NAME: Geonho Marco ID NUMBER: 122200
SURNAME: You

ECONOMETRICS AND DATA ANALYSIS

Home Assessment N.1 - Solution

Deadline: 18 September 2020, 11h59pm [Paris time]

Problem 1 - (Law of Large Numbers)

Let X be a random variable representing the roll of a fair dice. Write the R code to show the Law of Large Numbers in action. Build a Monte Carlo experiment in which you set the sample size n equal to 2, 25, 100 and 1000 and

1. show that the range of variation of the sample average $\bar{X}$ shrinks.
2. represent graphically the four cases as in slide 18 in my Class 1 presentation.

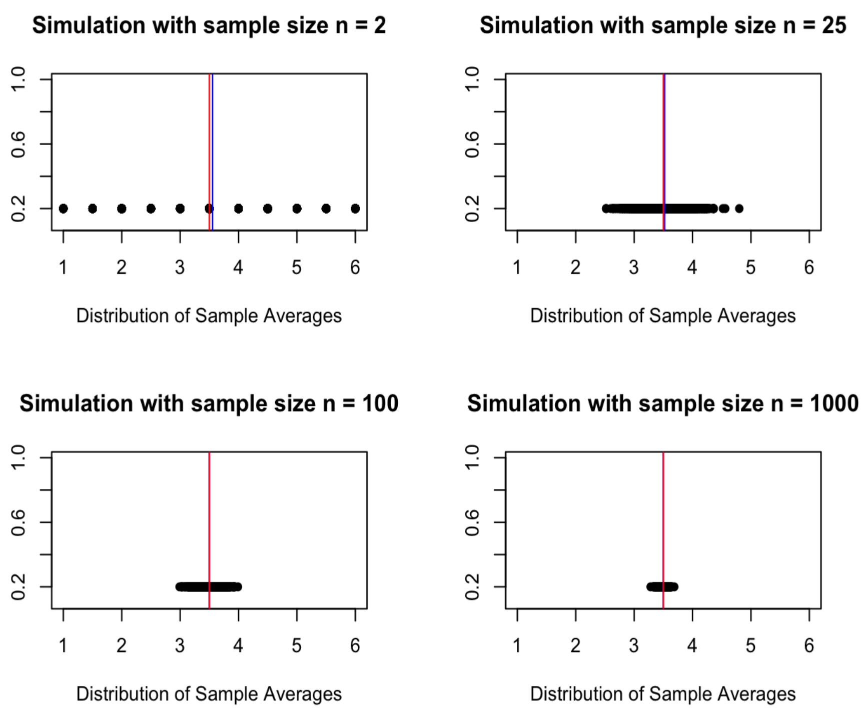
This problem asks to demonstrate the Law of Large Numbers in action using Monte Carlo experiment of rolling a dice 2, 25, 100, and 1000 times. The following section is the simulation built in R :

```r
# [0]
set.seed(1)
# [1]
LLN <- function(iteration, sample_size){
  # [2]
  mean_vector <- rep(0, iteration)
  title <- sprintf("Simulation with sample size n = %s", sample_size)
  # [3]
  for(n in 1:iteration){
    mean_vector[n] <- mean(sample(1:6, size = sample_size, replace = TRUE))
  }
  # [4]
  plot(mean_vector,
       rep(0.2, iteration),
       xlab = "Distribution of Sample Averages",
       ylab = "",
       xlim = c(1,6),
       ylim = c(0.1,1),
       pch = 16,
       main = title
       )
  # [5]
  abline(v = mean(mean_vector), col = "blue")
  abline(v = 3.5, col = "red")
}
par(mfrow=c(2,2))
# [6]
LLN(1000,2)
LLN(1000,25)
LLN(1000,100)
LLN(1000,1000)
```

