On Assessing Large Sample Properties of Estimators: An Empirical Approach

Khairul Islam and Sadia Sarker

Department of Mathematics and Statistics
Eastern Michigan University, Ypsilanti, MI 48197

Abstract

Large sample properties such as law of large number, central limit theorem, asymptotic probability distribution, etc. play a significant role in statistical education. Given students' background and mathematical skills, these properties are often very difficult to communicate with students theoretically at lower level statistics classes. This article addresses communicating these properties by simulation approaches, which enhance student's understanding and learning without any ambiguity and misconceptions.

Keywords: Law of large number, central limit theorem, asymptotic distribution, large sample properties, simulation.

1. Introduction

Many statistical procedures are stated as large sample properties. For examples, the law of large number, central limit theorem, the asymptotic sampling distribution of estimators (e.g., sampling distribution of mean and proportion), etc. [3-4] are widely used in statistical education. These properties are parts of many statistical inference procedures. While their mathematical proofs are very difficult or beyond the scope of the classes, some empirical evidences should be provided to students to better understand these concepts.

This article has been intended to communicate the large sample properties empirically via simulations by adapting an open source statistical software R [1-2]. Where appropriate, we provide sample code that are being used in teaching simulation approach using R.

2. Methodology

The following large sample concepts are of great use in introductory statistics classes.

- Law of large number for proportion
- Sampling distribution of proportion
- Law of large number for mean
- Sampling distribution of mean

2.1 Law of large number for proportion

The law of large numbers states that as the sample size gets larger and larger, an estimate of a population parameter gets closer and closer to the parameter. According to this law, the sample proportion gets closer to the population proportion as the sample size gets larger. The proof of this law is beyond the scope of elementary or introductory statistics

classes. The necessity is then to communicate this law empirically via simulation to provide some evidence supporting this property, which does not require any mathematical skills.

Let us consider tossing of a coin a finite number of times, and estimate the proportion of successes π by the sample proportion p. How do we provide evidence that as n gets larger and larger, p gets closer and closer to π as a consequence of the law of large number?

In real-life, how feasible is it to repeat the tossing of a coin a large number of times? Say, for example, we wish to toss the coin 5,000 times. It is very inconvenient to do this and keep the records of outcomes favoring a success and compute the proportion of successes, thereby. However, with simulation, it is an easy task, particularly with the help of the software R. It is possible to generate the proportion of successes over a sequence of repetitions of the sample size n and then to characterize associated properties.

An R version [1-2] of the coin with two possible outcomes "H" and "T" may look like:

```
coin=c("H","T")
```

Suppose getting an "H", while tossing the coin, refers to a success.

Drill Problems

- Toss a coin 20 times and report the output.

```
sample(coin,20, rep=T)
```

- Toss a coin 20 times and count the number of successes.

```
sum(sample(coin,20, rep=T)=="H")
```

- Toss a coin 20 times and find the proportion of successes.

```
sum(sample(coin,20, rep=T)=="H")/20
```

- Execute the code for computing proportion of successes from 20 tosses 5 times and report the proportion of successes.

```
sum(sample(coin,20, rep=T)=="H")/20
[1] 0.25
sum(sample(coin,20, rep=T)=="H")/20
[1] 0.65
sum(sample(coin,20, rep=T)=="H")/20
[1] 0.45
sum(sample(coin,20, rep=T)=="H")/20
[1] 0.35
sum(sample(coin,20, rep=T)=="H")/20
[1] 0.50
```

The first question that might be asked to students from here:
What do you observe and learn throughout this process?

The answer to this question is invaluable—the outcome of the execution is a random variable.

