ECOM20001 Econometrics 1

Tutorial 3 Semester 1, 2022

Chin Quek

Department of Economics

Introduction

- Distributions
- Law of Large Numbers
- Central Limit Theorem
- Conditional Probability Practice Problems

```
## Set the working directory for the tutorial file
setwd("your working directory")
```

Part 1: Normal, Chi-Square, t, F, Distributions

Pre-tutorial Question 1

i. Explain why the following equality holds from the code

- `pnorm(-1.65, mean=0, sd=1) = 1 - pnorm(1.65, mean=0, sd=1)`

ii. Explain the relationship between the output from the following two lines of code:

```
pnorm(1.96,mean=0,sd=1)
qnorm(0.975,mean=0,sd=1)
```

Summary of Distributions in R

Probability distribution functions usually have four functions associated with them. The functions are prefixed with a:

- `d`: for density
- `p`: for cumulative distribution
- `q`: for quantile function
- `r`: for random number generation

For a normal distribution:

- `dnorm(x, mean, sd)`: `dnorm(0,0,0.5)` gives the density (height of the PDF) of the normal distribution with mean $= 0$, sd $= 0.5$
- `pnorm(q, mean, sd)`: `pnorm(1.96,0,1)` gives the area under the standard normal curve to the left of 1.96
- `qnorm(p, mean, sd)`: `qnorm(0.975,0,1)` gives the value at which the CDF of the standard normal is 0.975.
- `rnorm(n, mean, sd)`: `rnorm(1000,3,0.25)` gives 1000 numbers from a normal distribution with mean 3 and sd $= 0.25$

Note: For all functions of the normal distribution, leaving out the mean and standard deviation would result in default values of `mean = 0` and `sd = 1`, i.e. a standard normal distribution.

Pre-tutorial Question 2: Conditional Distribution Practice Problems

Consider the following table which describes the joint probability distribution for all combinations of studying and performance. The outcome space for Studying (Y) and Performance (X) is:

- Y - Studying: Study Hard, Study Sometimes, Study Never
- X - Performance: High Grade , Medium Grade , Low Grade