Interpretation

[0]: We set a seed for replicability.
[1]: We define a function called LLN (Law of Large Numbers) receiving 2 parameters that are "iteration" and "sample-size".
[2]: We initialize a vector called "mean-vector" as a vector containing zeros of length equal to "iteration". Then we format the title of the graphs which will be drawn in the next steps.
[3]: We loop "iteration" times the followings:
i) draw a sample of experiment equivalent to rolling a dice "sample-size" times ;
ii) calculate the mean of the drawn sample ;
iii) store the calculated mean into the n-th row of "mean-vector" (n $\in$ [1:iteration]).
[4]: We plot the mean-vector using scatter plot.
[5]: We draw a red line for the population mean = 3.5 ($E[X] = \frac{1+2+3+4+5+6}{6} = 3.5$) and a blue line for the mean of different sample means.
[6]: We execute the LLN function for 4 different sample sizes (2,25,100,1000) with fixed iteration number equal to 1000. The result is as below.

Result



We can see that the more sample size grows, the lesser points are spread around the red line (population mean = 3.5). With a greater sample size, it is also much more likely for the blue line (sample mean) to collide the red line. We can conclude that the Law of Large Numbers is in action.

Problem 2 - (Central Limit Theorem)

Let X be a Normally distributed random variable with mean $\mu_X = -10$ and variance $\sigma_X^2 = 1$. Write the R code to show the Central Limit Theorem in action. Build a Monte Carlo experiment in which you set the sample size n equal to 2, 25, 100 and 1000 and

1. show that the distribution of $\frac{\bar{X}-\mu_X}{\sqrt{\sigma_X^2/n}}$ converges to a standardized Normal distribution, that is to a Normal distribution with mean and variance equal to 0 and 1 respectively.
2. represent graphically the four cases as in slide 23 in my Class 1 presentation.

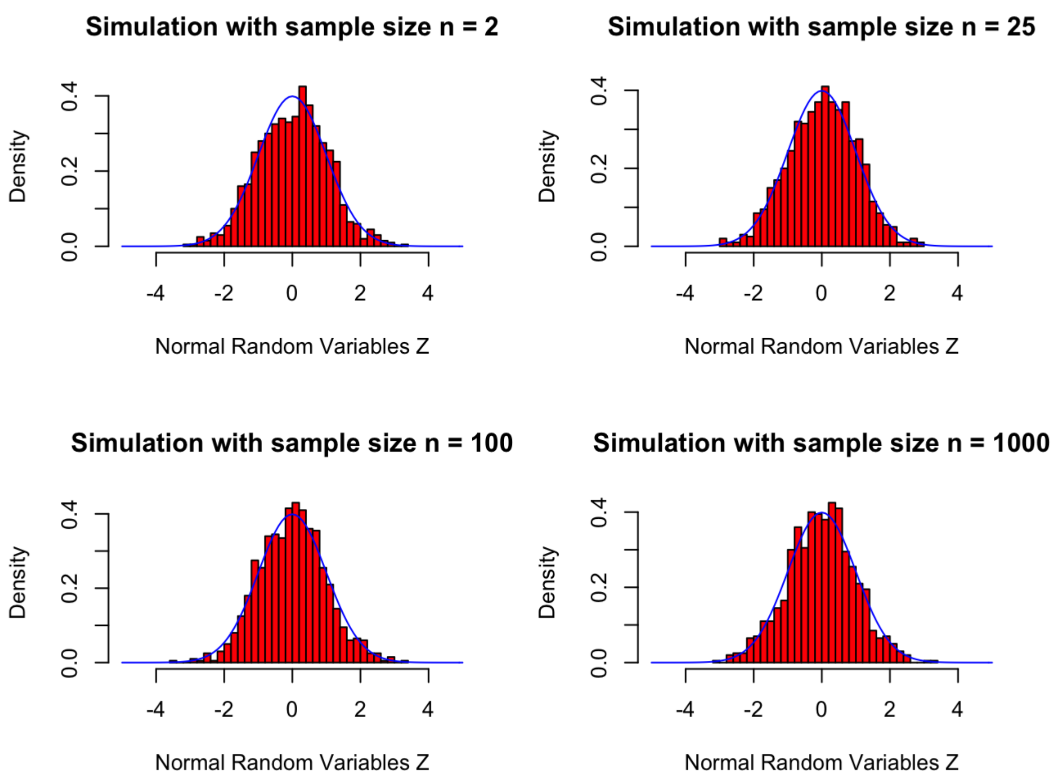
The problem asks to demonstrate, given samples from a normally distributed population of mean μ_X = -10 and variance σ_X^2 = 1, that $\frac{\bar{X}-\mu_X}{\sqrt{\sigma_X^2/n}}$ converges to a standardized normal distribution (where $\bar{X}$:= sample-mean and n := sample-size).

```r
# [0]
set.seed(2)
# [1]
CLT <- function(iteration, sample_size, mu, var){
  # [2]
  mean_X <- rep(0,iteration)
  Z <- rep(0,iteration)
  title <- sprintf("Simulation with sample size n = %s", sample_size)
  # [3]
  for(i in 1:iteration){
    X <- rnorm(sample_size, mean = mu, sd = sqrt(var))
    mean_X[i] <- mean(X)
    Z[i] <- (mean_X[i] - mu)/(sqrt(var/sample_size))
  }
  # [4]
  hist(Z,
       prob = TRUE,
       xlim = c(-5, 5),
       xlab = "Normal Random Variables Z",
       main = title,
       col = "red",
       breaks = 25,
       freq = FALSE)
  # [5]
  curve(dnorm(x, mean = 0, sd = 1),
        add=T,
        col="blue")
}
par(mfrow = c(2,2))
# [6]
CLT(1000,2,-10,1)
CLT(1000,25,-10,1)
CLT(1000,100,-10,1)
CLT(1000,1000,-10,1)
```