- Perform above execution using a loop:

```
for (i in 1:5){
sum(sample(coin,20, rep=T)=="H")/20}
```

An execution of the above code computes proportion of successes from 20 tosses 5 times, but it does not print or save the result of the execution. To print or save the result of above execution, we need to instruct results to be printed or stored:

- Print results of execution of a loop:

```
for (i in 1:5){
print(sum(sample(coin,20, rep=T)=="H")/20)}
```

- Store 5 proportions from above execution of the loop:

```
storage<-c()
for (i in 1:5){
storage[i]=sum(sample(coin,20, rep=T)=="H")/20}
```

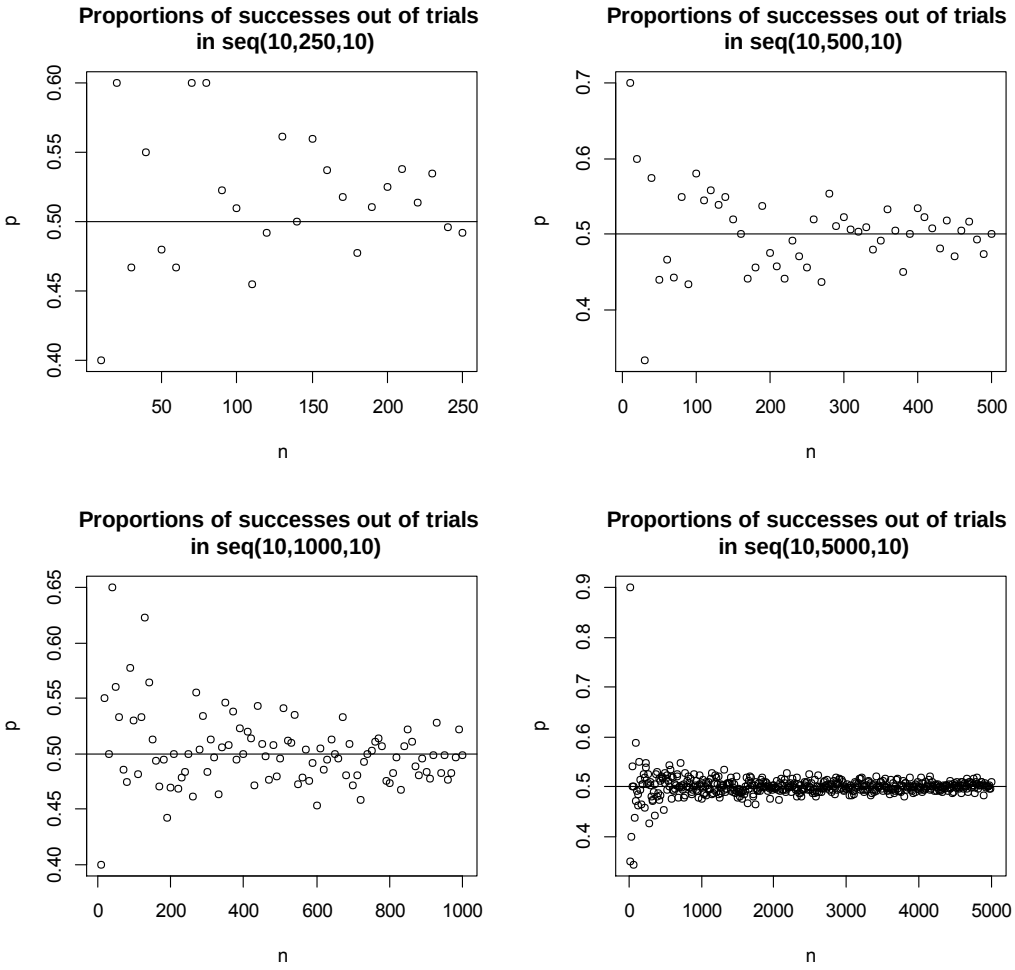
By storing the results of an execution of the loop, one can re-use the result for graphing or other intended computations. Note that an empty **storage** is defined using the function c() for this purpose, which gets updated by including results of the execution throughout the process.

