## The Central Limit Theorem

Suppose that a sample of size $n$ is selected from a population that has mean $\mu$ and standard deviation $\sigma$. Let $X_1, X_2, \cdots, X_n$ be the $n$ observations that are independent and identically distributed (i.i.d.). Define now the sample mean and the total of these $n$ observations as follows:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$T = \sum_{i=1}^{n} X_i$$

The *central limit theorem* states that the sample mean $\bar{X}$ follows approximately the normal distribution with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$, where $\mu$ and $\sigma$ are the mean and standard deviation of the population from where the sample was selected. The sample size $n$ has to be large (usually $n \geq 30$) if the population from where the sample is taken is nonnormal. If the population follows the normal distribution then the sample size $n$ can be either small or large.
To summarize: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.
To transform $\bar{X}$ into $z$ we use: $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Example: Let $X$ be a random variable with $\mu = 10$ and $\sigma = 4$. A sample of size 100 is taken from this population. Find the probability that the sample mean of these 100 observations is less than 9. We write $P(\bar{X} < 9) = P(z < \frac{9-10}{\frac{4}{\sqrt{100}}}) = P(z < -2.5) = 0.0062$ (from the standard normal probabilities table).

Similarly the central limit theorem states that sum $T$ follows approximately the normal distribution, $T \sim N(n\mu, \sqrt{n}\sigma)$, where $\mu$ and $\sigma$ are the mean and standard deviation of the population from where the sample was selected.
To transform $T$ into $z$ we use: $z = \frac{T-n\mu}{\sqrt{n}\sigma}$

Example: Let $X$ be a random variable with $\mu = 10$ and $\sigma = 4$. A sample of size 100 is taken from this population. Find the probability that the sum of these 100 observations is less than 900. We write $P(T < 900) = P(z < \frac{900-100(10)}{\sqrt{100}(4)}) = P(z < -2.5) = 0.0062$ (from the standard normal probabilities table).

Below you can find some applications of the central limit theorem.

**EXAMPLE 1**
A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean $\mu = 205$ pounds and standard deviation $\sigma = 15$ pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

**EXAMPLE 2**
From past experience, it is known that the number of tickets purchased by a student standing in line at the ticket window for the football match of $UCLA$ against $USC$ follows a distribution that has mean $\mu = 2.4$ and standard deviation $\sigma = 2.0$. Suppose that few hours before the start of one of these matches there are 100 eager students standing in line to purchase tickets. If only 250 tickets remain, what is the probability that all 100 students will be able to purchase the tickets they desire?

**EXAMPLE 3**
Suppose that you have a sample of 100 values from a population with mean $\mu = 500$ and with standard deviation $\sigma = 80$.

a. What is the probability that the sample mean will be in the interval $(490, 510)$?

b. Give an interval that covers the middle 95% of the distribution of the sample mean.

**EXAMPLE 4**
The amount of regular unleaded gasoline purchased every week at a gas station near $UCLA$ follows the normal distribution with mean 50000 gallons and standard deviation 10000 gallons. The starting supply of gasoline is 74000 gallons, and there is a scheduled weekly delivery of 47000 gallons.

a. Find the probability that, after 11 weeks, the supply of gasoline will be below 20000 gallons.

b. How much should the weekly delivery be so that after 11 weeks the probability that the supply is below 20000 gallons is only 0.5%?

Solutions:
**EXAMPLE 1**
We are given $n = 49, \mu = 205, \sigma = 15$. The elevator can transport up to 9800 pounds. Therefore these 49 boxes will be safely transported if they weigh in total less than 9800 pounds. The probability that the total weight of these 49 boxes is less than 9800 pounds is $P(T < 9800) = P(z < \frac{9800-49(205)}{\sqrt{49}15}) = P(z < -2.33) = 1 - 0.9901 = 0.0099$.

**EXAMPLE 2**
We are given that $\mu = 2.4, \sigma = 2, n = 100$. There are 250 tickets available, so the 100 students will be able to purchase the tickets they want if all together ask for less than 250 tickets. The probability for that is $P(T < 250) = P(z < \frac{250-100(2.4)}{\sqrt{100}2}) = P(z < 0.5) = 0.6915$.

**EXAMPLE 3**
We are given $\mu = 500, \sigma = 80, n = 100$.

a. $P(490 < \bar{x} < 510) = P(\frac{490-500}{\frac{80}{\sqrt{100}}} < z < \frac{490-500}{\frac{80}{\sqrt{100}}}) = P(-1.25 < z < 1.25) = 0.8944 - (1 - 0.8944) = 0.7888$.

b. $\pm 1.96 = \frac{\bar{x}-500}{\frac{80}{\sqrt{100}}} \Rightarrow \bar{x} = 484.32, \bar{x} = 515.68$. Therefore $P(484.32 < \bar{x} < 515.68) = 0.95$.

