

# Introductory Statistics for Economics

## ECON1013: LAB 2

Duong Trinh

University of Glasgow

20/02/2022

# Picture the Scenario

## ► Objective:

- illustrate how the statistic  $\bar{X}_n$  varies with repeated samples
- describe what might happen if we repeat the entire sampling process and compute the sample average again and again.
- illustrate the CLT and the LLN

## ► Implementation:

- We use a *pseudo-random number generator*, already built in R, to simulate the samples. Simulations allow the user to perform experiments and answer questions in a rapid manner.

# PRELIMINARIES

# The normal distribution

For each distribution, R provides functions that allow users calculate value of: quantile, probability and generate random variable of the distribution.

- ▶ To calculate the *probability* of  $P(X \leq 0.05)$  where  $X \sim N(0, 1)$ :

```
pnorm(q = 0.05, mean = 0, sd = 1)
```

- ▶ To calculate the *probability* of  $P(X \leq 1)$  where  $X \sim N(4, 3^2)$ :

```
pnorm(q = 1, mean = 4, sd = 3)
```

# The normal distribution

For each distribution, R provides functions that allow users calculate value of: quantile, probability and generate random variable of the distribution.

- ▶ To calculate the *quantile* at  $p = 0.05$  of  $N(0, 1)$ :

```
qnorm(p = 0.05, mean = 0, sd = 1)
```

- ▶ To calculate the *quantile* at  $p = 0.05$  of  $N(4, 3^2)$ :

```
qnorm(p = 0.05, mean = 4, sd = 3)
```

# The normal distribution

For each distribution, R provides functions that allow users calculate value of: quantile, probability and generate random variable of the distribution.

- ▶ To calculate value of *density function* of  $N(0, 1)$  at 0.4:

```
dnorm(x = 0.4, mean = 0, sd = 1)
```

- ▶ To calculate value of *density function* of  $N(4, 3^2)$  at 0.4:

```
dnorm(x = 0.4, mean = 4, sd = 3)
```

# The normal distribution

For each distribution, R provides functions that allow users calculate value of: quantile, probability and generate random variable of the distribution.

- ▶ To generate randomly 1000 numbers from  $N(0, 1)$ :

```
R <- rnorm(n = 1000, mean = 0, sd = 1)
head(R)
length(R)
summary(R)
```

## Loops (for) and conditional expressions (if)

- ▶ Create the vector GENDER where the entries are either male or female

```
GENDER = c('male', 'male', 'female', 'male', 'female')
```

- ▶ Define the random variable  $X$  taking value 1 if the individual is a male, and 0 if a female.

```
X=numeric(0)
for (i in 1:length(GENDER)) {
  if (GENDER[i] == 'male') {
    X[i] = 1
  }
  else {
    X[i]=0
  }
}
```

## Exercise 1: The Central Limit Theory

## Exercise 1: The Central Limit Theory

- ▶ Simulate 10,000 samples of size  $n$  from an exponential distribution with mean  $\mu = 10$ . Consider different sample sizes, e.g.  $n = 10, 30, 500$ .
- ▶ For each sample, compute the standardized sample average:

$$\tilde{X} = \frac{\bar{X} - \mathbb{E}(\bar{X})}{\sigma_{\bar{X}}}$$

- ▶ Compute the empirical quantiles of the sample distribution of  $\tilde{X}$ . Compare the results with the theoretical quantiles of the standard normal random variable  $Z$ .
- ▶ Compute quantile  $z$  satisfying  $P(Z < z) = 0.975$ . Calculate the percentage of values of  $\tilde{X}$  greater than  $z$ .

Comment on your results.

## Exercise 1: The Central Limit Theory

Specify the parameters of the experiment

```
# "n" is the sample size. Try n = 10, 30, 500
n = 10
# "R" is the number of replications
R = 10000
# Parameter exponential
mu = 10 # mean
```

To reproduce results, setting a seed in R means to initialize a pseudorandom number generator

```
set.seed(12) # take an (arbitrary) integer
```

## Exercise 1: The Central Limit Theory

Simulate  $R$  samples of size  $n$  from an exponential distribution with mean  $\mu = 10$  and compute the standardized sample average for each sample:

$$Y = \frac{\bar{X} - \mathbb{E}(\bar{X})}{\sigma_{\bar{x}}}$$

```
# Create vector Y (pre-allocation)
Y = numeric(0)
# Loop through R replications
for (i in 1:R) {
  X = rexp(n,1/mu) # The sample will change for every "i"
  # Compute standardized sample average
  Y[i] = (mean(X)-mu)/(mu/sqrt(n))
}
```