Putting them all together, we can verify the law of large number for proportion via simulation and graphical representation by executing the following chunk of R codes:

```
p=c();
n=seq(10,250,10);
L=length(n);
for (i in 1:L){p[i]=mean(sample(coin,n[i], rep=T)=="H")}
plot(n,p, main="Proportions of successes out of trials \n
in seq(10,250,10)")
abline(h=0.5)
```

Using above code, we estimate proportion of successes from a sequence of number of tosses between 10 and 250, with an increment of 10. The main objective is to note the behavior of proportions of successes as n increases. To make the behavior of proportion of successes with the increase in sample size n notable and visible, the proportions of successes so obtained are plotted against the number of tosses (i.e., the sample sizes n) in Figure 1.

Figure 1. Proportions of successes plotted against number of trials in varying sequence



Above figures indicate that the estimates of the proportion stabilize around the true value of $\pi = 0.50$ as n gets larger and larger, which is the lesson of the law of the large number applied to the proportion of successes.

2.2 Sampling distribution of proportion

For a population with proportion π, the probability distribution of the sample proportions p over all possible samples of a given size n is called the sampling distribution of p. The mean and standard deviation of the sampling distribution of p over all possible samples of size n, denoted by μ_p and σ_p, respectively, are given by

$$\mu_p = \pi$$

$$\sigma_p = \sqrt{\frac{\pi \times (1-\pi)}{n}}$$

The standard deviation of the sampling distribution of a statistic is generally termed as the standard error (S.E.). Therefore, one can write

$$\sigma_p = S.E.(p) = \sqrt{\frac{\pi \times (1-\pi)}{n}}$$

The sampling distribution of p is approximately normal with mean μ_p and standard deviation σ_p provided $n\pi \geq 10$ and $n(1-\pi) \geq 10$, and the approximation gets better and better as n gets larger and larger. This property is called the central limit theorem for proportion p.

While the central limit theorem is of great use in statistical inference via confidence interval estimate and tests of hypotheses regarding proportions, the proof is beyond the scope of elementary or introductory statistics class. Therefore, an ideal approach could to make some effort to provide some evidence in the support of this statement via simulation and graphical representation.

To make this effort realized, let us consider samples of sizes n=100, 200, 500, 1000 and investigate the sampling distribution of the proportion p over 10,000 samples and explore how approximate the distribution is to the normal population with specified mean and standard error.

Note that the true sampling distribution of p for a sample of size n will have mean and standard error as follows:

$\mu_p = \pi = 0.50$ and $\sigma_p = \sqrt{\frac{\pi \times (1-\pi)}{n}} = \sqrt{\frac{0.5 \times 0.5}{n}}$

