A graphical approach for verification of the central limit theorem

Suresh Kumar Veluchamy
University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds
Part of the Higher Education Commons

Repository Citation
Veluchamy, Suresh Kumar, 'A graphical approach for verification of the central limit theorem' (2005). UNLV Retrospective Theses & Dissertations. 2203.
https://digitalscholarship.unlv.edu/rtds/2203

This Thesis is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
A GRAPHICAL APPROACH FOR VERIFICATION OF THE CENTRAL LIMIT THEOREM

by

Suresh Kumar Veluchamy

Master of Science Degree in Computer Science
University of Nevada, Las Vegas
2001

Bachelor in Computer Science and Engineering
College of Engineering, Guindy
Anna University, Madras, India
1999

A thesis submitted in partial fulfilment
of the requirements for the

Master of Science Degree in Mathematical Sciences
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
December 2005
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
UNLV

Thesis Approval
The Graduate College
University of Nevada, Las Vegas

November 15, 2005

The Thesis prepared by
Suresh Veluchamy

Entitled
A Graphical Approach for Verification of The Central Limit Theorem

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Sciences

Examination Committee Chair

Dean of the Graduate College
ABSTRACT

A Graphical Approach for Verification of the Central Limit Theorem

by

Suresh Kumar Veluchamy

Dr. Ashok K. Singh, Examination Committee Chair
Professor of Mathematical Sciences
University of Nevada, Las Vegas

The distribution of the sample mean, when sampling from a normally distributed population, is known to be normal. When sampling is done from a non-normal population, the above result holds when the number of samples (n) is sufficiently large. This important result is known as the Central Limit Theorem (CLT). The CLT plays a very important role in statistical inference. The logical question that arises is: how large does n have to be before the CLT can be used? No one answer is available in the statistical literature, since n depends on the extent of nonnormality present in the underlying population. A rule of thumb given in almost every introductory applied statistics text is that n = 30 is sufficient for most cases. In this thesis, the method of bootstrap is used to develop a graphical approach to determine if the CLT will be valid for any given random sample. A computer program in C#.NET is developed and Monte Carlo simulation is used to demonstrate the program.
# TABLE OF CONTENTS

ABSTRACT ............................................................................................................................ iii

LIST OF TABLES .................................................................................................................. vi

LIST OF FIGURES .............................................................................................................. vii

ACKNOWLEDGEMENTS .................................................................................................. viii

CHAPTER 1 INTRODUCTION ...........................................................................................1
  1.1 Laplace’s Initial Result ...............................................................................................2
  1.2 Liapounov Theorem .....................................................................................................3
  1.3 Lindeberg Theorem ......................................................................................................4
  1.4 A Simple Proof of the CLT .........................................................................................4
  1.5 Proof of the Central Limit Theorem ...........................................................................5
  1.6 Product of Random Variables .................................................................................5

CHAPTER 2 DISTRIBUTIONS CONSIDERED ................................................................7
  2.1 The Log-Normal Distribution .....................................................................................7
  2.2 The Weibull Distribution ............................................................................................8
  2.3 Gamma distribution ..................................................................................................10
  2.4 The Uniform Distribution .........................................................................................12
  2.5 Inverse Transform Sampling Method .......................................................................13

CHAPTER 3 MODIFIED BOOTSTRAP CLT VERIFICATION ....................................15
  3.1 The Modified Bootstrap Method ..............................................................................15

CHAPTER 4 SIMULATION MODE ..................................................................................18
  4.1 Results From Simulation Mode .................................................................................18
  4.2 Conclusions Drawn From Simulation Mode Results .................................................19

CHAPTER 5 DETERMINATION MODE .........................................................................20
  5.1 Results from Determination Mode ...........................................................................20

CHAPTER 6 CONCLUSIONS ............................................................................................24

APPENDICES ................................................................................................................... cdrom
  A.1. Lognormal ................................................................................................................25
  A.2 Weibull .......................................................................................................................36
  A.3. Gamma .....................................................................................................................47
  A.4. Verification ...............................................................................................................61
  A.5. Verification ...............................................................................................................63
LIST OF TABLES

