Research Article

Sampling Distribution and Simulation in R

Tofik Mussa Reshid
Department of Statistics, Werabe University, P.O. Box 46 Werabe, Ethiopia
Email: toffamr@gmail.com Tel: +251911878075

Simulation plays important role in many problems of our daily life. There has been increasing interest in the use of simulation to teach the concept of sampling distribution. In this paper we try to show the sampling distribution of some important statistic we often found in statistical methods by taking 10,000 simulations. The simulation is presented using R-programming language to help students to understand the concept of sampling distribution. This paper helps students to understand the concept of central limit theorem, law of large number and simulation of distribution of some important statistic we often encounter in statistical methods. This paper is about one sample and two sample inference. The paper shows the convergence of t-distribution to standard normal distribution. The sum of the square of deviations of items from population mean and sample mean follow chi-square distribution with different degrees of freedom. The ratio of two sample variance follow F-distribution. It is interesting that in linear regression the sampling distribution of the estimated parameters are normally distributed.

Key Words: simulation, central limit theorem, law of large number, normal distribution, t-distribution, chi-square distribution, F-distribution, Regression

INTRODUCTION

Simulation plays important role in many problems of our daily life. By simulation we can do a lot and it is the heart of statistics. When some experiments are conducted, its validity can be checked by using simulation. Sometimes empirical solutions are unattainable in such case we solve by using simulations. This paper discusses how computer simulations are employed for sampling distribution and to make inference. The objective of this paper is to show how to make inference for one sample and two sample data. It provides researchers and students with brief description of commonly used statistical distributions using simulation. The paper tries to prove the distribution of some important statistic by using computer simulation. As computers become more readily available to educators, there is wide speculation that teaching inference via dynamic, visual simulations may make statistical inference more accessible to introductory students (Moore, 1997).

In statistical method we teach students about sampling distribution of statistic, as a matter of fact sampling distribution is necessary for statistical inference. Using simulation to teach the sampling distribution of the mean is widely recommended but rarely evaluated (Ann E. Watkins et al, 2014). Students find the concept of sampling distribution difficult to grasp. Students find the concept of sampling distribution, specifically, the sampling distribution of the mean and central limit theorem, difficult to understand (Ann E. Watkins et al, 2014). When we Come to regression students find more difficult the concept of sampling distribution. This paper presents simulation studies focused on the difficulties students experience when learning about sampling distribution. It gives good understanding for students, researchers and teachers about sampling distribution and central limit theorem using simulation.

In this paper it is tried to explain the simulation of sampling distributions of some important statistic we often encounter in statistical methods. One way to use simulations is to allow students to experiment with a simulation and to discover the important principles on their own (David M. Lane 2015). To compute population parameters, the entire population needs to be known. However, in many cases it is difficult to find the entire population items. In this case inference can be drawn based on sample statistic.

Simulation in this paper is such that drawing samples in an experiment over and over again it tends to reveals certain pattern. We try to check the validity of central limit theorem and law of large numbers. Moreover, in this paper it is attempted to see some important sample statistic distributions using simulation.
This paper has provided a tool for sampling distribution and simulation in R programming. It gives a good understanding for students in simulation and R syntax code. Copy the code and paste in to R will run the program. Sometimes simulation is impressive due to its result. for example, the distribution of the sample variance is different when the population mean is unknown. And the distribution of the sample mean is different when the population variance is unknown. It is shown that the convergence of student t-distribution, central limit theorem, law of large numbers, and sampling distribution in regression.

**Statistic and Simulation**

A statistic is a function of sampling units drawn from a population. Any function of random variables of sampling units is also a random variable used to estimate the corresponding population parameter. Let \( u \) denote statistic then it is a function of items in the sample. That is \( u = g(x_1, x_2, \ldots, x_n) \). Sampling distribution of a statistic is the probability distribution of the sample statistic based on all possible simple random samples from the population. Through the use of simulation, we want to demonstrate the properties of sampling distributions of some statistic such as the mean, variance and regression parameters. Simulation in this paper is such that takes \( n \) samples and compute statistic \( u \) and repeat \( N \) times. The simulation procedure follows the following steps.