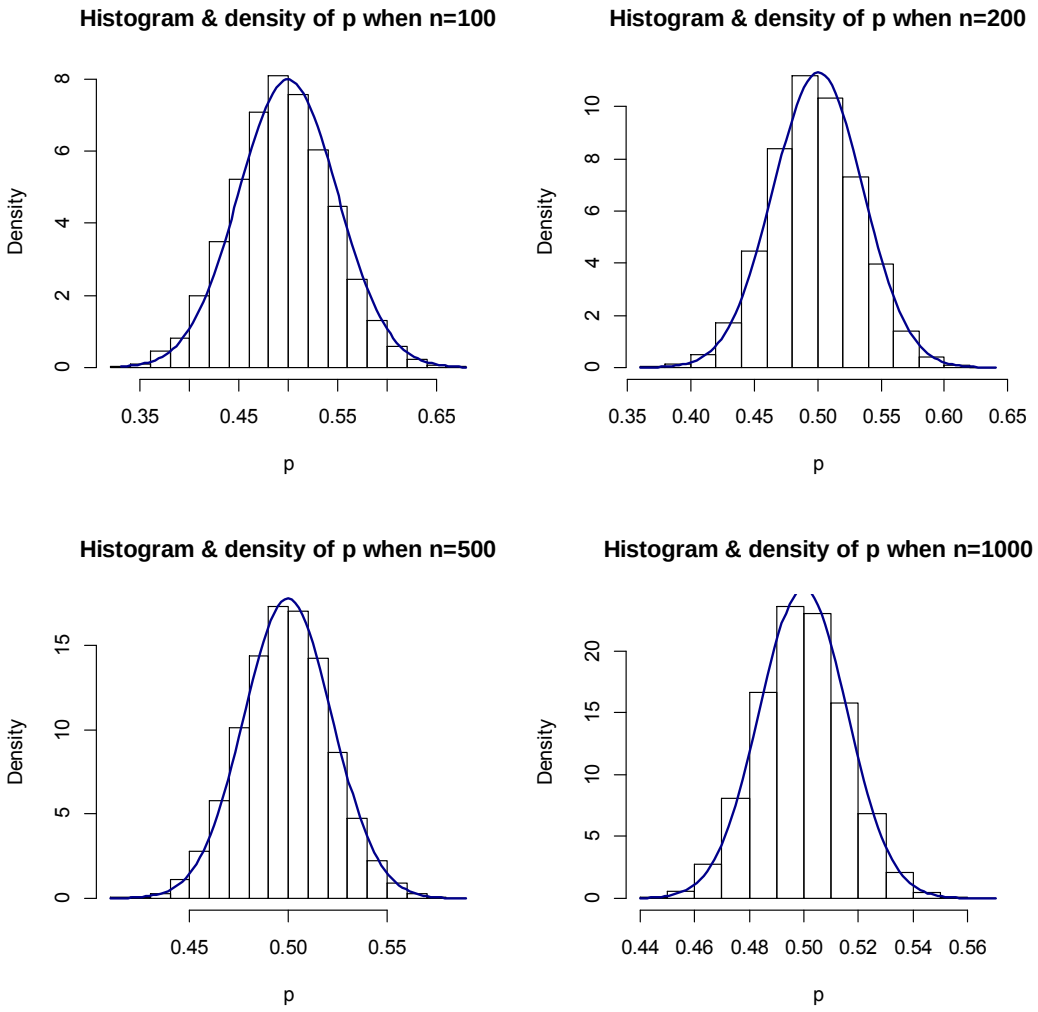
The table below summarizes true mean (μ_p) and standard error (σ_p) and their estimated values ($\hat{\mu}_p$ and $\hat{\sigma}_p$) from simulated sampling distribution of p for sample sizes n =100, 200, 500, and 1000 over 10,000 samples.

n	μ_p	$\hat{\mu}_p$	σ_p	$\hat{\sigma}_p$
100	0.50	0.49963	0.05000	0.04994
200	0.50	0.50044	0.03536	0.03524
500	0.50	0.49984	0.02236	0.02244
1000	0.50	0.49976	0.01581	0.01594

Figure 2 below is graphical representation of the histogram and density plot corresponding the simulated sampling distribution with the configuration of the above table.

It is evident from Figure 2 that the histograms and superimposed normal density plots with parameters as simulated estimates seem approximately normal. Also, as n gets larger, the variability over simulated sampling distribution gets smaller and thereby approximation might get better and better with the increasing n. The graphical visualization of Figure 2 leaves no doubt that the sampling distribution of p, as n gets larger and larger, is approximately normal.

Figure 2. Histogram of sampling distribution of proportion of successes for varying sample sizes



2.3 Law of large number (LLN) for sample mean

Given a sample of size n from a population with mean μ, the sample mean gets closer and closer to the parameter μ as the sample size gets n larger and larger.

Let us verify the LLN for simulation from a normal distribution with mean 10 and standard deviation 5, and a discrete Poisson distribution with mean 10, denoted by P(10).