Table 1  Probability Models with Varying Levels of Skewness..........................18
Table 2  Number of samples n needed for CLT to hold ..................................19
Table 3  Vehicle emission data for CO (gms/mile) ...........................................20
Table 4  Descriptive statistics computed for CO data.........................................20
Table 5  EPA Guidance Document Dataset (1996)................................................23
LIST OF FIGURES

Figure 1  Log-normal pdf's for $\mu = 5$ and $\alpha = 0.5, 1, \text{ and } 2$ .................................8
Figure 2  Weibull pdf's for $\beta = 1, \alpha = 1, 2, 3$ (b = $\beta$, a = $\alpha$)........................................9
Figure 3  Weibull pdf's for $\alpha = 2, \beta = 2, 3, 4$ (b = $\beta$, a = $\alpha$).....................................10
Figure 4  Graphs of the Gamma pdf's G($\beta , \alpha$) (a = $\alpha$, b = $\beta$)...............................................11
Figure 5  Main Screen in Determination Mode ........................................................................17
Figure 6  Main Screen in Simulation Mode ..............................................................................17
Figure 7  Output from the determination mode for data in Table 3 with m = 10 ............21
Figure 8  Output from the determination mode for data in Table 3 with m = 25 .............22
Figure 9  Output from the determination mode for data in Table 3 with m = 45 .............22
Figure 10 Output from the determination mode for data in Table 5 with m = 9 ..............23
ACKNOWLEDGEMENTS

A very sincerely thank you goes out to the following people:

☐ To my advisor, thesis committee chair, and mentor, Dr. Ashok K. Singh: your intelligence and extraordinary talent has been very fascinating to me. Your compassion provided me the right atmosphere to learn and grow. I enjoyed every moment we worked together. You are truly unique. Thank you so much for investing your time in me, and for all of your help and guidance while I have been a graduate student here at UNLV.

☐ To Dr. Dieudonne Phanord, Dr. Rohan Dalpatadu, and Dr. Laxmi Gewali: thank you for taking the time out of your busy schedules to serve on my thesis examination committee.
CHAPTER 1

INTRODUCTION

Central limit theorems are a set of weak-convergence results in probability theory, expressing the fact that any sum of many independent identically distributed random variables has an approximate normal distribution. These results explain why the normal distribution is central to statistical methods.

The most important and famous result is simply called The Central Limit Theorem (CLT); it is concerned with sums of independent and identically distributed (iid) variables with finite mean and variance. Several generalizations exist which replace the requirement for identical distribution by some condition, which guarantees that none of the variables exert a much larger influence than the others. Two such conditions are the Lindeberg condition and the Lyapunov condition. There are other generalizations that even allow some "weak" dependence of the random variables.

Abraham de Moivre (1733) introduced the concept of the normal distribution as an approximation to the binomial distribution. This result is now called the theorem of de Moivre – Laplace. (see Feller, 1950).

The de-Moivre-Laplace theorem is a special case of the CLT, and states that the distribution of the number of successes in n independent Bernoulli trials approaches a normal distribution as n is increased.
The probability distribution function (pdf) of the arithmetic mean for distributions with finite support was available even before 1910. For example, the pdf of the sum

\[ S_n = X_1 + X_2 + \ldots + X_n \]

where \( X_1, X_2, \ldots, X_n \) are independent discrete uniform \( \text{U}[-a, +a] \) random variables was available in the following form:

\[
P(S_n = s) = \frac{\sum_{j=1}^{n} (-1)^j \binom{n}{j} \frac{n+s+na-1-(2a+1)j}{n-1}}{(2a+1)^n}
\]

where

\[ 0 \leq j \leq \left\lfloor \frac{s+na}{2a+1} \right\rfloor \]

Simpler approximations to such results were obviously needed. Laplace (1810) generalized De Moivre’s work to sums of any discrete random variables with finite mean and variance, and used the characteristic function to rigorously prove the following result, which has come to be known as the first version of CLT:

### 1.1 Laplace’s Initial Result

Let \( X_j \) be integer-valued random variables defined on integers \( \{-a, \ldots, 0, \ldots, +a'\} \) with probability \( f\left(\frac{x}{a+a'}\right) \), then for large \( n \), the sum

\[ S_n = X_1 + X_2 + \ldots + X_n \]

is approximately normal \( N(n\mu, n\sigma^2) \) where \( E(X) = \mu, \text{var}(X) = \sigma^2 \).