i. Draw \( n \) samples from a population of mean \( \mu \) and variance \( \sigma^2 \)

ii. Compute statistic \( u \)

For example, \( \bar{x}_n = \sum_{i=1}^n x_i \) and \( s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \)

iii. Repeat \( i \) and \( ii \) \( N \) times, where \( N \) is the number of times a sample is drawn.

iv. Compute the expected value

For example, \( E(\bar{x}_n) = \frac{\sum_{i=1}^n x_i}{n} \) and \( Var(\bar{x}_n) = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{N} = \frac{\sigma^2}{nN} \)

where \( \sigma^2 \) is population variance

Let \( X_1, X_2, \ldots, X_n \) are independently, identically, and normally distributed with mean \( \mu \) and finite variance \( \sigma^2 \) then inference about population mean can be made as follows:

1. If the population variance \( \sigma^2 \) is known the sampling distribution of the sample mean is obtained through normal distribution for any sample size. \( Z = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \) This is called standard normal distribution denoted by \( Z \) (Paul L., 2015)

2. The other case is if the population variance \( \sigma^2 \) is unknown we use student t-distribution. t-distribution arises when the estimate of the mean of a normally distributed population in a situation where the sample size is small and the population variance is unknown.

It is developed by William Seally Gosset (1908). Its density of \( n \) degrees of freedom is

\[
 f(t) = \frac{1}{\sqrt{n \pi}} \frac{1}{\Gamma(\frac{n+1}{2})} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}
\]

Let \( X_1, X_2, \ldots, X_n \) are independently, identically and normally distributed with mean \( \mu \) and variance \( \sigma^2 \) then the sample mean inference can be made through \( T = \frac{\bar{x}_n - \mu}{S/\sqrt{n}} \sim t(n-1) \) this follows student t-distribution (Paul L., 2015).

**Figure 1:** Simulation approaches to standard normal distribution

If the population variance is known the sampling distribution of \( Z \) defined above is standard normal distribution for any sample size. The simulation does not depend on sample size \( n \), only the variance of the sample mean decreases as sample size \( n \) increases because \( Var(\bar{x}_n) = \frac{\sigma^2}{nN} \)

**Figure 2:** Shows the density of \( N=10000 \) simulations from normal distribution by taking a random sample of \( n=5 \) items and statistic \( T \) is computed. The figure shows the statistic \( T \) follows student t-distribution and it compares t-distribution with standard normal distribution. The R code for the simulation is given as:

```r
f=function(N,n){
  m=matrix(0,N);v=m;z=m
  for(i in 1:N){
    x=rnorm(n,8,6);m[i]=mean(x);v[i]=var(x)
    z[i]=(m[i]-8)/6*sqrt(n);return(z)
  }
}
Z=f(10000,20)
plot(density(Z),xlim=c(-5,5),ylab=c(“Z”,lwd=1,col=“blue”,lty=1,main=“”))
lines(sort(Z),dnorm(sort(Z)),col=“green”,lwd=2,lty=2)
Z=f(10000,5)
lines(sort(Z),dnorm(sort(Z)),col=“red”,lwd=3,lty=3)
legend(1.6,0.4,c(“simulation n=20”,“simulation n=5”,“normaldist”),col=c(“blue”,“green”,“red”),lwd=c(“1”,“2”,“2”),lty=c(1,2,3))
```
When the population variance is unknown the statistics $T$ is student's t-distribution with $n-1$ degrees of freedom. However, as the degree of freedom of student's t-distribution increases it goes to standard normal distribution as shown in figure 2. Mathematically the following equation verifies it.

\[
\lim_{n \to \infty} \frac{1}{\sqrt{n}} \left( 1 + \frac{t^2}{n} \right)^{-\frac{n+1}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} 
\]

This is standard normal distribution.

If the population variance is unknown and the sample size is large (most texts take $n>30$) the sampling distribution of $T$ is approximately standard normal distribution. As degrees of freedom increase making inference on t-distribution is the same as making inference using standard normal distribution.