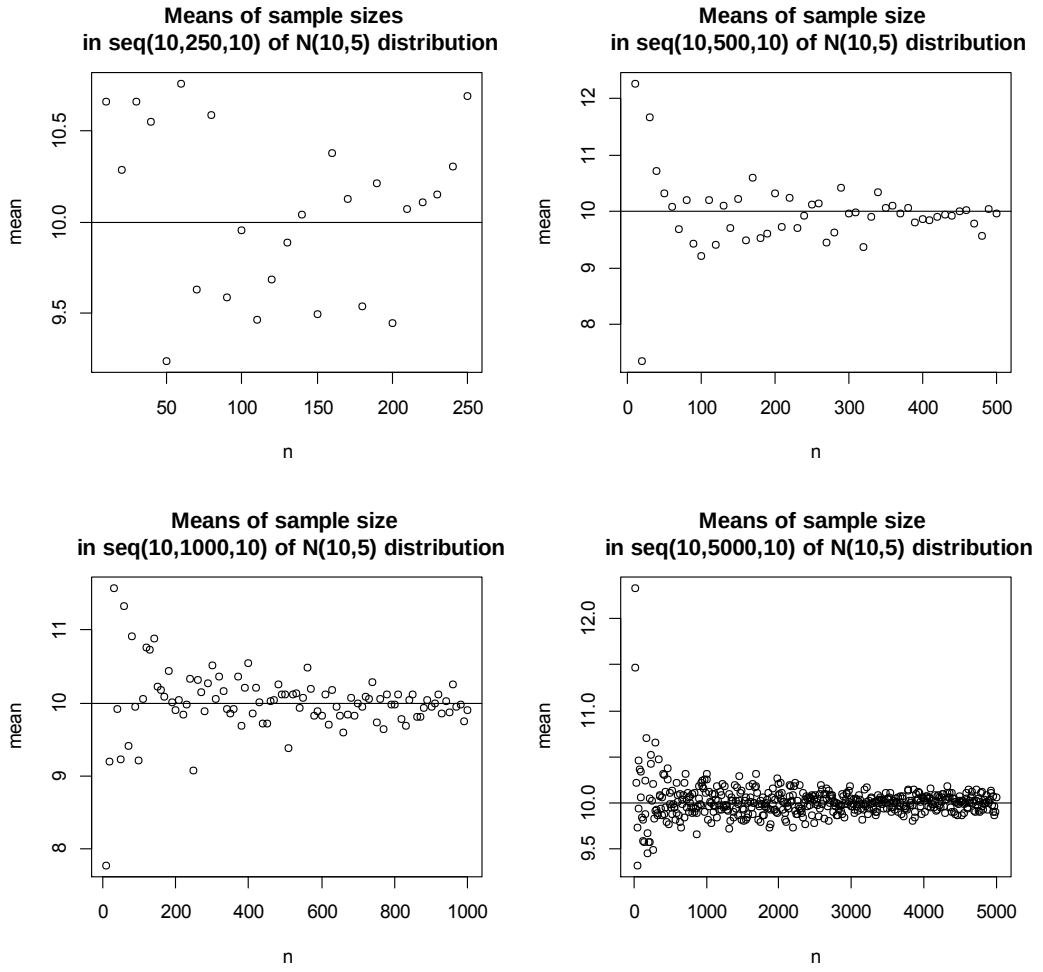
To simulate a sample of size n from a normal distribution with mean 10 and standard deviation 5, we use the rnorm(n, mean=10, sd=5) function available in R:

```
mean=c();
```

```
n=seq(10,250,10);
L=length(n);
for (i in 1:L){mean[i]=mean(rnorm(n[i], mean=10,sd=5))}
plot(n,mean, main="Means in samples of size \n in
seq(10,250,10)")
abline(h=10)
```

In Figure 3, we plot simulated means from N(10,5) distribution for varying sample sizes.

Figure 3. Means of samples from N(10, 5) distribution plotted against sample sizes in varying sequences



As appears in Figure 3, the estimates of the mean μ form N(10,5) distribution stabilize around the true value of $\mu = 10$ as n gets larger and larger, which is the lesson of the law of the large number applied to the estimate of the mean.