Laplace further argued that, since the limiting distribution does not involve \( a' \) and \( a \), the result is valid for discrete distributions with infinite support if the moments are finite. He also proved, but not formally, that his result holds true for continuous random
variables with bounded support. It is generally accepted in the literature, however, that the classical CLT is due to Laplace:

The normal density was derived by Gauss in 1809 in his "Theoria motus corporum celestium" ("Theory of Motion of the Heavenly Bodies Moving About the Sun in Conic Sections").

Laplace's CLT was generalized by Poisson in 1824. He proved the result for iid random variables, first for the sum, then for a linear combination, and also for the sum of independent variables with different distributions. Poisson also noticed that the CLT did not hold for the Cauchy distribution. Bessel gave a different proof using Dirichlet's result. Cauchy in 1853 used the method of characteristics functions to prove a version of the CLT for continuous random variables with finite support.

The CLT still needed to be proved for distributions with infinite support, the conditions on moments under which the CLT holds were still unknown, and the convergence rate was not derived. Russian probabilists during 1870 to 1910 provided the solutions to these problems: Chebychev and Markov used the method of moments, and Liapounov used the characteristic function to give the first "rigorous" proof of general CLT. The following convergence results are given in Ash(1972).

1.2 Liapounov Theorem

Let \( X_1, X_2, ..., X_n \) be independent, identically distributed (iid) random variables with

\[
E(X_i) = m_i < \infty, Var(X_i) = \sigma_i^2 < \infty.
\]

Let \( T_n = \frac{S_n - E(S_n)}{c_n} \), where \( S_n = \sum_{i=1}^{n} X_i, c_n^2 = Var(S_n) = \sum_{i=1}^{n} \sigma_i^2 \). If for some \( \delta > 0 \),

\[
\frac{T_n}{\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).
\]
\[ \sum_{i=1}^{n} \frac{E|X_i - m_i|^2}{c_n^{2+\delta}} \to 0 \text{ as } n \to \infty, \text{ then} \]
\[ T_n \overset{d}{\to} N(0,1). \]

1.3 Lindeberg Theorem

Let \( X_1, X_2, \ldots, X_n \) be independent, identically distributed (\textit{iid}) random variables with
\[ E(X_i) = m_i < \infty, \text{Var}(X_i) = \sigma_i^2 < \infty. \]

Let \( T_n = \frac{S_n - E(S_n)}{c_n} \), where \( S_n = \sum_{i=1}^{n} X_i \sigma_i^2 = \text{Var}(S_n) = \sum_{i=1}^{n} \sigma_i^2. \) If for every \( \varepsilon > 0, \)
\[ \sum_{k=1}^{n} \frac{\int_{|x-m_k|^2 \le \varepsilon c_n} (x-m_k)^2 dF_k(x)}{c_n^{2}} \to 0 \text{ as } n \to \infty, \text{ then} \]
\[ T_n \overset{d}{\to} N(0,1). \]

We now state and prove the simple form of the CLT.

