallest size of *n* where the central limit theorem holds in this distribution.

Method

This section describes the steps taken in order to test the distribution of sample mean from a $|^3_2$ distribution in which the mean *m* = 3 and variance v^2 = 6. The following algorithm was used to conduct the study.

1. Use the SAS generator RANDGEN (SAS Institute 1999) to generate samples of size *n* = 15 from a $|^3_2$ distribution. A total of *N* = 300 samples were generated.
2. From each *i*-th sample (where *i* =1, …, 15), the sample mean, $\overline{\text{V}}_i$, is calculated. Therefore a total of 300 sample mean values were obtained.
3. Observe the mean, standard deviation, skewness and kurtosis of the 300 sample mean values obtained in step (ii).
4. Compare the observed mean, standard deviation, skewness and kurtosis with their theoretical values.
5. Perform these normality tests: Kolmogorov-Smirnov, Shapiro-Wilk, Cramer-von Mises and Anderson-Darling.
6. Repeat steps (i) to (v) for *n* = 20, 25, 30, 35 and 40.

Results and discussion

Table 1 presents the observed values of various measurement against their theoretical values for sample sizes *n* = 15, 20, 25, 30, 35, 40. The observed means and the observed standard deviations were all very close to their theoretical values for all sample sizes. The skewness and kurtosis values indicate the deviation of the distribution from the normal distribution. The observed values of skewness range from 0.1654 to 0.5340 while the observed values of kurtosis range from -0.0392 to 0.7128 indicating all the distributions deviate from the normal distribution.

Formal test for the normality of the distribution of the sample mean are carried out using four tests, namely the Kolmogorov-Smirnov, Shapiro-Wilk, Cramer-von Mises and Anderson-Darling. The *p*-values of those tests are shown in Table 2. The nominal value was set at 0.05. From Table 2, the *p*-values are larger than 0.05 for all tests when sample size was at least 20 indicating the distribution for the sample mean is normal. While for *n* = 15, the *p*-values of all the tests are less than the nominal value, indicating the distribution for the sample mean is not normal. From the results in Table 2, we can conclude that the sample mean from a $|^3_2$ distribution is normally distributed when the sample size is at least *n* = 20.

Table 1 Theoretical and observed values for various sample sizes.

Measure	Theoretical Values	Observed values from various sample sizes					
		n = 15	*n* = 20	*n* = 25	*n* = 30	*n* = 35	*n* = 40
Mean	3	3.0171	3.0279	3.0213	3.0122	3.0017	3.0010
Std. dev. when							
n = 15	0.6325	0.6378					

$n = 20$	0.5477		0.5380				
$n = 25$	0.4899			0.4848			
$n = 30$	0.4472				0.4389		
$n = 35$	0.4140					0.4215	
$n = 40$	0.3873						0.3898
Skewness	0	0.5340	0.2274	0.2015	0.1826	0.1654	0.2752
Kurtosis	0	0.7128	-0.0469	-0.2345	-0.0392	-0.0592	-0.0412

Table 2 The *p*-values of various normality tests

Normality Test	Sample Sizes					
	$n = 15$	$n = 20$	$n = 25$	$n = 30$	$n = 35$	$n = 40$
Shapiro-Wilk	0.0010	0.2972	0.3223	0.4505	0.4858	0.1112
Kolmogorov-Smirnov	0.0100	0.1500	0.1500	0.0466	0.1500	0.0695
Cramer-von Mises	0.0126	0.1508	0.2500	0.0886	0.2500	0.1544
Anderson-Darling	0.0177	0.1762	0.2500	0.1389	0.2500	0.1292

To further support the results of the normality tests, we also used the normal probability plot or *p*-*p* plot. The *p*-*p* plots of the sample means for sample sizes n = 15, 20, 25, 30, 35, 40 are shown in Figure 1(a) to (f). The linear pattern of the plots provides evidence that the sample means are normally distributed.