## Exercise 1: The Central Limit Theory

```
# Empirical quantiles
```

```
quantile(Y)
```

```
##          0%        25%        50%        75%       100%
## -2.6105220 -0.7338397 -0.1024090  0.5983728  5.6335527
```

```
# Population quantiles of a standard normal distribution
```

```
qnorm(p = 0.25, mean = 0, sd = 1)
```

```
## [1] -0.6744898
```

```
qnorm(p = 0.5, mean = 0, sd = 1)
```

```
## [1] 0
```

```
qnorm(p = 0.75, mean = 0, sd = 1)
```

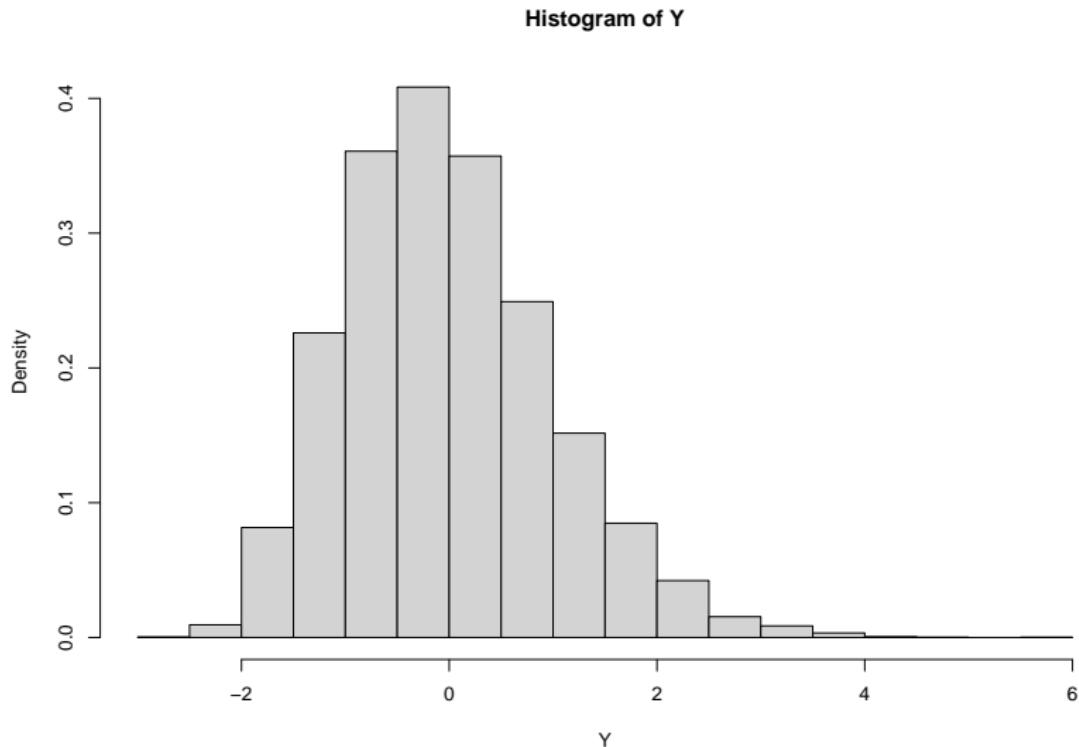
```
## [1] 0.6744898
```

## Exercise 1: The Central Limit Theory

To assess the how well the sampling distribution of  $X$  is approximated by the standard normal we compare the quantile of  $X$  and the quantile for a standard normal distribution.