**Law of large number**

Increasing sample size increases precision. That is the estimate close to population parameter with decreasing variability. This is called law of large number (Casella G. et al 2001). \(\lim_{n \to \infty} P (|\bar{X}_n - \mu| = 0) = 1\). This theorem is shown in Figure 3 using simulation.

Figure 3: Shows sample size versus sample mean. When the sample size increases the variability of the sample mean decreases. We draw samples from normal distribution of mean 8 and standard deviation 6. The expectation of the sample mean is equal to the population mean. The R code is given as:

```r
f=function(N){
m=matrix(0,N);L=m;U=m
for(i in 1:N){x=rnorm(i,8,6);
m[i]=mean(x)
L[i]=8-1.96*6/sqrt(i)
U[i]=8+1.96*6/sqrt(i)}
m=data.frame(m,L,U)
return(m)}
m=f(1000)
plot(m[,1],ty="l",col="blue",xlab="sample size",ylab="sample mean")
lines(m[,2],ty="l",col="black")
lines(m[,3],ty="l",col="black")
lines(c(0,1000),c(8,8),lty=1,lwd=2,col="red")
```

Figure 3: The sample mean decreases variability when the sample size increases.

For any sample size the expectation of the sample mean is equal to the population mean. However, increasing sample size decreases variability. For finite population the sampling distribution of the sample mean is exactly equal to the population mean if all \(\binom{N}{n}\) samples are collected. When the population is large it is not possible to take all sampling combinations. In any case increasing sample size the mean of the sample mean is more closer to population mean. (David M. Lane, 2015) found that the difference between the simulated sampling distribution mean and the population mean decreased as a function of sample size. From simulation we found that the mean of the simulation mean is equal to the population mean and the standard deviation of the simulation mean is \(\frac{\sigma^2}{\sqrt{nN}}\). As sample size increases the sample statistic become less variable and more closely estimate the corresponding population parameter (Jennifer Noll et al., 2014).

**Central limit theorem**

Now we will see the distribution of the sample mean when the underlying distribution is not normal. In many cases the population is not normal. Since the central limit theorem is only be stated and not proved evidence of its operation will need to be given to the students.

If we have large sample size coming from mean \(\mu\) and variance \(\sigma^2\) then the sampling distribution of quantity \(Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\) is standard normal distribution without considering the
original distribution of random variable $X_1, X_2, \ldots X_n$. A shape that is normal if the population is normal, for other distributions with finite mean and variance, the shape becomes more normal as $n$ increases (Monica E. Brussolo, 2018).

We try to see the central limit theorem by taking simulation of different non normal distributions. We mentioning some examples such as drawing $n$ samples from binomial, exponential and gamma distributions. Let us draw non normal sample and standardize it. Let us take symmetrical distribution say binomial $n=50$ and $p=0.5$ and another highly skewed pattern say exponential $\lambda = 50$ and gamma distribution shape parameter $\alpha = 30$ and scale parameter $\beta = 0.5$. The samples are drawn from binomial, exponential and gamma distributions.

Figure 4 shows the density of $N=10000$ simulations of a sample $n=100$ for each of binomial, exponential and gamma distribution and statistic $Z$ is computed.

For large simulation, in this case $N=10000$ for all distributions $Z$ turn to standard normal distribution approximately. If our sample size is sufficiently large the sampling distribution goes to standard normal distribution without considering the distribution of the original pattern. Another important statistic we often found in statistical method is the sample variance. The inference made about population variance is through chi-square distribution. Chi-square distribution is one of the most widely used probability distribution in inferential statistics, notably in hypothesis testing and construction of confidence intervals. Chi-square distribution is used in goodness of fit test, test of independency, likelihood ratio test, logarithm rank test and Cochran Mantle Haenszel test.

Let $z_1, z_2, \ldots z_n$ are independent standard normal distributions then the sum of their square is chi square distribution with $n$ degrees of freedom. Suppose $X_1, X_2, \ldots X_n$ are identically, independently normally distributed we need to make inference on population variance, $\sigma^2$. Therefore, we have a quantity $Q = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{s} \right)^2$ which follows chi square distribution with $n-1$ degrees of freedom (R. Lyman Ott et al, 2010).