1.4 A Simple Proof of the CLT

Theorem (CLT): Let \( X_1, X_2, \ldots, X_n \) be independent, identically distributed (\textit{iid}) random variables with \( E(X) = \mu < \infty, \text{Var}(X) = \sigma^2 < \infty, \) then
\[ Z_n = \frac{\sum_{j=1}^{n} X_j - n \mu}{\sigma \sqrt{n}} \to N(0,1) \text{ as } n \to \infty \]
or
\[ \frac{\sum_{j=1}^{n} X_j - n \mu}{\sigma \sqrt{n}} \le z \]
\[ P \left( \frac{\sum_{j=1}^{n} X_j - n \mu}{\sigma \sqrt{n}} \le z \right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \]
uniformly in \( z, \) as \( n \to \infty. \)
1.5 Proof of the Central Limit Theorem

For any random variable, $Y$, with zero mean and unit variance ($\text{var}(Y) = 1$), the moment generating function (mgf) of $Y$ is, by Taylor's theorem,

$$m_Y(t) = E(e^{itY}) = 1 + \frac{t^2}{2} + o(t^2)$$

Let $Y_i = \frac{X_i - \mu}{\sigma}$.

Then,

$$E(Y_i) = 0, \text{Var}(Y_i) = 1,$$ and

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\sum_{i=1}^{n} (X_i/n) - \mu}{\sigma / \sqrt{n}} = \frac{\sum_{i=1}^{n} (X_i - \mu)/n}{\sigma / \sqrt{n}} = \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} = Z_n \ (\text{say})$$

Hence,

$$m_{Z_n}(t) = \prod_{i=1}^{n} m_{Y_i}(\frac{t}{\sqrt{n}}) = [1 + \frac{t^2}{2} + o(t^2)]^n \xrightarrow{t \to 0} e^{t^2/2}$$

In words, the mgf of $Z_n$ approaches the mgf of the standard normal distribution, which proves the theorem.

1.6 Product of Random Variables

The logarithm of a product is the sum of the logs of the individual terms in a product, so the log of a product of random variables tends to have a normal distribution, which implies that the product itself has a log-normal distribution. Many physical quantities (especially mass or length, which are a matter of scale and cannot be negative) can be thought of as the product of different random factors, so they follow a log-normal distribution.
It is stated in every applied statistics book that, in inference problems concerning the population mean, the underlying probability distribution must be approximately normal. A closer look at all of the formulas for computing confidence intervals for a population mean shows that, what is actually needed is that the sampling distribution of the sample mean be approximately normal. The purpose of this thesis is to develop a method and a computer program that can be used to verify whether the sampling distribution of the sample mean is approximately normal or not, i.e., whether the CLT holds for the sample or not. The proposed method, based upon bootstrap sampling from the given sample, is explained in Chapter 3 of this thesis. Monte Carlo simulation is used to illustrate the procedure. It is generally known that, when sampling from a symmetric probability distribution, a small number of observations is sufficient for the CLT to hold. For this reason, simulation is done from probability distributions that range from symmetric to heavily skewed. Chapter 2 of this thesis describes these probability distribution models.
CHAPTER 2

DISTRIBUTIONS CONSIDERED

The Log-Normal, Weibull, Gamma and Uniform probability distributions are used in this study.

2.1 The Log-Normal Distribution

The log-normal distribution (Gilbert, 1987) is a probability distribution which is closely related to the normal distribution: the random variable (rv) X follows a log-normal distribution

\[ \text{LN}(\mu, \sigma) \] if and only if \( \ln(X) \) follows a normal distribution with mean \( \mu \) and standard deviation \( \sigma \) \([N(\mu, \sigma)]\).

A variable might be modelled as log-normal if it can be thought of as the product of many small independent factors. A typical example is the long-term return rate on a stock investment: it can be considered as the product of the daily return rates. The log-normal distribution is very commonly used in environmental science and engineering.

The log-normal distribution \( \text{LN}(\mu, \sigma) \) has probability density function

\[
f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, x > 0
\]

where \( \mu \) and \( \sigma \) are the mean and standard deviation of \( \ln(X) \). The parameters of interest of \( \text{LN}(\mu, \sigma) \) are given below:
$E(X) = e^{\mu + \frac{3}{2}\sigma^2}$
Median = $e^{\mu}$

Coefficient of variation (CV) = $\sqrt{e^{\sigma^2} - 1}$

$\text{Var}(X) = CV \times e^{2\mu + \sigma^2}$

Skewness = $(CV)^3 + 3(CV)$

Figure 1 shows graphs of the log-normal probability density functions (pdf) for $\mu = 5$ and varying $\sigma$. It is clear from the above expression for skewness that the log-normal distribution approaches a symmetric distribution as $\sigma$ approaches 0.