# Exercise 1: The Central Limit Theory

```
hist(Y, prob = T)
```



## Exercise 1: The Central Limit Theory

```
# Compute the quantile z satisfying P(Z<z) = 0.975
z = qnorm(p = 0.975, mean = 0, sd = 1)
z

## [1] 1.959964

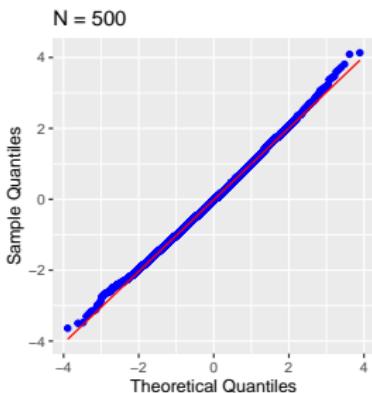
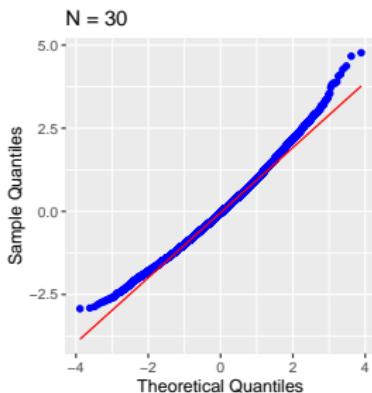
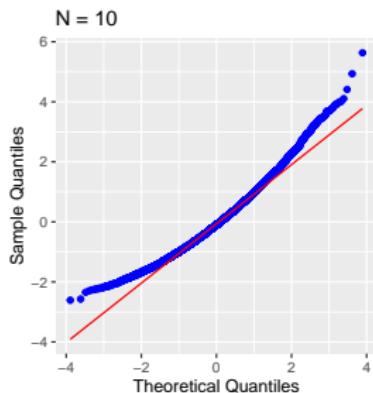
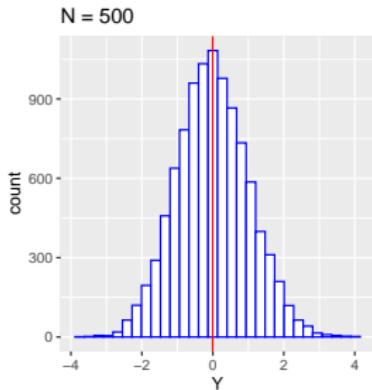
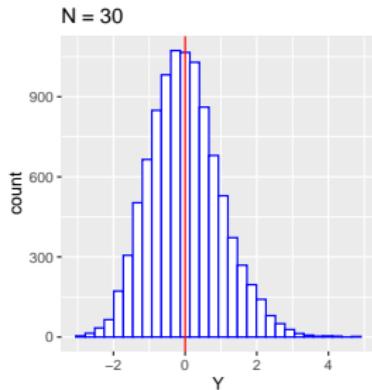
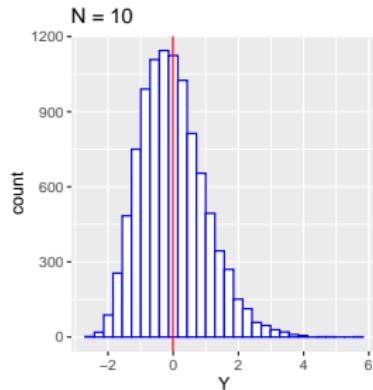
# Create vector (pre-allocation) to count how many times Y>z
count = numeric(0)
for (i in 1:R) {
  if (Y[i] > z) {
    count[i] = 1
  }
  else {
    count[i] = 0
  }
}
sum(count)/R*100 # The percentage

## [1] 3.79
```

## Exercise 1: The Central Limit Theory

If the sample is well approximated by a standard normal distribution, the realization of  $Y$  should be greater than  $z : (1 - 0.975 * R) = 250$  times, or 2.5% of the times.

# Exercise 1: The Central Limit Theory



## The Central Limit Theory - Reflection

The distribution of the sample average  $\bar{X}_n$  will be approximately normal as long as the sample size  $n$  is large enough.

## Exercise 2: The Law of Large Numbers

## Exercise 2: The Law of Large Numbers

- ▶ Compute the sample averages

$$\bar{x}_n = \frac{\sum_{i=1}^n x_i}{n}$$

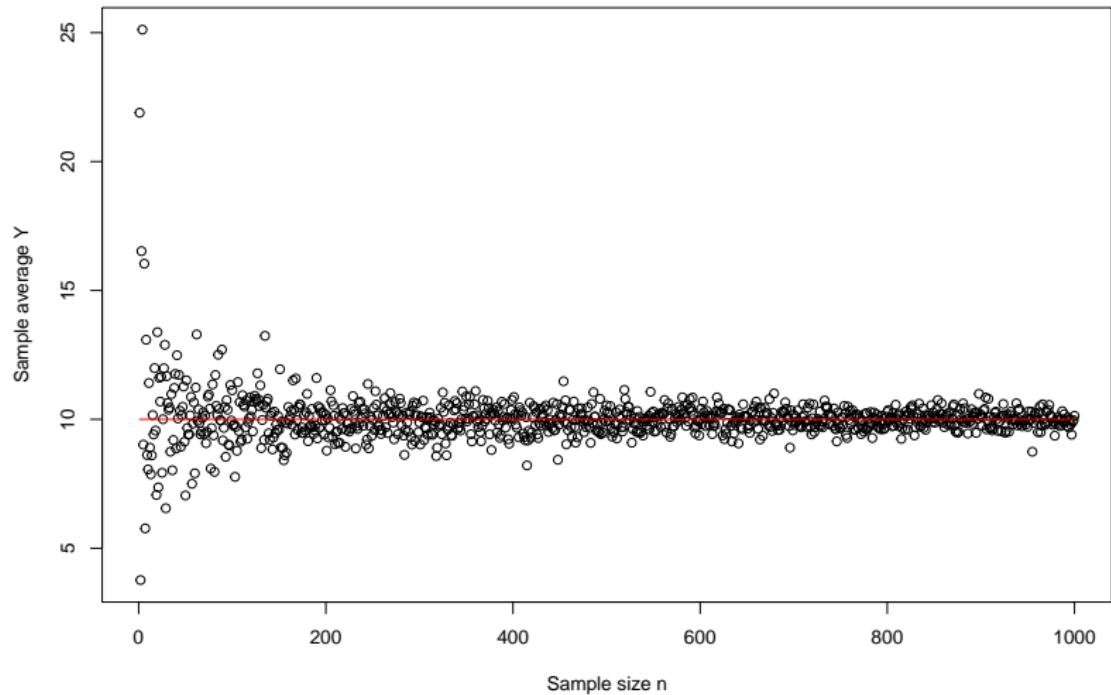
for  $n$  draws from the exponential distribution considered in Exercise 1, for  $n = 1, 2, \dots, 1000$ .