Figure 5: shows the density of large simulation $N=10,000$ items from normal distribution, for a random sample $n=10$, the sampling distribution of $Q$ is exactly chi-square distribution. The R code is given below

\begin{verbatim}
f=function(N,n){
z=i=1:N{
b=rbn(100,50,0.5)
z[i]=(mean(b)-25)/(sqrt(1.25/100))
return(z)}
z=f(10000)
plot(density(z),xlim=(min(z),max(z)),ylim=c(0.05,0.5),lty=1,
lwd=1,col="black",ylab="z",main="")
lines(sort(z),dnorm(sort(z),col="red",lty=2,lwd=3)
f=function(N){
z=matrix(0,N)
for(i in 1:N)
    z[i]=(mean(b)-0.02)/(sqrt(0.0004/100))
return(z)}
Z=f(10000)
plot(density(z),xlim=(min(z),max(z)),ylim=c(0.05,0.5),lty=1,
lwd=2,col="green",ylab="z")
f=function(N,n,a,b){
g=matrix(0,N)
for(i in 1:N)
g[i]=(mean(b)-a*b)/sqrt(a^2/b^2/n)
return(g)}
g=f(10000,1000,30,0.5)
lines(density(g),col="purple",ylab="z",lty=3,lwd=3)
legend(1.4,0.45,c("bin simulation","exp simulation","gamma simulation","std norm"),lty=c(1,2,3,3),col=c("black","green","purple","red"))

Figure 4: Standardized binomial, exponential and gamma approaches to standard normal distribution.

Figure 5: Simulation of chi-square distribution
\end{verbatim}

As we can see from Figure 5 when the population mean is unknown the sampling distribution of the sample variance loses one degrees of freedom (it is $n-1$). In the above equation, $Q$, if the sample mean is substituted by known population mean the probability distribution is chi-square distribution with a different degrees of freedom.
Let $X_1, X_2, \ldots, X_n$ are identically, independently normally distributed we need to make inference on population variance, $\sigma^2$. Therefore, we have a quantity $Q = \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2$ follows chi square distribution with $n$ degrees of freedom (R. Lyman Ott et al, 2010).

Figure 6: shows the density of $Q$ is chi-square distribution with $n$ degrees of freedom. As we can see from the figure $Q$ is chi-square distribution with $n$ degrees of freedom. The R code is:

```r
f=function(N,n1,n2){
  z=matrix(0,N)
  for(i in 1:N){
    z[i]=(mean(rnorm(n1,8,6)))/(6*sqrt(1/n1+1/n2))
  }
  return(z)
}
```

Figure 7: shows the density of 10,000 simulations from two populations of equal mean and equal known variance. We compute statistic $Z$ for random sample of $n_1 = 10$ and $n_2 = 12$, the simulation probability density is the same as standard normal distribution approximately. The following is the R code of the simulation.

```r
t=matrix(0,N)
  for(i in 1:N){
    t[i]=(mean(rnorm(n1,8,6)))/(6*sqrt(1/n1+1/n2))
  }
  return(t)
}
```

Two sample inference

Sometimes we need to make inference on two populations based on sample. When sample is taken from two groups the process of inference about population mean difference is very similar to one sample inference. Let $X_1, X_2, \ldots, X_n$ iid normal with mean $\mu_1$ and variance $\sigma_1^2$ and $Y_1, Y_2, \ldots, Y_n$ are iid normal with mean $\mu_2$ and variance $\sigma_2^2$ then we need to make inference on $\mu_1 - \mu_2$.

The distribution of the statistic

$$\frac{(\bar{X} - \bar{Y}) - d}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

where $d = \mu_1 - \mu_2$, is depends on whether the population variance is known or unknown (R. Lyman Ott et al, 2010). If the two population has common known variance $\sigma^2$ then the statistic $Z = \frac{(\bar{X} - \bar{Y}) - d}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ is standard normal distribution.
We have shown that when the variance of the two populations is equal and unknown, the density of the simulation is the same as student t-distribution with \( n_1 + n_2 - 2 \) degrees of freedom (R. Lyman Ott et al, 2010). Then \((1 - \alpha)100\%\) CI for \( \mu_1 - \mu_2 \) is

\[
\bar{X} - \bar{Y} \pm t_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

Note: if \( n_1 + n_2 - 2 > 30 \) making inference on both Z and t-distribution is the same.