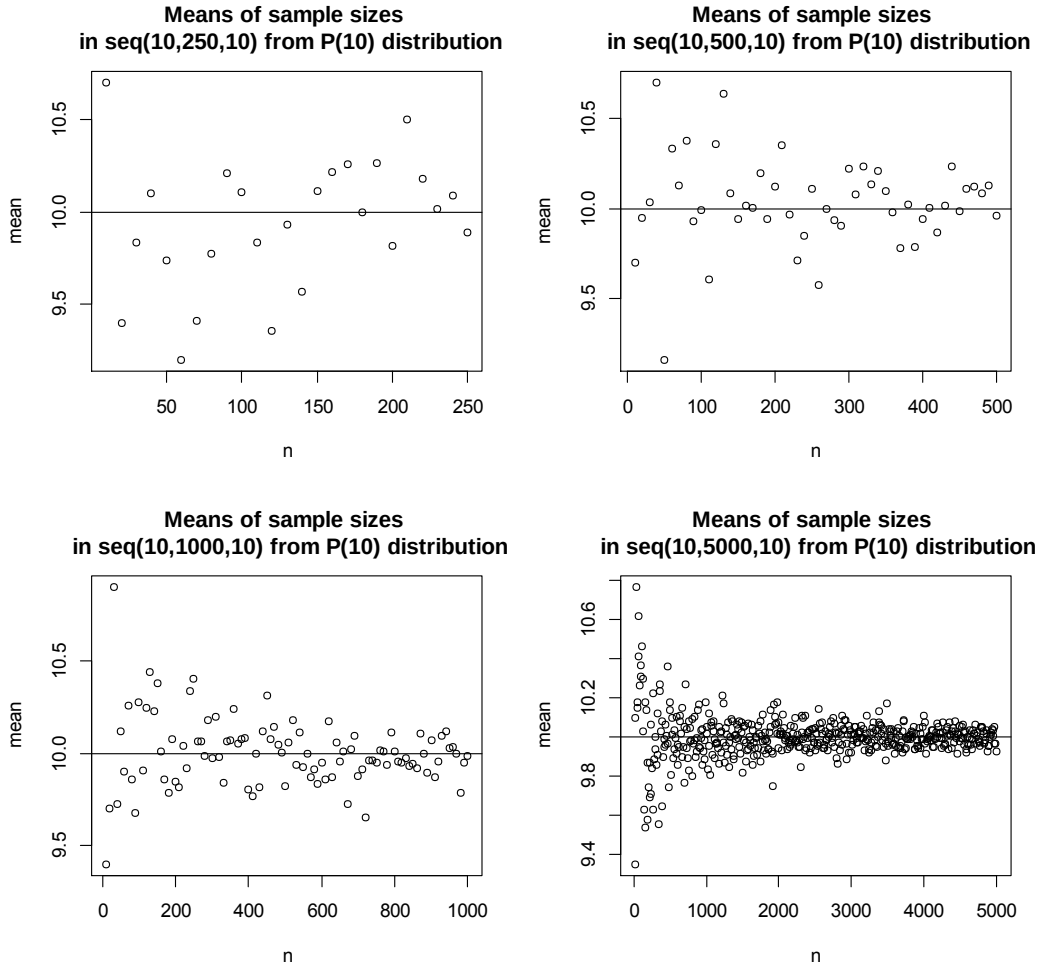
In order to simulate a sample of size n from a discrete Poisson distribution with mean 10, we use R function rpois(n, 10) and execute the following code:

```
set.seed(1)
```

```
par(mfrow = c(2, 2))
mean=c();
n=seq(10,250,10);
L=length(n);
for (i in 1:L){mean[i]=mean(rpois(n[i], 10))}
plot(n,mean, main="Means of sample sizes \n in seq(10,250,10) from P(10)")
abline(h=10)
```

In Figure 4, we demonstrated plots of simulated means from P(10) distribution for varying sample sizes.