- ▶ Plot the sample averages. What do you conclude about the variability of the sample average?

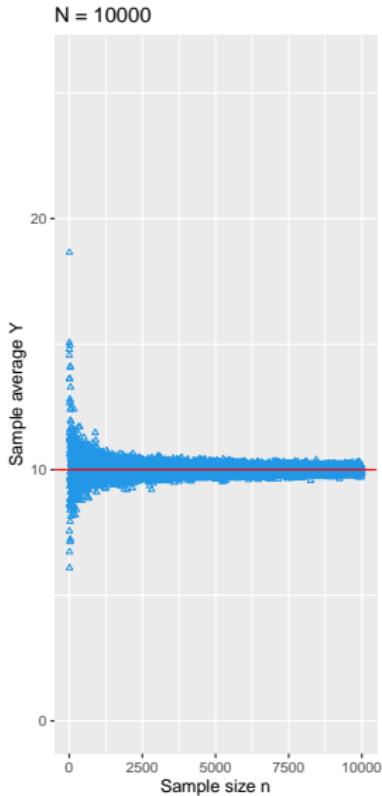
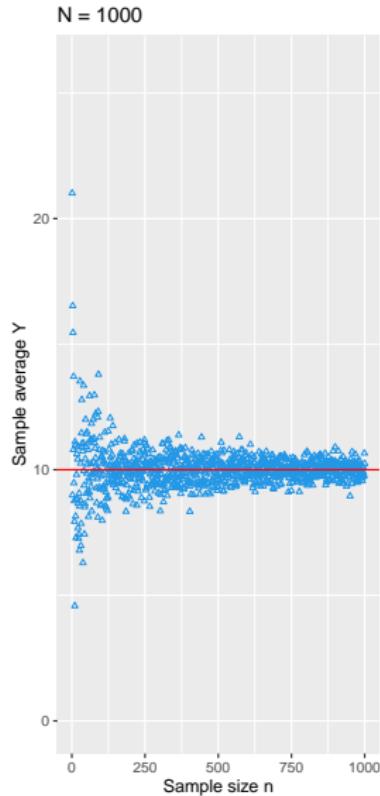
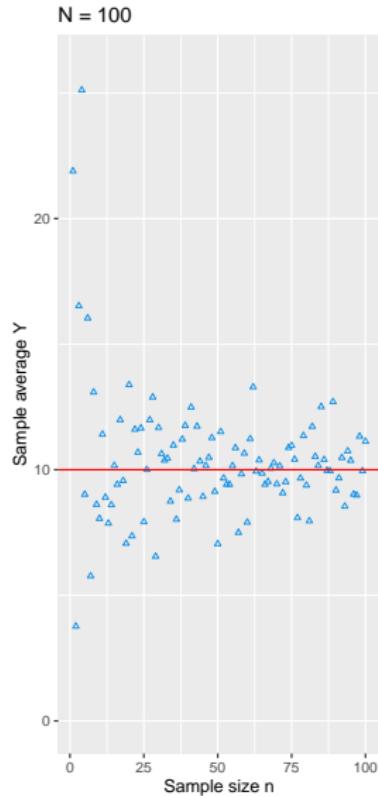
## Exercise 2: The Law of Large Numbers

```
set.seed(12) # to reproduce results
N = 1000 # maximum sample size
mu = 10 # parameter exponential
Y = numeric(0) # pre-allocation
for (i in 1:N) {
  X = rexp(i,1/mu) # 'i' draws from Exp(1/mu)
  Y[i] = mean(X) # sample average
}
plot(Y, xlab = "Sample size n", ylab = "Sample average Y")
lines(rep(mu,N), col = 'red')
```

## Exercise 2: The Law of Large Numbers



## Exercise 2: The Law of Large Numbers



## The Law of Large Numbers - Reflection

Given a random sample of size  $n$  from a population, the sample average  $\bar{X}_n$  will approach the population mean  $\mu$  as the sample size  $n$  becomes large.