If the two populations have different and known variance inference is made using standard normal distribution. Interval estimation for mean difference \( \mu_1 - \mu_2 \) when \( \sigma_1^2 \neq \sigma_2^2 \) We make the assumptions of both populations are normally distributed and The samples are independent \( \sigma^2 X - \sigma^2 Y = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \) and the statistic \( Z = \frac{(\bar{X} - \bar{Y}) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \) is standard normal distribution.

Then \((1 - \alpha)100\%\) CI for \( \mu_1 - \mu_2 \) is

\[
\bar{Y}_1 - \bar{Y}_2 \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

If the population variances \( \sigma_1^2 \) and \( \sigma_2^2 \) are unknown and different We estimate the variance of the sample mean difference from sample

\[
\hat{\sigma}^2 \bar{Y}_1 - \bar{Y}_2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}
\]

and then the statistic

\[
T = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}
\]

Where \( v = \frac{(n_1 - 1)(n_2 - 1)}{c_1(n_1 - 1) + c_2(n_2 - 1)} \) degrees of freedom (Lyman et al. 2010).

Where \( c_1 = \frac{S_1^2}{S_1^2 + S_2^2} \) and \( c_2 = \frac{S_2^2}{S_1^2 + S_2^2} \) note \( c_1 + c_2 = 1 \)

Figure 8: simulation of t-distribution for two population.

We make the difference \( d \) for each data pair. \( d_i = X_i - Y_i \)

The mean of the differences is \( \bar{d} = \frac{\sum d_i}{n} \) and \( s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} \)

The distribution of \( T = \frac{\bar{d} - \bar{d}}{s_d/\sqrt{n}} \)

Figure 10: the density of \( N=10000 \) simulation density for sample size \( n=10 \) from two normal population is identical to student’s t-distribution

The following R syntax code can do the simulation

```r
f=function(n){
  t=matrix(0,n)
  for(i in 1:n){
    x=x1+rnorm(n1,8,6)
    y=x2+rnorm(n2,8,4)
    d=x-y
t[i]=(mean(x1)-mean(x2))/(sqrt(var(x1)/n1+var(x2)/n2))
  return(t)
}
t=f(10000)
plot(density(t),col="red",lwd=1,lt=1,xlab="t")
n1=10;n2=12
c1=36/(10)/(36/10+16/12)
c2=1-c1
df=(n1-1)*(n2-2)/((n1-1)*c1^2+(n2-1)*c2^2)
lines(sort(t),dt(sort(z),(df)),col="green",lwd=2,lt=2)
legend(1.2,0.36,c("simulation","t-dist(17.7df)"),col=c("red","green"),lwd=c("1","2"))
```

Figure 9: the density of simulations compared with t-distribution

Figure 9 shows the density of 10,000 simulations is approximately the same as students t-distribution with \( \nu \) degrees of freedom (R. Lyman Ott et al, 2010). Then \((1 - \alpha)100\%\) CI for \( \mu_1 - \mu_2 \) is

\[
\bar{X} - \bar{Y} \pm t_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}
\]

Note: if \( \nu > 30 \) then \( t(\nu) \approx Z \) therefore making inference on both is the same.

**Paired sample**

When we have paired samples say \( X \) and \( Y \) we need to make inference on their difference mean. We will first find the difference \( d \) for each data pair. \( d_i = X_i - Y_i \)

The mean of the differences is \( \bar{d} = \frac{\sum d_i}{n} \) and \( s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} \)

The distribution of \( T = \frac{\bar{d} - \bar{d}}{s_d/\sqrt{n}} \)

Figure 10: the density of \( N=10000 \) simulation density for sample size \( n=10 \) from two normal population is identical to student’s t-distribution