Figure 4. Means of samples from P(10) distribution plotted against sample sizes in varying sequences



Again, as appears in Figure 4, the estimates of the mean μ form P(10) distribution stabilize around the true value of $\mu = 10$ as n gets larger and larger, and thus simulated means

conform to the law of the large number applied to the estimate of the mean from P(10) distribution.

2.4 Sampling distribution of mean

The probability distribution of the sample mean $\bar{x}$ over all possible samples of a given size is called the sampling distribution of $\bar{x}$. For a population with mean μ and the standard deviation σ, the mean $\mu_{\bar{x}}$ and standard deviation (i.e. standard error) $\sigma_{\bar{x}}$ of the sampling distribution of $\bar{x}$ over all possible samples of size n are:

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The sampling distribution of $\bar{x}$ is either (a) normal if the population the samples come from is normal, irrespective of the sample size n, or (b) approximately normal for large sample size n if the population the samples come from is not normal. This property is what we call the Central Limit Theorem for $\bar{x}$.

The table below summarizes true mean (μ_p) and standard error (σ_p) and their estimated values ($\hat{\sigma}_p$ and $\hat{\sigma}_p$) from simulated sampling distribution for n =100, 200, 500, 1000 from N(mean=5, sd=5) distribution.

n	$\mu_{\bar{x}}$	$\hat{\mu}_{\bar{x}}$	$\sigma_{\bar{x}}$	$\hat{\sigma}_{\bar{x}}$
25	10	9.9943	1.0000	0.9939
50	10	9.9976	0.7071	0.7066
100	10	10.0007	0.5000	0.5039
500	10	10.0018	0.2236	0.2264

The following code has been implemented to generate 10,000 sample of size n=25.

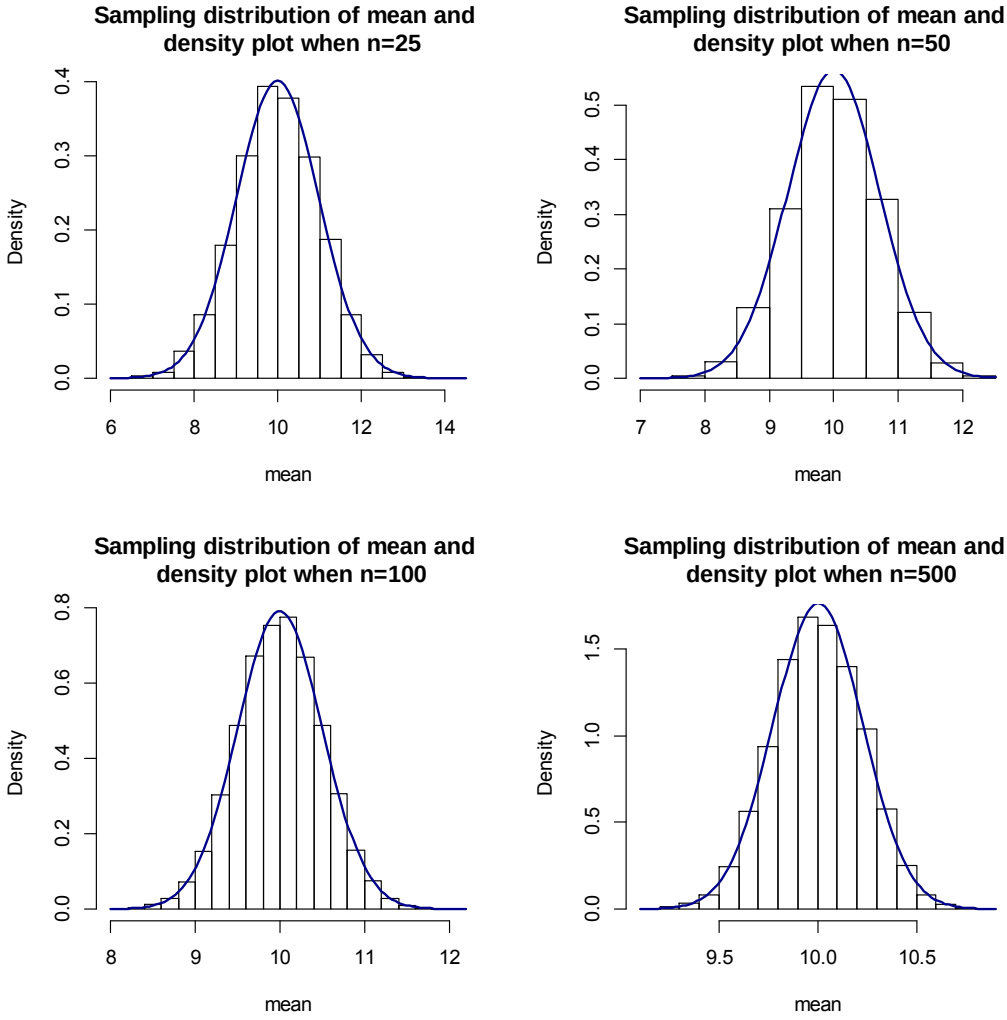
```
set.seed(1)
par(mfrow = c(2, 2))
mean=c();
for (i in 1:10000){mean[i]=mean(rnorm(25, mean=10,sd=5))}
hist(mean, freq=F,main="Sampling distribution of mean and \n
density plot when n=25")
curve(dnorm(x,     mean(mean),     sd(mean)),     add=TRUE,
col="darkblue", lwd=2)
round(c(mean(mean), sd(mean)),digits=5)
```