**EXAMPLE 4**
We are given that $\mu = 50000, \sigma = 10000, n = 11$. The starting supply is 74000 gallons and the weekly delivery is 47000 gallons. Therefore the total supply for the 11-week period is $74000 + 11 \times 47000 = 591000$ gallons.

a. The supply will be below 20000 gallons if the total gasoline purchased in these 11 weeks is more than $591000 - 20000 = 571000$ gallons. Therefore we need to find $P(T > 571000) = P(z > \frac{571000-11(50000)}{\sqrt{11}10000}) = P(z > 0.63) = 1 - 0.7357 = 0.2643$.

b. Let $A$ be the unknown schedule delivery. Now the total gasoline purchased must be more than $74000 + 11 \times A - 20000$. We want this with probability 0.5%, or $P(T > 74000 + 11A - 20000) = 0.005$. The $z$ value that corresponds to this probability is 2.575. So, $2.575 = \frac{74000+11A-20000-11(50000)}{\sqrt{11}10000} \Rightarrow A = 52854.8$. The weekly delivery must be 52854.8 gallons.

## Central limit theorem - proof

For the proof below we will use the following theorem.
**Theorem:**
Let $X_n$ be a random variable with moment generating function $M_{X_n}(t)$ and $X$ be a random variable with moment generating function $M_X(t)$. If

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$$

then the distribution function (cdf) of $X_n$ converges to the distribution function of $X$ as $n \to \infty$.

**Central limit theorem:**
If $X_1, X_2, \cdots, X_n$ are i.i.d. (independent and identically distributed) random variables having the same distribution with mean $\mu$, variance $\sigma^2$, and moment generating function $M_X(t)$, then if $n \to \infty$ the limiting distribution of the random variable $Z = \frac{T-n\mu}{\sigma\sqrt{n}}$ (where $T = X_1 + X_2 + \cdots + X_n$) is the standard normal distribution $N(0,1)$.

Proof:

$$M_Z(t) = M_{\frac{T-n\mu}{\sigma\sqrt{n}}}(t) = Ee^{\frac{T-n\mu}{\sigma\sqrt{n}}t} = e^{-\frac{n\mu}{\sigma\sqrt{n}}t} M_T(\frac{t}{\sigma\sqrt{n}})$$

But $T = X_1 + X_2 + \cdots + X_n$. From earlier discussion the mgf of the sum is equal to the product of the individual mgf. Here each $X_i$ has mgf $M_X(t)$. Therefore,

$$M_T(\frac{t}{\sigma\sqrt{n}}) = \left[M_X(\frac{t}{\sigma\sqrt{n}})\right]^n$$

and so $M_Z(t)$ is equal to

$$M_Z(t) = e^{-\frac{n\mu}{\sigma\sqrt{n}}t}\left[M_X(\frac{t}{\sigma\sqrt{n}})\right]^n$$

One way to find the limit of $M_Z(t)$ as $n \to \infty$ is to consider the logarithm of $M_Z(t)$:

$$ln\ M_Z(t) = -\frac{\sqrt{n}\ \mu}{\sigma}t + n\ ln\ M_X(\frac{t}{\sigma\sqrt{n}})$$

Expanding $M_X(\frac{t}{\sigma\sqrt{n}})$, using the following (also see handout on mgf)

$$M_X(t) = \sum_x P(x) + \frac{t}{1!}\sum_x xP(x) + \frac{t^2}{2!}\sum_x x^2P(x) + \frac{t^3}{3!}\sum_x x^3P(x) + \cdots$$

we get

$$ln\ M_Z(t) = -\frac{\sqrt{n}\ \mu}{\sigma}t + n\left(ln\left[1 + \frac{\frac{t}{\sigma\sqrt{n}}}{1!}EX + \frac{(\frac{t}{\sigma\sqrt{n}})^2}{2!}EX^2 + \frac{(\frac{t}{\sigma\sqrt{n}})^3}{3!}EX^3 + \cdots\right]\right)$$