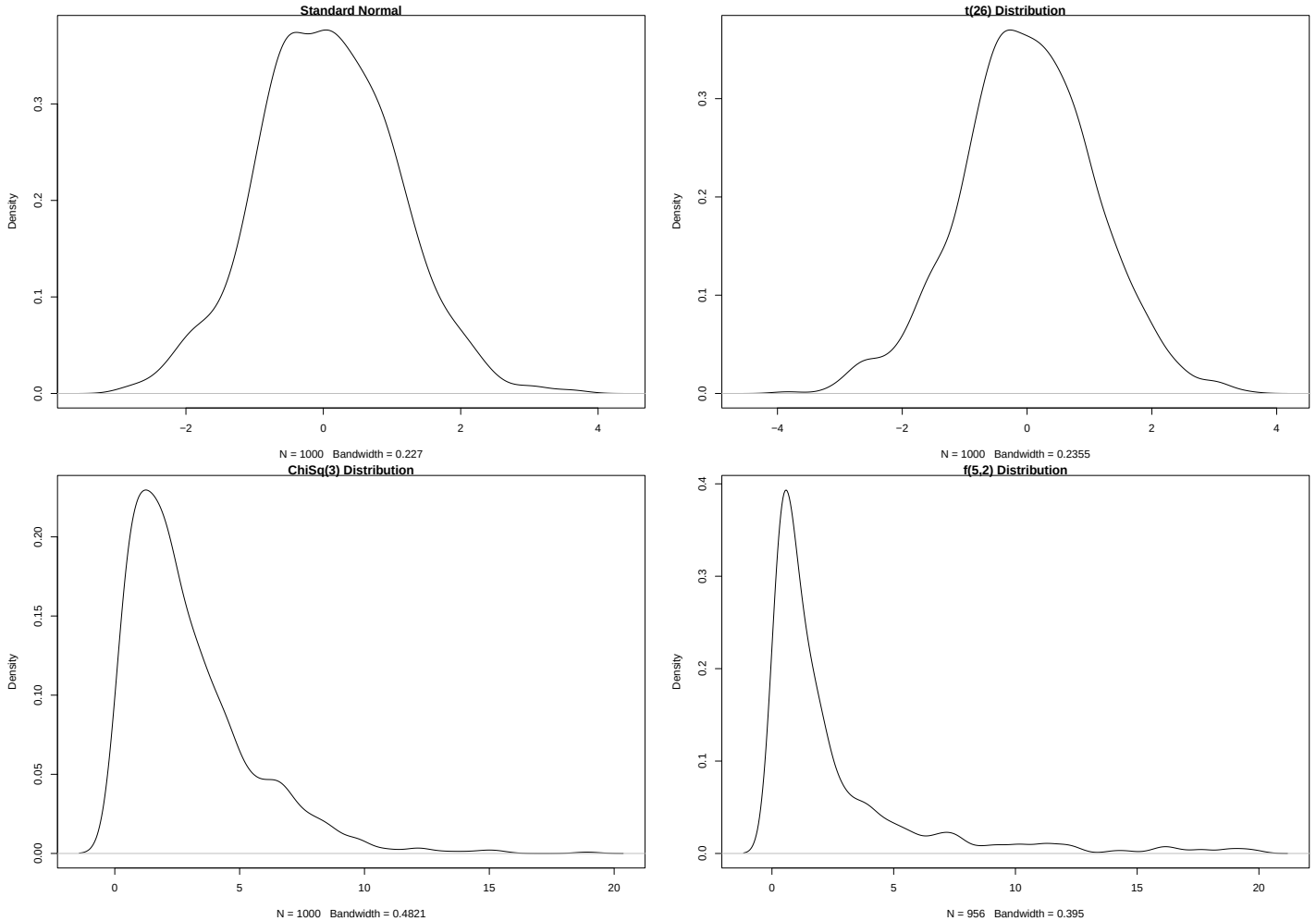
	High Grade	Medium Grade	Low Grade	Total
Study Hard	0.20	0.10	0.02	0.32
Sometimes	0.07	0.30	0.10	0.47
Never Study	0.01	0.05	0.15	0.21
Total	0.28	0.45	0.27	1.00

a. What is the marginal distribution for studying?
b. What is the marginal distribution for performance?
c. What is the probability distribution of Performance, conditional on Studying hard?
d. What is the probability distribution of Performance, conditional on Studying Sometimes?
e. What is the probability distribution of Studying, conditional on Medium Grade?
f. What is the probability distribution of Studying, conditional on Low Grade?
g. Using an example from the table above, show that Studying and Performance are not independently distributed.

In-tutorial Question 1:

Comment on the shape of the **Normal** (`mean=0, sd=1`), **Chi-Square** (`df=3`), **t** (`df=26`), and **F** (`df1=5, df2=2`) distributions.

Discuss whether each distribution is symmetric, right or left skewed, and whether you would expect the mean of the distribution to equal the median, be smaller than the median, or larger than the median.



In-tutorial Question 2:

Using the code chunk below, compute the sampling distribution of the mean from an underlying sample that is **Chi-Square** with `df=3` for sample sizes of `nobs=1000`.

Note

You do not need to know how to use loops and matrix operations in ECOM20001. This is denoted by ***** OPTIONAL (start) ****** and ***** OPTIONAL (end) ****** in the code chunk.

In-tutorial Question 2:

Using the code chunk below, compute the sampling distribution of the mean from an underlying sample that is **Chi-Square** with `df=3` for sample sizes of `nobs=1000`.

Note

You do not need to know how to use loops and matrix operations in ECOM20001. This is denoted by ***** OPTIONAL (start) ****** and ***** OPTIONAL (end) ****** in the code chunk.

Steps in R:

- Generate distribution of sample means

1. Generate `nobs=1000` sample of x's from the distribution using `rchisq` (for random number generation)
2. Repeat 1. K times to get K samples of x's from the same distribution
3. Obtain the mean for each sample of x's and store in a matrix "means"

Optional

R codes for Steps 2 and 3

a. What is the variance of the sampling distribution of the means?

b. Suppose a sample average is "close" to the true value if it is within 0.3 of the true value. What percentage of sample means lies within 0.3 of the true population mean of 3?

a. What is the variance of the sampling distribution of the means?

b. Suppose a sample average is "close" to the true value if it is within 0.3 of the true value. What percentage of sample means lies within 0.3 of the true population mean of 3?