In Figure 5, we provide a graphical representation of the histogram and density plot corresponding the simulated sampling distribution from a N(mean=10,sd=5) population for varying sample sizes based on 10,000 simulated samples.

From Figure 5, it follows that the histograms and superimposed normal density plots with parameters as simulated estimates seem approximately normal. Also, as n gets larger, the

variability over simulated sampling distribution gets smaller and thereby approximation might get better and better with the increasing n.

Figure 5. Histograms and density plots of sampling distributions of means from N(mean=10, sd=5) for varying sample sizes.



3. Conclusion

Introducing large sample properties in elementary statistics class is a challenging task. Discussing large sample properties without any evidence to support them leave many students confused with underlying large sample concepts. Note that large sample properties are being used for making inference regarding unknown population mean and proportion. While providing a theoretical evidence is not possible, students can have empirical evidence of all large sample properties via simulation, to some extent. It can be used to provide other large sample ideas such as consistency of estimators [5-7], as required in the curricula. For example, theoretically speaking, an estimator $\hat{\theta}_n$ of the parameter θ is consistent if $\lim_{n\to\infty} E[\hat{\theta}_n] = \theta$ and $\lim_{n\to\infty} Var[\hat{\theta}_n] = 0$, the proof of which is beyond the scope of the class. Such property can be presented empirically via simulation with very little effort from the instruction. Particularly, with an open source software such as R, such

efforts seem to be an easy task. It provides students with necessary introduction to the programming skill early, and they can explore many interesting facts that raise further questions towards enhancing learning.

References

[1]. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[2]. Verzani, J. 2005. Using R for Introuctory Statistics. Chapman & Hall/CRC.
[3]. Moore, D. 2009.The Basic Practice of Statistics. Fifth Edition, Freeman, W. H. & Company.
[4]. Bowerman, B.L., Orris, J.B. and Murphree, E. 2014. Essential of Business Statistics. Fifth Edition, McGraw-Hill Education.
[5]. Casella, G. and Berger, R.L. 2002. *Statistical Inference*. Second Edition, Duxbury.
[6]. Hogg, R.V., McKean, J.W. and Craig, A.T. 2013. *Introduction to Mathematical Statistics*. Seventh Edition, Prentice Hall.
[7]. Rosner, B. 2016. Fundamentals of Biostatistics. Eight Edition, Cengage Learning.