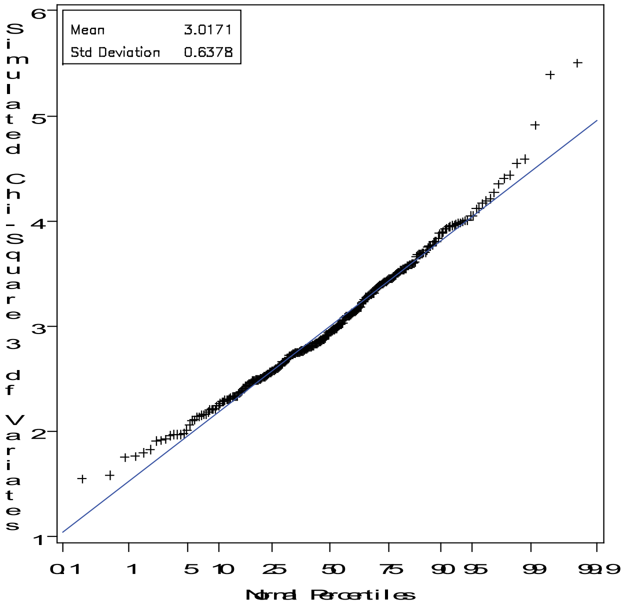

Figure 1(a) *p*-*p* plot for sample mean when n = 15

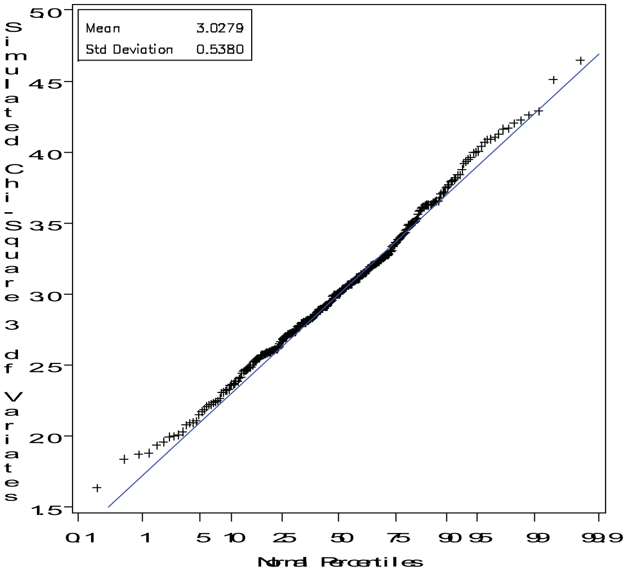

Figure 1(b) *p-p* plot for sample mean when n = 20

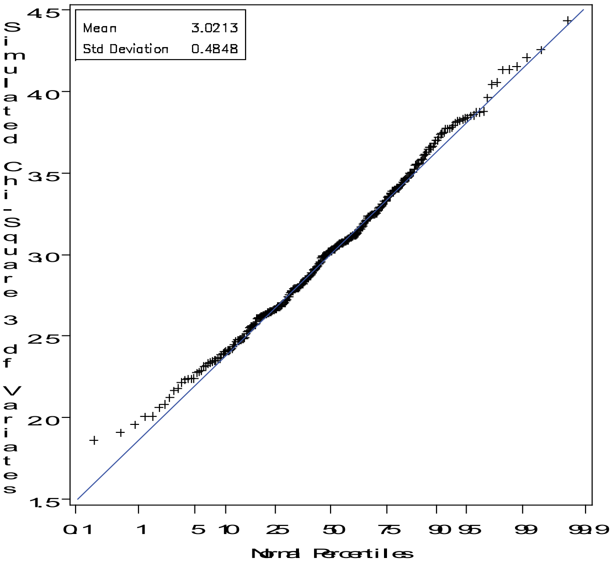

Figure 1(c) *p-p* plot for sample mean when n = 25

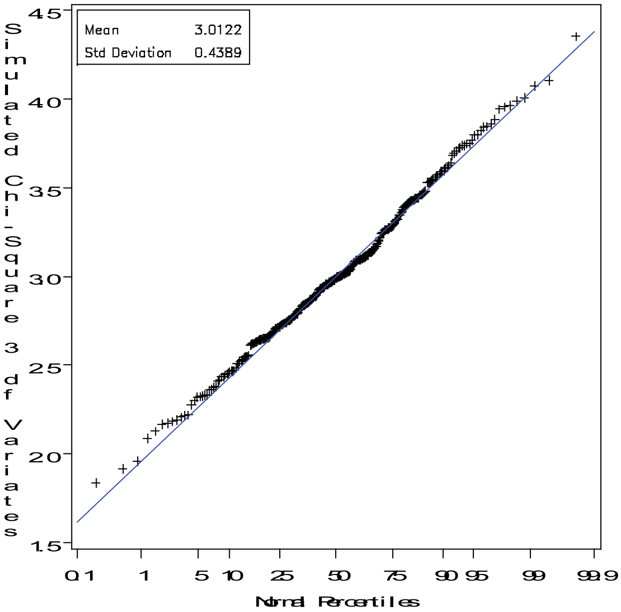

Figure 1(d) *p-p* plot for sample mean when *n* = 30

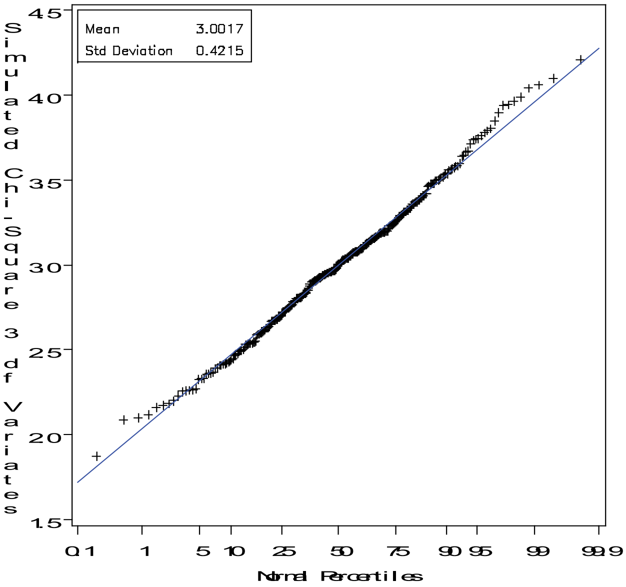

Figure 1(e) *p-p* plot for sample mean when *n* = 35

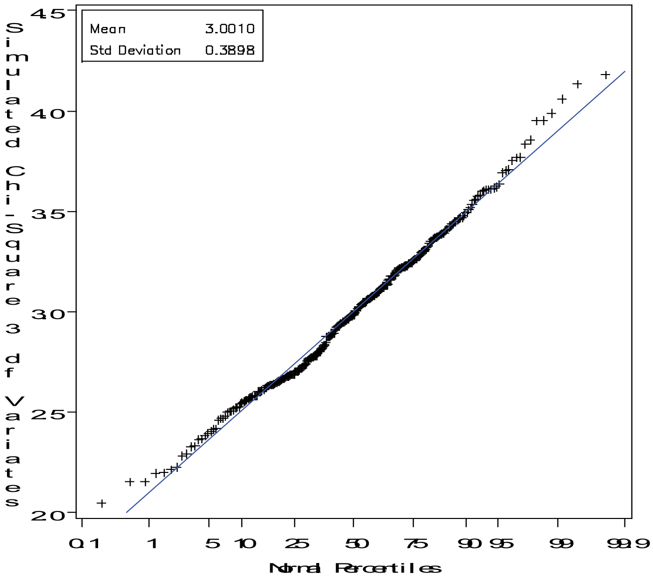

Figure 1(f) *p-p* plot for sample mean when $n = 40$

Conclusion

The central limit theorem is known to be true when the size of samples taken is sufficiently large. In our study, we tried to find out the exact size of sample in which the central limit theorem actually holds in a particular distribution. The theorem was tested on a χ_3^2 distribution by using samples generated by SAS routine. Based on the study, we discovered that in a χ_3^2 distribution, the central limit theorem holds when the sample size, n, is at least 20.

Acknowledgement

We greatly appreciate the helpful comments of the anonymous referees and editor. Their comments have contributed in the improvement of this article. The work that led to the publication of this paper was funded by the Fundamental Research Grant Scheme of the Malaysian Ministry of Higher Education, and supported by the Universiti Sains Malaysia and Universiti Utara Malaysia.