- To compute the variance of the sampling distribution of the mean, use `var()` command
- To compute the percentage of sample means lying within 0.3 of the true value of the mean, use the formula

$$\text{pct} = 100 * \text{sum}(\text{err} < 0.3)/K \qquad \text{where err} = \text{abs}(\text{means - truevalue})$$

Generate nobs=1000 sample of x's from a χ^2_3 dist; compute the mean

```
nobs=1000
x=rchisq(nobs,df=3)
x_mean=mean(x)
print(x_mean)    # First mean
```

```
## [1] 3.036069
```

```
## Do it again:
x=rchisq(nobs,df=3)
x_mean=mean(x)
print(x_mean)    # Second mean
```

```
## [1] 3.098726
```

```
## Do one more time:
x=rchisq(nobs,df=3)
x_mean=mean(x)
print(x_mean)    # Third mean
```

```
## [1] 2.90796
```

Generate distribution of sample means (Optional)

```
nobs=1000    # sample size number of observations
df=3         # degree of freedom
K=500        # number of repetitions

## Creates a variable K=500 rows for saving the 500 means
means=matrix(0,K,1)

## Draw K=500 random samples each with nobs=1000 observations from a
# Chi-Square distribution with df=3, and save the mean from each sample

for (k in 1:K){
  x=rchisq(nobs,df=3) # draw the sample
  means[k]=mean(x)    # save the mean
}
```

- You now have K=500 means stored in the variable "means"

```
# Recall that the Chi-Square distribution has an E[x]=df, that is,
# the expected value is simply the degrees of freedom of the distribution
# So the true value of the mean is truevalue=df, which is 3 in our example

## Compute true value
truevalue=df

# Compute the percentage of the K=500 sample averages in the means
# variable that are with 0.3 of the true value of 3, that is the number
# of sample averages that lie between 2.7 and 3.3

## Compute abs value of the error between sample averages and true value
err=abs(means-truevalue)

# Here, we've chosen an error rule within 0.3 of the true value

## Compute percentage of sample averages within 0.3 of the true value;
pct=100*sum(err<0.3)/K

## Compute variance of the K=500 sample averages with var()
varmeans=var(means)
```

Results for nobs = 1000

```
## Print LLN results
print("Number of observations in each random sample")
print(nobs)
print("Percentage of Sample Averages within 0.3 of the True Value")
print(pct)
print("Variance of the K=500 Sample Means")
print(varmeans)
```

```
## [1] 1000

## [1] 100

##              [,1]
## [1,] 0.006258738
```

a. What is the variance of the sampling distribution of the means?

- 0.006

b. Suppose a sample average is "close" to the true value if it is within 0.3 of the true value. What percentage of sample means lies within 0.3 of the true population mean of 3?

- 100%

a. What is the variance of the sampling distribution of the means?

- 0.006

b. Suppose a sample average is "close" to the true value if it is within 0.3 of the true value. What percentage of sample means lies within 0.3 of the true population mean of 3?