![Plots of lognormal probability density functions](image)

Figure 1 Log-normal pdf's for $\mu = 5$ and $\sigma = 0.5, 1, \text{ and } 2$

2.2 The Weibull Distribution

The *Weibull distribution* (named after Wallodi Weibull) $\text{WEI}(a, b)$ is a continuous probability distribution with the probability density function (see Gilbert, 1987):
and cumulative distribution function (cdf)
\[ F(x; \alpha, \beta) = 1 - e^{-(x/\beta)^\alpha}, x > 0 \]

where \( \alpha > 0 \) is the shape parameter and \( \beta > 0 \) is the scale parameter of the distribution.

The parameters of interest of \( \text{WEI}(\alpha, \beta) \) are given below:

\[ E(X) = \beta \Gamma(1 + \alpha^{-1}) \]
\[ \text{Var}(X) = \beta^2 [\Gamma(1 + 2\alpha^{-1}) - \Gamma^2(1 + \alpha^{-1})] \]
\[ \text{Skewness} = \frac{2\Gamma^3(1 + \alpha^{-1}) - 3\Gamma(1 + \alpha^{-1})\Gamma(1 + 2\alpha^{-1}) + \Gamma(1 + 3\alpha^{-1})}{[\Gamma(1 + 2\alpha^{-1}) - \Gamma^2(1 + \alpha^{-1})]^{3/2}} \]

Figures 2-3 show the probability density functions (pdf) of the Weibull Distribution for selected values of \( \alpha \) and \( \beta \).
The Exponential distribution (when $\alpha = 1$) and Rayleigh distribution (when $\alpha = 2$) are two special cases of the Weibull distribution. Weibull distributions are often used to model the time until a given technical device fails.

2.3 Gamma distribution

The Gamma Distribution $G(\alpha, \beta)$ is given by (Gilbert, 1987):

$$f(x; \beta, \alpha) = \frac{1}{\Gamma(\beta)\alpha^\beta} x^{\beta-1} e^{-x/\alpha} \text{ for } x>0$$

where $\beta > 0$ is the shape parameter and $\alpha > 0$ is the scale parameter of the gamma distribution. The cumulative distribution function can be expressed in terms of the incomplete gamma function.

$$F(x; \beta, \alpha) = \int_0^x f(u; \beta, \alpha)du = \frac{\gamma(\beta, x/\alpha)}{\Gamma(\beta)}$$
The expected value, variance, and skewness of a gamma random variable X are:

\[E(X) = \beta \alpha\]
\[Var(X) = \beta \alpha^2\]
\[Skewness = \frac{2}{\sqrt{\beta}}\]

For \(\beta = 1\), the gamma distribution is an exponential distribution with parameter \(\alpha\). If \(\beta\) is an integer, the gamma distribution is an Erlang distribution (so named in honor of A.K. Erlang) and is the probability distribution of the waiting time of the \(\beta\)-th "arrival" in a one-dimensional Poisson process with intensity \(1/\alpha\).

If \(\beta\) is a half-integer and \(\alpha = 2\), then the gamma distribution is a chi-square distribution with \(2\beta\) degrees of freedom.

Figure 4 shows graphs of the gamma pdf’s \(G(\beta, \alpha)\) for \(\alpha=1,2,3\) and \(\beta=1,1,2\).

Figure 4 Graphs of the Gamma pdf’s \(G(\beta, \alpha)\) (\(a = \alpha, b = \beta\))
2.4 The Uniform Distribution

There are two types of uniform distribution: discrete and continuous.