References

Anderson, T. W. and Slove. S. L. (1974). *Introductory Statistical Analysis*. Houghton Mifflin Company.

Bluman, A. G. (2007). *Elementary Statistics a Step by Step Approach*. Sixth Edition. New York: McGraw Hill.

Chen, L. (1995). Testing the mean of skewed distributions. *Journal of the American Statistical Association*, 90, 767-772.

Freund, J. E. and Simon, G. A. (1995) *Statistics A First Course.* Sixth Edition. New Jersey: Prentice Hall.

Johnson, R. and Kuby, D. (2007). *Elementary Statistics.* Tenth Edition. Belmont, CA: Thomson

Lovric, M. M. (2007). *Note on the correct interpretation of the Central Limit Theorem.* Retrieved 24th May, 2011, from http://www.bmj.com/content/310/6975/298/reply.

Mann, P. S. (2005). *Introductory Statistics.* Fifth Edition. Noida, India: Wiley.

Milton, J. S. and Arnold, J. C. (2003). *Introduction to Probability and Statistics* (4th ed.). New York: McGraw Hill.

Newbold, P., Carlson, W. L. and Thorne, B. (2007). *Statistics for Business and Economics* (6th ed.). New Jersey: Pearson Education.

SAS Institute Inc. (1999). *SAS /IML User's Guide version 8.* Cary, NC: SAS Institute Inc.

Smith, P. J. (1998). *Into Statistics* (2nd ed.). Springer.

Smith, Z. R. and Wells, C. S. (2006). *CLT and sample size*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Kerhonkson, New York.