The following R syntax code can do the simulation

```r
f=function(n){
  t=matrix(0,n)
  for(i in 1:n){
    x=x1=morm(n1,12,6)
    y=x2=morm(n2,12,8)
    d=x-y
t[i]=(mean(x1)-mean(x2))/(sqrt(var(x1)/n1+var(x2)/n2))
  return(t)
}
t=f(10000)
```
The estimate of \( \rho \) from the sample is \( r \) defined as
\[
\rho = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2} \sqrt{\sum_{i=1}^{n} y_i^2 - n \bar{y}^2}}
\]

Let \( X_1, X_2, \ldots, X_n \) be a simple random sample from a normal distribution of mean \( \mu_1 \) and variance \( \sigma_1^2 \) and let \( Y_1, Y_2, \ldots, Y_n \) be a simple random sample from a normal distribution of mean \( \mu_2 \) and variance \( \sigma_2^2 \) and suppose that \( X_1, X_2, \ldots, X_n \) and \( Y_1, Y_2, \ldots, Y_n \) are independent samples then the sampling distribution \( T \) is students t-distribution with \( n-2 \) degrees of freedom (R. Lyman Ott et al, 2010).

Figure 12 shows simulation density of \( T \) for a random sample of \( n = 10 \) from two normal populations are taken. The sampling distribution of \( T \) is students t-distribution with 8 degrees of freedom. We use the following R code
\[
f=function(N,n){
 t=matrix(0,N);for(i in 1:N){x=rnorm(n,8,6); y=rnorm(n,7,4)
 r=cor(x,y);t[i]=r*sqrt(n-2)/sqrt(1-r^2)}
 return(t);}
f1=function(N,n1,n2)
 F=matrix(0,N);for(i in 1:N)
 F[i]=var(rnorm(n1,18,6))*49/36/var(rnorm(n2,12,7)))
 return(F)}
f1(10000,10,14)
 plot(density(t),xlim=c(min(t),max(t)),ylim=c(0,0.6),xlab="T ",ylab="density",main="")
 lines(sort(t),dt(sort(t),8),col="red",lwd=2,lty=2)
 legend(1.5,0.4,c("simulation","t-dist(8) df"),col=c("blue","red"),lwd=c(1,2),lty=c(1,2))
From Figure 12 we can see that the sampling distribution of T is students t-distribution with n-2 degrees of freedom.

**Inference on regression**

In this section we will simulate the simple linear regression model in order to verify the probabilistic behaviour of the resulting least square statistic. Let \( X_1, X_2, \ldots, X_n \) are independent(explanatory) variables and \( Y_1, Y_2, \ldots, Y_n \) be dependent(response) variables and let \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\) be simple random samples, then we want a relation \( Y = X \beta + \varepsilon \) here \( \beta \) is unknown unless the population is taken. So we want to estimate \( \beta \) from the sample. The estimated \( \hat{\beta} = (X'X)^{-1}X'Y \) our objective here is that we want to show the sampling distribution of \( \hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2) \) when \( \sigma^2 \) is known and the distribution is different when \( \sigma^2 \) is unknown(Sanford, W. 2005). The distribution of studentized \( \hat{\beta} \) is student t-distribution

For simple regression \( \beta = (\beta_0, \beta_1)' \) and \( \hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{\sum (x_i-\bar{x})^2} \) \( \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum (x_i-\bar{x})^2} \)

Let us go to the simulation. Assume we have \( N=40 \) observations the data and the R code are given in Appendix. The least square estimates and their variance is given in Table 1.

**Table 1:** The least square estimates of simple linear regression parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>estimated</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>2.965</td>
<td>2.30645</td>
</tr>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>1.054</td>
<td>0.01684804</td>
</tr>
</tbody>
</table>

But these quantities are unknown for us. What we have is only sample.