2.4.1 The Discrete Case

In the discrete case, if there are \( N \) possible outcomes \( x_1, x_2, \ldots, x_N \) which are distributed uniformly, then the probability of the outcome \( x_n \) is:

\[
P(x_n) = \frac{1}{N}
\]

A simple example of the discrete uniform distribution is throwing a fair die (or, "a dice"). The possible values of \( x \) are 1, 2, 3, 4, 5, 6; and each time the die is thrown, the probability of a given score is 1/6.

2.4.2 The Continuous Case

In the continuous case, the uniform distribution is also called the rectangular distribution because of the shape of its probability density function (see below). It is parameterized by the smallest and largest values that the uniformly-distributed random variable can take, \( a \) and \( b \). The probability density function of the uniform distribution \( U(a, b) \) is thus:

\[
f(x; a, b) = \begin{cases} 
\frac{1}{b-a}, & a \leq x \leq b \\
0, & \text{otherwise}
\end{cases}
\]

and the cumulative distribution function is:

\[
F(x; a, b) = \begin{cases} 
0 & \text{for } x < a \\
\frac{x-a}{b-a} & \text{for } a \leq x \leq b \\
1 & \text{for } x > b
\end{cases}
\]

The expected value, variance, and skewness of a random variable following \( U(a,b) \) distribution are:
The standard uniform distribution is the continuous uniform distribution with \( a=0, b=1 \).

2.4.3 Sampling from a Uniform Distribution

In this thesis, Monte Carlo simulation is used to demonstrate the proposed bootstrap method for verification of the CLT. Many programming languages have the ability to generate pseudorandom number sequences which are effectively distributed according to the standard uniform distribution \( U(0, 1) \). If \( u \) is a value sampled from the standard uniform distribution, then the value \( a + (b - a)u \) follows the uniform distribution \( U(a, b) \).

2.4.4 Uses of the Uniform Distribution

Although the uniform distribution is not commonly found in nature, it is very useful for sampling from other specified distributions. A general method is the probability integral transform method, which uses the cumulative distribution function (CDF) of the target random variable. This method is very useful in theoretical work. Since simulations using this method require inverting the CDF of the target variable, alternative methods have been devised for the cases where the CDF is not known in closed form. One such method is rejection sampling.

2.5 Inverse Transform Sampling Method

The *inverse transform sampling method* is a method of generating random numbers from any probability distribution, given its cumulative distribution function (CDF). The method is based on the following well-known result:

\[
E(X) = \frac{a+b}{2} \\
Var(X) = \frac{(b-a)^2}{12} \\
Skewness = 0
\]
Let X be a random variable whose distribution can be described by the cdf F(x). Then the transformation $Y = F(X) \sim U(0, 1)$.

The inverse transform sampling method works as follows:

(1) Generate a random number $u \sim U(0, 1)$.

(2) Set $F(x) = u$ and solve for $x$: $x = F^{-1}(u)$ has the cdf $F(x)$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
CHAPTER 3

MODIFIED BOOTSTRAP CLT VERIFICATION

In this chapter, the details of the proposed method for verification of the CLT are given. Chapter 4 consists of results obtained for several simulated examples. The problem considered in this thesis is as follows:

Given a random sample \( \{x_1, x_2, \ldots, x_n\} \), verify if the probability distribution of \( \bar{x} \) is approximately normal, i.e., verify if the sample size \( n \) is large enough for the CLT to hold.

3.1 The Modified Bootstrap Method

The steps of the proposed method are given below:

1) Select a random sample of size \( m \leq n \) with replacement from the given sample \( \{x_1, x_2, \ldots, x_n\} \). Let this sample be denoted by \( \{x^*_1, x^*_2, \ldots, x^*_m\} \).

2) Calculate \( \bar{x}_m^* \), the mean of the above bootstrap sample.

3) Repeat the step 1 - 2 a large number (say B) times, and obtain \( \{\bar{x}_m^*: k = 1, 2, \ldots, B\} \).