Interpretation

[0]: We set a seed for replicability.
[1]: We create a function named CLT (Central Limit Theorem) receiving parameters such as "iteration", "sample-size", "mu", and "var".
[2]: We initialize a vector called "mean-X" and "Z" for length equal to "iteration" containing only zeros, and a title for graphs later on.
[3]: We set a loop that executes "iteration" times of following tasks:
i) draw samples from a Normal population of mean equal to "mu" and standard deviation equal to square root of "var";
ii) store the mean of drawn sample in i-th row of mean-X vector;
iii) store the normalized mean of drawn sample in i-th row of Z vector.
[4]: We plot a histogram for random variable Z.
[5]: We draw the density curve for standard normal distribution.
[6]: We execute CLT function for 4 different sample sizes (2,25,100,1000) for a normal random variable of mu = -10 and var = 1 with fixed number of iteration equal to 1000. The result is as below.

Result



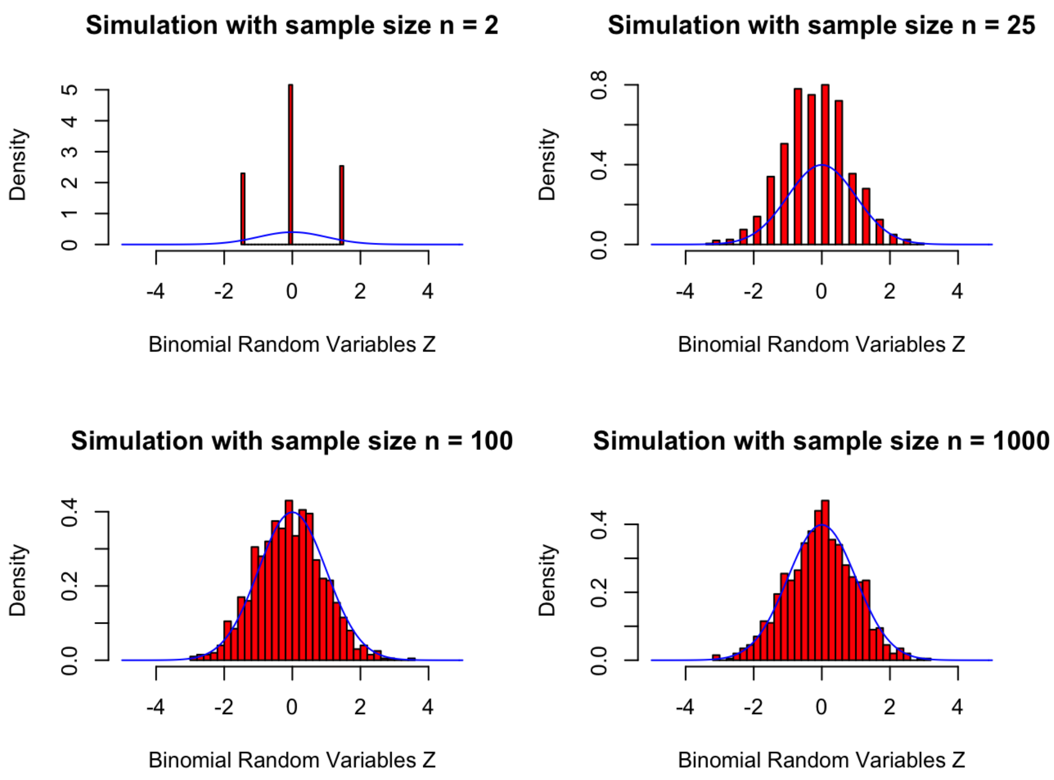
The result illustrates that histogram of each sample of different size fits well into the curve. This result will always be reproduced for any sample size or number of iteration because from the beginning we chose the Normal random variable X. The very purpose of the CLT is to show that with a large sample size, the random variable converges to a normal distribution in probability. But we already **assumed** the distribution beforehand.