Let us take sample size say \( n=20 \) and let us see the distribution of \( \hat{\beta} \). the R code for this simulation is given in appendix. We take \( N=10000 \) simulations.

\[
E(\hat{\beta}_0) = \frac{\sum \hat{\beta}_0}{N} = 2.96799
\]

\[
Var(\hat{\beta}_0) = \frac{\sum (\hat{\beta}_0 - E(\hat{\beta}_0))^2}{N-1} = 2.3163576
\]

\[
E(\hat{\beta}_1) = \frac{\sum \hat{\beta}_1}{N} = 1.054187
\]

\[
Var(\hat{\beta}_1) = \frac{\sum (\hat{\beta}_1 - E(\hat{\beta}_1))^2}{N-1} = 0.0160689
\]

These values are approximately the same as the population parameters we found in the table. Then the distribution of the statistic \( \frac{\hat{\beta}-\beta}{\sqrt{Var(\hat{\beta})}} \) depends on whether the variance \( \sigma^2 \) is known or unknown.

![Figure 13](image1.png)

**Figure 13:** shows the density of 10000 simulation for a random sample of \( n=20 \) items when \( \sigma^2 \) is unknown.

**Figure 13** the studentized \( \hat{\beta} \) is \( \sim t(n-2) \)

When \( \sigma^2 \) is unknown the statistic

\[
T = \frac{\hat{\beta}-\beta}{\sqrt{Var(\hat{\beta})}}
\]

is student t-distribution with \( n-2 \) degrees of freedom.

If the population variance \( \sigma^2 \) is known the statistic

\[
Z = \frac{\hat{\beta}-\beta}{\sigma\sqrt{\frac{1}{n}}}
\]

is standard normal distribution.

**Figure 14:** shows the density of 10000 simulation for a random sample of \( n=20 \) items. If the population variance is known then the statistic \( Z \) is normal distribution.

Figure 14 the standardized \( \hat{\beta} \) is \( \sim N(0, 1) \)

Standardized estimate distribution is approximately standard normal distribution.

The variance of the sampling distribution of \( \hat{\beta} \) converges to the population parameters we found in the table.

**Figure 15:** the convergence of the variance of estimated coefficients
The structure of the variance of $\hat{\beta}_1$ is the same as the structure of the variance of $\hat{\beta}_0$.

This is due to the fact that $\text{Var}(\hat{\beta}_0) = \text{constant} + \text{constant} \times \text{Var}(\hat{\beta}_1)$.

DISCUSSION

Learning statistics are important for college students (Leslie C.). However, many students have difficulties understanding statistics concepts such as sampling distribution and confidence intervals (Leslie C). Students in introductory statistics class often struggle to understand the fundamental concepts of sampling distribution, central limit theorem, confidence interval and hypothesis testing (Leslie C, Moore D.S. 1997). Sampling distribution is the gateway to statistical inference (Leslie C). This paper is about one sample and two sample inference using computer simulation. It is demonstrated that the simulation of the sampling distribution of some basic statistics we often find in introductory statistics using R programing language (Leslie C, Moore D.S. 1997). R is increasingly being used as a tool for statistics education (Leslie C). We use several typical examples to illustrate how to conduct computer simulation using R. Simulation with the help of computer can be very effective tool in getting a good grasp of sampling distribution (Leslie C). The use of computer simulation in the teaching of introductory statistics can help undergraduate students understand difficult or abstract statistics concepts (Moore D.S. 1997). Simulation is essential to gain a good understanding about the concept of sampling distribution for students in statistics class (Leslie C). The tool of computer simulation is an essential part of understanding statistics (Moore D.S. 1997).

CONCLUSION

In this paper, we try to verify the sampling distribution of sample statistic that we found in statistical methods. We try to see central limit theorem and law of large number by mentioning counter example. Normality assumption is the basis for all distributions. For large sample size the distribution of the sample mean is normally distributed without considering the distribution of the original variable. The mean of normally distributed pattern is normal with decreasing variance as a function of sample size. In a situation where small sample size and unknown variance we use student t-distribution. However, as the degree of freedom increases t-distribution turns to normal distribution. The ratio of sample variance to population variance multiplied by degrees of freedom is chi-square distribution with $n-1$ degrees of freedom. F-distribution arises when we are interested in the ratio of two variances. In regression if the population variance is unknown the estimates follow t-distribution.

REFERENCES


Leslie Chandrakantha: Understanding Statistics Concepts using Simulation in R.