4) Plot the sample histogram of the bootstrap means \( \{\bar{x}_m^*: k = 1, 2, \ldots, B\} \). If th
histograms appear to be Gaussian, then the CLT does not seem to hold for the given sample.

The above algorithm was programmed in C#.NET. The software package developed for this project has two modes:

1) Determination Mode: In this mode, the program can be used to verify whether the CLT holds for a random sample. Figure 5 shows the main screen of the program in the determination mode, in which Total Sample Size is \( n \) (the size of the input sample), Sub-sample Size is \( m \) and Simulation size is \( B \), the number of bootstrap runs.

2) Simulation Mode: In this mode, the program will generate, from one of the probability models of the previous Chapter, random samples of specified size. The sample size can be increased to a specified value. Figure 6 shows the Main Screen in this mode, By running the program in Simulation Mode we demonstrate that the CLT holds for small sample sizes if the underlying distribution is symmetric, and that as skewness is increased, large sample sizes are necessary for the CLT to hold.
Figure 5  Main Screen in Determination Mode

Figure 6  Main Screen in Simulation Mode
CHAPTER 4

SIMULATION MODE

The software program can be run in Simulation Mode to get an idea of how large \( n \) must be before the CLT holds.

4.1 Results From Simulation Mode

The program was run in Simulation Mode for the log-normal, Weibull, and Gamma distributions. These three probability models are chosen for this work as they can be used as viable models for data sets of varying levels of skewness. The following table shows the probability models of varying levels of skewness used in this thesis:

<table>
<thead>
<tr>
<th>Skewness</th>
<th>Log-Normal</th>
<th>Weibull</th>
<th>Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.168097</td>
<td>( \sigma = 0.05593 )</td>
<td>( \mu = 5 )*</td>
<td>( \beta = 3 )</td>
</tr>
<tr>
<td>0.358618</td>
<td>( \sigma = 1.1856 )</td>
<td>( \mu = 5 )*</td>
<td>( \beta = 2.5 )</td>
</tr>
<tr>
<td>1.999982</td>
<td>( \sigma = 5.5138 )</td>
<td>( \mu = 5 )*</td>
<td>( \beta = 1 )</td>
</tr>
<tr>
<td>11.35274</td>
<td>( \sigma = 1.20523 )</td>
<td>( \mu = 5 )*</td>
<td>( \beta = 0.4 )</td>
</tr>
<tr>
<td>190.1084</td>
<td>( \sigma = 1.86243 )</td>
<td>( \mu = 5 )*</td>
<td>( \beta = 0.2 )</td>
</tr>
</tbody>
</table>

* This parameter does not affect the skewness of the distribution.

Appendices show the screen captures of program outputs. The following table summarizes the results obtained from the program runs in the simulation mode, for the three skewed probability models considered in this thesis.
Table 2 Number of samples $n$ needed for CLT to hold

<table>
<thead>
<tr>
<th>Skewness</th>
<th>Log-Normal</th>
<th>Weibull</th>
<th>Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.168097</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>0.358618</td>
<td>$\approx$ 40</td>
<td>$\approx$ 40</td>
<td>$\approx$ 40</td>
</tr>
<tr>
<td>1.999982</td>
<td>85</td>
<td>$&gt;150$</td>
<td>$\approx$ 85</td>
</tr>
<tr>
<td>11.35274</td>
<td>150</td>
<td>$\approx$ 200</td>
<td>175</td>
</tr>
<tr>
<td>190.1084</td>
<td>$&gt;200$</td>
<td>$&gt;200$</td>
<td>$&gt;200$</td>
</tr>
</tbody>
</table>

4.2 Conclusions Drawn From Simulation Mode Results

The following conclusions can be drawn from the Table 2:

1) The number of samples $n$ required for the CLT to hold for the sample of size $n$ does not seem to depend on the form of the probability distribution.

2) A small number $n$ of samples is sufficient for the CLT to hold, when the skewness is low.

3) When skewness is very large, samples of sizes as large as 200 are not enough to guarantee that the CLT holds.
CHAPTER 5

DETERMINATION MODE

The software program can be run in determination mode to determine whether the CLT holds true for any given sample of size \( n \).