A better alternative for this simulation could be, for instance, a simulation where we draw samples from a binomial random variable. To facilitate the demonstration, we set the probability of success (outcome $= 1$) equal to 0.5 hence for failure (outcome $= 0$) it will be 0.5 as well. Since there are only 2 possible outcomes 1 and 0 both having the same probability of realization, the mean of such experiment is, when repeated once, equal to $1 \times 0.5 = 0.5$ and the variance is equal to $1 \times 0.5 \times 0.5 = 0.25$.

```
CLT_binomial <- function(iteration, sample_size, mu, var){
  mean_X <- rep(0,iteration)
  Z <- rep(0,iteration)
  title <- sprintf("Simulation with sample size n = %s", sample_size)

  for(i in 1:iteration){
    # sampling binomial random variable X of outcome = 1 with probability = 0.5
    X <- rbinom(sample_size, 1, 0.5)
    mean_X[i] <- mean(X)
    Z[i] <- (mean_X[i] - mu)/(sqrt(var/sample_size))
  }
  hist(Z,
       prob = TRUE,
       xlim = c(-5, 5),
       xlab = "Binomial Random Variables Z",
       main = title,
       col = "red",
       breaks = 25,
       freq = FALSE)
  curve(dnorm(x, mean = 0, sd = 1),
        add = T,
        col = "blue")
}
par(mfrow = c(2,2))
CLT_binomial(1000, 2, 0.5, 0.25)
CLT_binomial(1000, 25, 0.5, 0.25)
CLT_binomial(1000, 100, 0.5, 0.25)
CLT_binomial(1000, 1000, 0.5, 0.25)
```

This alternative shows the following 4 graphs:



We can clearly see that the first(top-left) and the second(top-right) graphs are not fit to the density curve of standard normal distribution. However, with the sample size n = 100 and n = 1000, we have distributions that are fit to the N(0,1) density curve. In this case, we can confidently say that the increase in sample size allows a random variable to converge to a Normal distribution in probability. If such random variable is normalized, then it converges to a standard Normal distribution.