Leslie Chandrakantha: Sampling Distribution of the Mean in R

Leslie Chandrakantha, understanding sampling distribution using simulation in R


Monica E. Brussolo (2018): Understanding the Central Limit Theorem the Easy Way: A Simulation Experiment, http://creativecommons.org/licenses/by/4.0/


Accepted 29 February 2020


Copyright: © 2020 Reshid TM. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.
Appendix: The R code for Regression part

```r
# REGRESSION
x = c(13.1259289863519, 9.97280258173123, 3.60393659118563, 9.76076080370694,
     13.4466646108776, 18.6615084544756, 7.07075995905325, 10.7340311293656,
     11.0934200785123, 6.00321174040437, 5.06187068019062,
     8.30410914169624, 15.6528919194825, 19.340568847023, 17.6761783366092,
     15.3968393956311, 8.13349168887362, 2.47491842834279, 13.7670393986627,
     12.9030547216535, 9.61479113763198, 11.0504930313833, 13.6165081835352,
     15.227578259812, 8.8772931876952, 8.9937518001534, 6.91483139339834,
     4.3792885319672, 13.8287740319779, 8.4807922132313, 17.2897333060391)

y = c(18.3284716199676, 11.9983356793452, 3.19005389478917, 17.9857143032932,
     17.5261241893626, 24.2918720781243, 10.5747370562273, 6.15778938503531,
     9.6412808997099, 14.0942357409376, 14.802017684097, 21.1935663470971,
     23.6971419993407, 22.438711803928, 10.7544176247085, 12.7307036296116,
     17.9085304531413, 12.7612311923915, 14.6016582112767, 14.2730117991017,
     11.4572381322187, 11.6969638051808, 11.0486753962352, 21.0292074375214,
     24.6398452952169, 11.13948313939834, 6.80099139708624, 21.4242608411097,
     4.92690380123521, 24.9058315809615)

mm = lm(y ~ x); summary(mm)

f = function(n) {
  b0 = matrix(0, n); b1 = b0; v0 = b0; v1 = b0; vb0 = b0; vb1 = b0
  for (i in 1:n) {
    s = sample(1:40, 20); Y = y[s]; X = x[s]
    m = lm(Y ~ X)
    b0[i] = m$coeff[1]; b1[i] = m$coeff[2]; r = residuals(m)
    vb0[i] = 1/20 + (mean(X))^2/sum((X - mean(X))^2) * var(r) * 19/18
    vb1[i] = 1/sum((X - mean(X))^2) * var(r) * 19/18
    v0[i] = var(b0[1:i]); v1[i] = var(b1[1:i])
  }
  b = data.frame(b0, b1, vb0, vb1, v0, v1)
  return(b)
}

b = f(10000)
mean(b[, 1]); mean(b[, 2]); mean(b[, 3]); mean(b[, 4]); mean(b[, 1:10000, 5]); mean(b[, 2:10000, 6])
t0 = (b[, 1] - 2.965)/sqrt(b[, 5]); t1 = (b[, 2] - mean(b[, 2]))/0.1298

#### Sigma is unknown

plot(density(t0), xlim = c(min(t0), max(t0)), ylab = expression(density~of~studentized(hat(bet[0]))), xlab = expression(studentized(hat(beta[0]))));
lines(sort(t0), dt(sort(t0), 18), col = "black", lwd = 1)

plot(density(t1), xlim = c(min(t1), max(t1)), ylab = expression(density~of~studentized(hat(bet[1]))), xlab = expression(studentized(hat(beta[1]))));
lines(sort(t1), dt(sort(t1), 18), col = "black", lwd = 1)

#### Sigma is known

plot(density(z0), xlim = c(min(z0), max(z0)), ylab = expression(density~of~standardized(hat(bet[0]))), xlab = expression(standardized(hat(beta[0]))));
lines(sort(z0), dnorm(sort(z0), mean = 0, sd = 1), col = "black", lwd = 1)

plot(density(z1), xlim = c(min(z1), max(z1)), ylab = expression(density~of~standardized(hat(bet[1]))), xlab = expression(standardized(hat(beta[1]))));
lines(sort(z1), dnorm(sort(z1), mean = 0, sd = 1), col = "black", lwd = 1)
```

Sampling Distribution and Simulation in R