- 100%

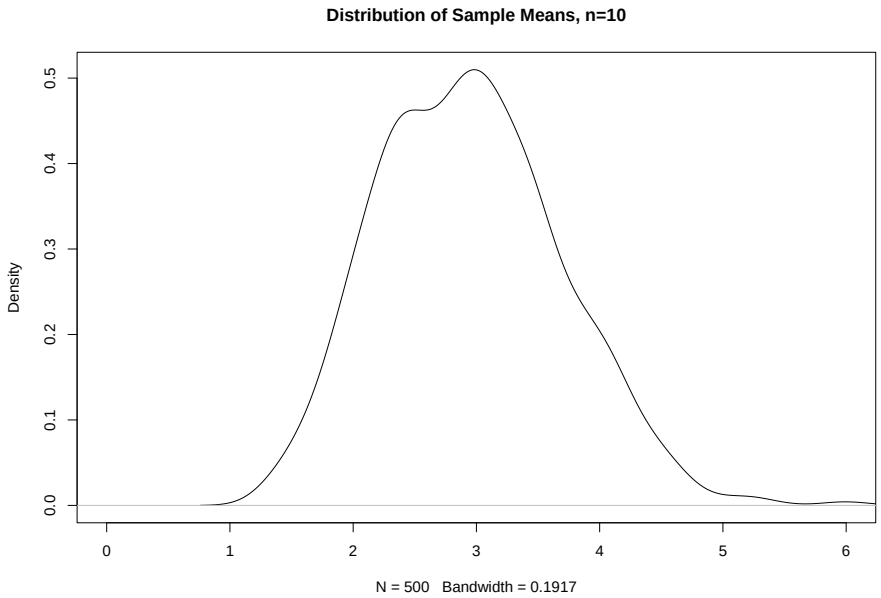
Note

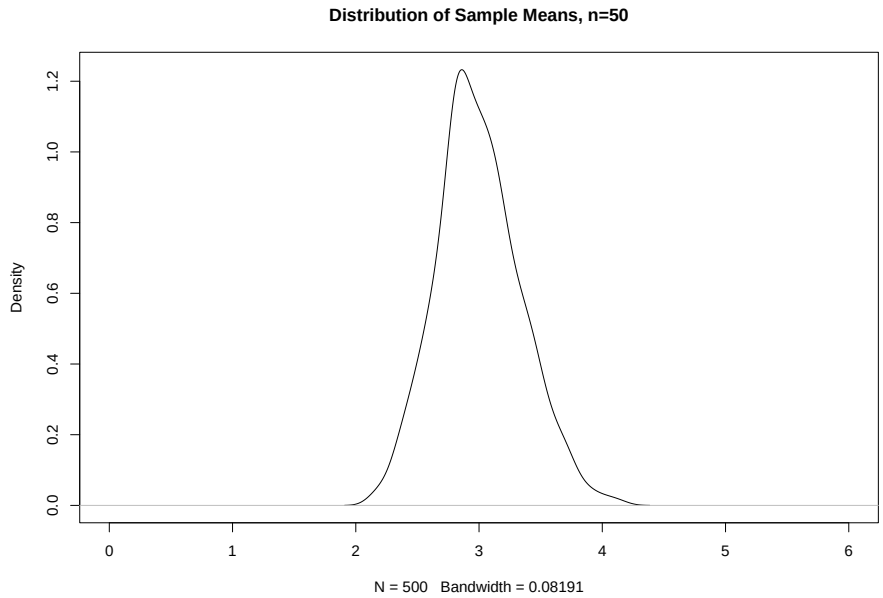
You may obtain a slightly different answer to parts a and b. This is due to the random generation of samples of x's.

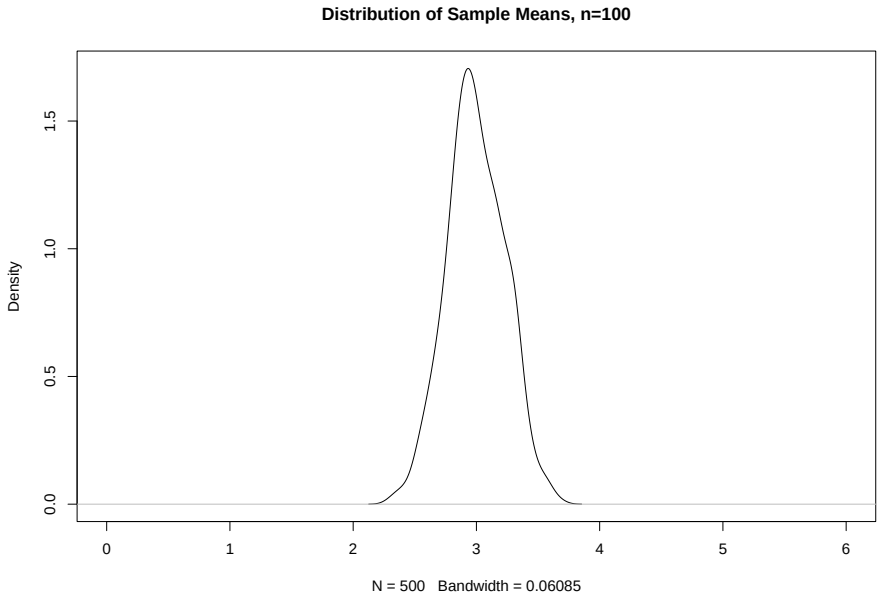
In-tutorial Question 3:

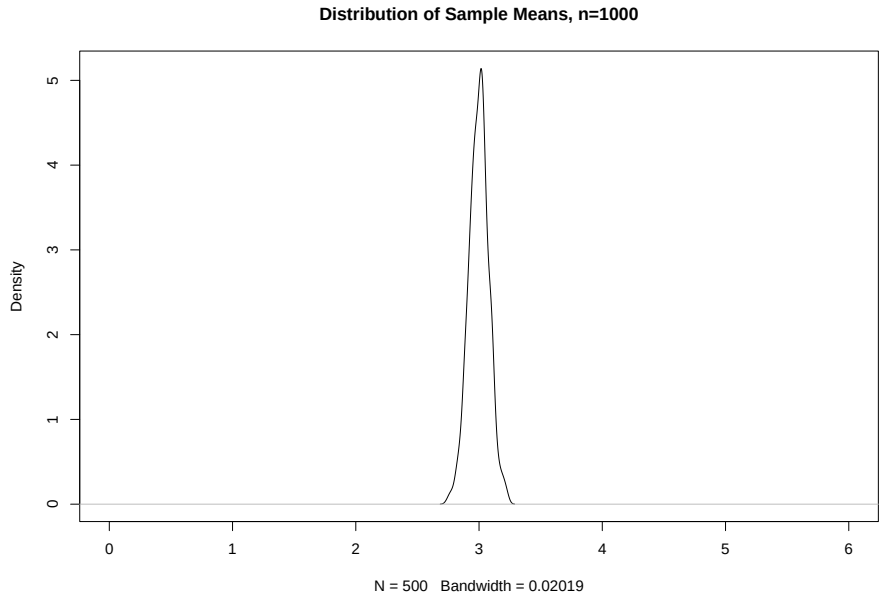
Consider the table of results and graphs of the distribution of sample means produced below. Explain how the results highlight how the Law of Large Numbers (LLN) and Central Limit Theorem (CLT) work as the number of observations increases from $n = 10$ to $n = 1000$.

n	Variance of the means	% of sample means lying within 30% of the true value of the mean
10	0.60	30%
50	0.12	65%
100	0.06	80%
1000	0.006	100%









- From the results, what can we conclude?

Law of large numbers and Central limit theorem

- The **LLN** says that the sample average will be more likely to be close to the true value of the mean as the number of observations (n) becomes large.
- The **LLN** is illustrated here: as n increases, the fraction of sample means that you compute from random sampling is more likely to be close (e.g., within 0.3 as our "close" rule) to the true value of the mean.
- The fact that the variance of the sample mean falls as n grows is further revealing of the **LLN** in action as n grows.
  - We get more precise estimates of the underlying population true value of the mean from the sample average.
  - In other words, a given sample average is more likely to be close to the true mean value as n rises
- The **CLT** is illustrated by the 4 graphs. These show that the distribution of the sample means becomes more symmetric and closer to a normal distribution as the number of observations grows.