5.1 Results from Determination Mode

In this chapter, results obtained from the determination mode of the program are given for two examples.

5.1.1 Determination Mode: Vehicle emission data

Vehicle emission data (from Gilbert, 1987; p. 144) for CO (gms/mile) obtained for 45 randomly selected vehicles is given below:

<table>
<thead>
<tr>
<th>Table 3 Vehicle emission data for CO (gms/mile)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.01</td>
</tr>
<tr>
<td>14.67</td>
</tr>
<tr>
<td>8.6</td>
</tr>
<tr>
<td>4.42</td>
</tr>
<tr>
<td>4.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4 Descriptive statistics computed for CO data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>CO</td>
</tr>
</tbody>
</table>
Figure 7 shows the output from the determination mode for sub-sample size $m = 10$. In other words, the estimated sampling distribution of the mean when sample size is 10 is shown in this figure. The sampling distribution appears to be asymmetric, and hence non-normal.

Figure 7 Output from the determination mode for data in Table 3 with $m = 10$

Figures 8-9 show the output from the determination mode for sub-sample size $m = 25, 45$ respectively. The sampling distribution still appears to be asymmetric, and hence non-normal. In other words, the sample size of 45 is not enough to guarantee the CLT for data of Table 3, and hence normal-theory based statistical procedures for the true population mean should not be used for this dataset. This is consistent with our findings in Chapter 4.
Figure 8  Output from the determination mode for data in Table 3 with m = 25

Figure 9  Output from the determination mode for data in Table 3 with m = 45
5.1.2 Determination Mode: EPA Guidance Document

Table 5  EPA Guidance Document Dataset (1996)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>82.39</td>
<td>105.52</td>
<td>86.62</td>
</tr>
<tr>
<td>103.46</td>
<td>98.37</td>
<td>91.72</td>
</tr>
<tr>
<td>104.93</td>
<td>113.23</td>
<td>108.21</td>
</tr>
</tbody>
</table>

The output from the determination mode for sub-sample size \( m = 9 \) is shown in Figure 10. The distribution of the sample mean appears to be normal for this dataset.

Figure 10  Output from the determination mode for data in Table 5 with \( m = 9 \)
CHAPTER 6

CONCLUSIONS

The fact that the sampling distribution of the mean of a sample of size \( n \) is approximately normal is a well-know result (CLT). What is not known is that how large must the sample size \( n \) be for this result to hold. The quantification of this critical sample size was the goal of this thesis. Results shown in Chapter 4 clearly demonstrate that the thesis has achieved its goal. The computer program developed for this thesis can be run in the simulation mode to get an idea of how large \( n \) must be before CLT holds. The computer program can be run in the determination mode for a given sample, to determine if the sample mean of the given sample will have an approximate normal distribution. If the program run in the determination mode shows that the distribution of the sample mean of size \( n \) is approximately normal, then the standard statistical procedures based on the t-distribution can be used for testing hypotheses and confidence interval estimation problems for the given sample.
REFERENCES


VITA

Graduate College
University of Nevada, Las Vegas

Suresh Kumar Veluchamy

Local Address:

4214 Chatham Circle #3
Las Vegas, NV 89119

Home Address:

41/6 Third Main Road
Gandhi Nagar, Adyar
Chennai, India

Degrees:

Master of Science Degree in Computer Science, 2001
University of Nevada, Las Vegas

Bachelor in Computer Science and Engineering, 1999
College of Engineering, Guindy, Anna University, India

Thesis Title:

A Graphical Approach for Verification of the Central Limit Theorem

Thesis Examination Committee:

Chairperson, Dr. Ashok K. Singh, Ph.D.
Committee Member, Dr. Dieudonné Phanord, Ph.D.
Committee Member, Dr. Rohan Dalpatadu, Ph.D.
Graduate Faculty Representative, Dr. Laxmi P. Gewali, Ph.D.