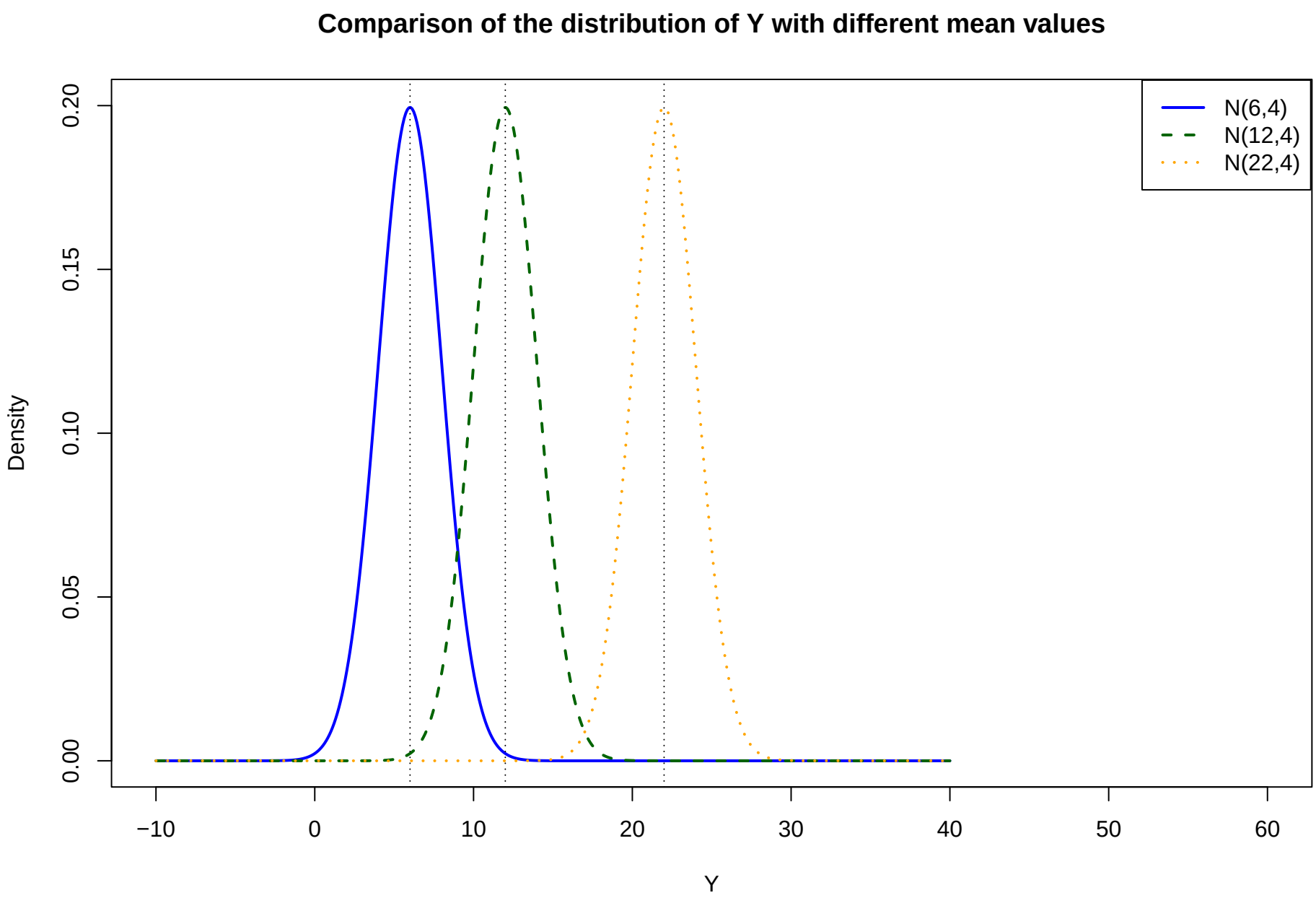
In-tutorial Question 4:

Suppose you have a random variable X that is i.i.d. (independent and identically distributed) from a $N(\mu_X, 1)$ distribution, and another random variable Y that is defined as follows: $Y = 2 + 2X$.

a. What is the distribution of Y?

b. Graphically plot the distribution of Y for different values of μ_X ($\mu_X = 2, 5, 10$).

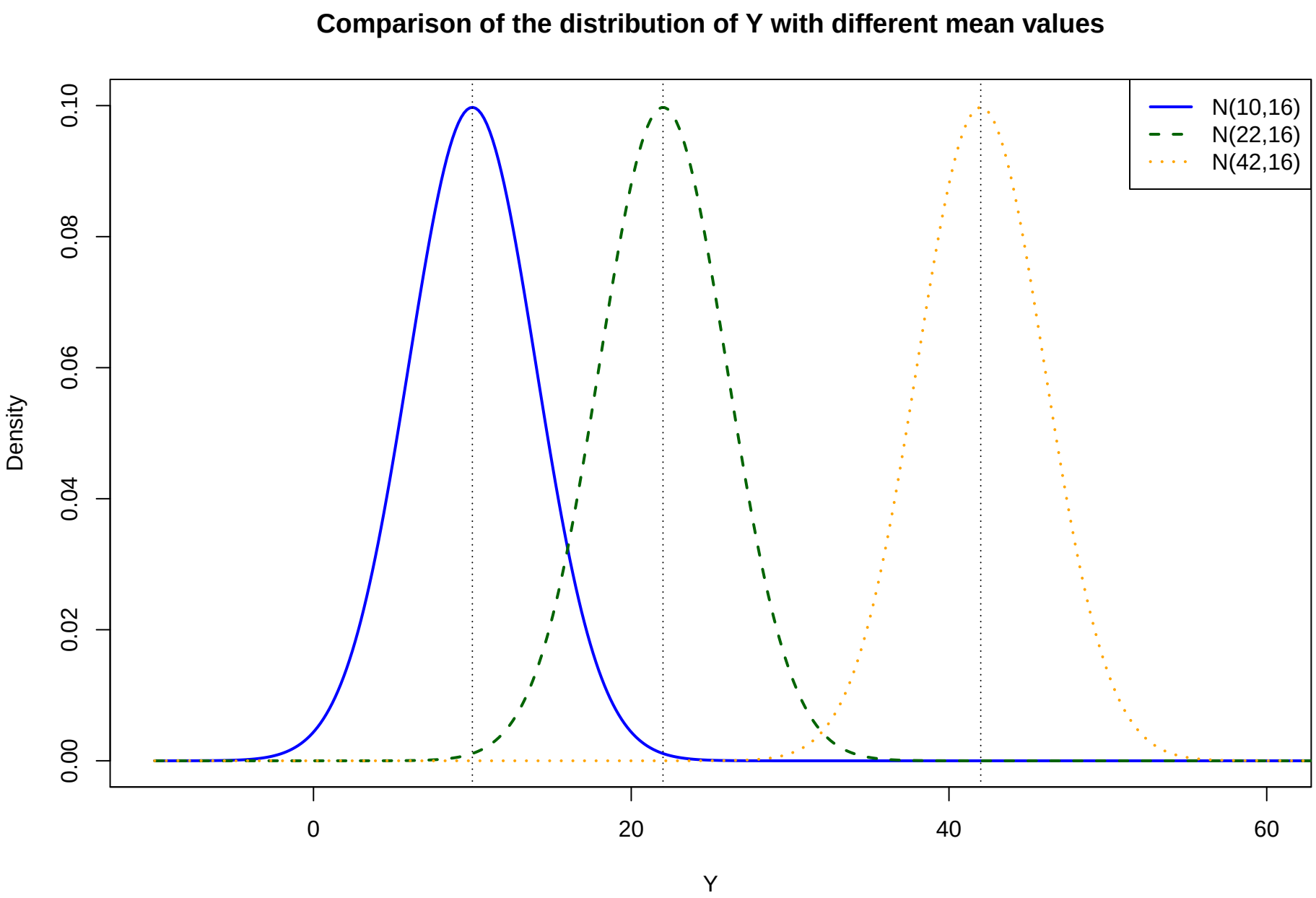
What is happening to the distribution of Y for these different μ_X values?


Comparison of the distribution of Y with different mean values
N(6,4)
N(12,4)
N(22,4)
Density
0.00
0.05
0.10
0.15
0.20
−10
0
10
20
30
40
50
60
Y

c. Suppose Y was instead distributed as $Y = 2 + 4X$.

- What is the distribution of Y now?

- Again, graphically plot the distribution of Y for different values of μ_X ($\mu_X = 2, 5, 10$) and compare your results to what you found in part b.
- What can you conclude about the magnitude of the shifts in the distribution of Y as a function of different μ_X values as the magnitude of the slope in the linear function that defines Y increases?


Comparison of the distribution of Y with different mean values
Density
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
Y
N(10,16)
N(22,16)
N(42,16)