Now using the series expansion of $ln(1+y) = y - \frac{y^2}{2} + \frac{y^3}{3} - \frac{y^4}{4} + \cdots$ where $y = \frac{\frac{t}{\sigma\sqrt{n}}}{1!}EX + \frac{(\frac{t}{\sigma\sqrt{n}})^2}{2!}EX^2 + \frac{(\frac{t}{\sigma\sqrt{n}})^3}{3!}EX^3 + \cdots$ we get:

$$\begin{aligned} ln\ M_Z(t) = & - \frac{\sqrt{n}\ \mu}{\sigma}t + n\left[\frac{\frac{t}{\sigma\sqrt{n}}}{1!}EX + \frac{(\frac{t}{\sigma\sqrt{n}})^2}{2!}EX^2 + \frac{(\frac{t}{\sigma\sqrt{n}})^3}{3!}EX^3 + \cdots\right] \\ & - \frac{1}{2}\left[\frac{\frac{t}{\sigma\sqrt{n}}}{1!}EX + \frac{(\frac{t}{\sigma\sqrt{n}})^2}{2!}EX^2 + \frac{(\frac{t}{\sigma\sqrt{n}})^3}{3!}EX^3 + \cdots\right]^2 \\ & + \frac{1}{3}\left[\frac{\frac{t}{\sigma\sqrt{n}}}{1!}EX + \frac{(\frac{t}{\sigma\sqrt{n}})^2}{2!}EX^2 + \frac{(\frac{t}{\sigma\sqrt{n}})^3}{3!}EX^3 + \cdots\right]^3 - \cdots \end{aligned}$$

Factoring out the powers of $t$ we obtain:

$$\begin{aligned} ln\ M_Z(t) &= \left(-\frac{\sqrt{n}\ \mu}{\sigma} + \frac{\sqrt{n}\ EX}{\sigma}\right)t + \left(\frac{EX^2}{2\sigma^2} - \frac{(EX)^2}{2\sigma^2}\right)t^2 \\ &+ \left(\frac{EX^3}{6\sigma^3\sqrt{n}} - \frac{EX\ EX^2}{2\sigma^3\sqrt{n}} + \frac{(EX)^3}{3\sigma^3\sqrt{n}}\right)t^3 + \cdots \end{aligned}$$

Because $EX = \mu$ and $EX^2 - (EX)^2 = \sigma^2$ the last expression becomes

$$ln\ M_Z(t) = \frac{1}{2}t^2 + \left(\frac{EX^3}{6} - \frac{EX\ EX^2}{2} + \frac{(EX)^3}{3}\right)\frac{t^3}{\sigma^3\sqrt{n}} + \cdots$$

We observe that as $n \to \infty$ the limit of the previous expression is

$$\lim_{n\to\infty} ln\ M_Z(t) = \frac{1}{2}t^2$$

and therefore

$$\lim_{n\to\infty} M_Z(t) = e^{\frac{1}{2}t^2}.$$

But this is the mgf of the standard normal distribution. Therefore the limiting distribution of $\frac{T-n\mu}{\sigma\sqrt{n}}$ is the standard normal distribution $N(0,1)$.

♣♡♠♢ ♣♡♠♢ ♣♡♠♢ ♣♡♠♢ ♣♡♠♢

## Central limit theorem - Examples

**Example 1**
A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean $\mu = 205$ pounds and standard deviation $\sigma = 15$ pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

**Example 2**
From past experience, it is known that the number of tickets purchased by a student standing in line at the ticket window for the football match of *UCLA* against *USC* follows a distribution that has mean $\mu = 2.4$ and standard deviation $\sigma = 2.0$. Suppose that few hours before the start of one of these matches there are 100 eager students standing in line to purchase tickets. If only 250 tickets remain, what is the probability that all 100 students will be able to purchase the tickets they desire?

**Example 3**
Suppose that you have a sample of 100 values from a population with mean $\mu = 500$ and with standard deviation $\sigma = 80$.

a. What is the probability that the sample mean will be in the interval $(490, 510)$?

b. Give an interval that covers the middle 95% of the distribution of the sample mean.

**Example 4**
The amount of mineral water consumed by a person per day on the job is normally distributed with mean 19 ounces and standard deviation 5 ounces. A company supplies its employees with 2000 ounces of mineral water daily. The company has 100 employees.

a. Find the probability that the mineral water supplied by the company will not satisfy the water demanded by its employees.

b. Find the probability that in the next 4 days the company will not satisfy the water demanded by its employees on at least 1 of these 4 days. Assume that the amount of mineral water consumed by the employees of the company is independent from day to day.

c. Find the probability that during the next year (365 days) the company will not satisfy the water demanded by its employees on more than 15 days.

**Example 5**
Supply responses *true* or *false* with an explanation to each of the following:

a. The probability that the average of 20 values will be within 0.4 standard deviations of the population mean exceeds the probability that the average of 40 values will be within 0.4 standard deviations of the population mean.

b. $P(\bar{X} > 4)$ is larger than $P(X > 4)$ if $X \sim N(8, \sigma)$.

c. If $\bar{X}$ is the average of $n$ values sampled from a normal distribution with mean $\mu$ and if $c$ is any positive number, then $P(\mu - c \leq \bar{X} \leq \mu + c)$ decreases as $n$ gets large.

**Example 6**

An insurance company wants to audit health insurance claims in its very large database of transactions. In a quick attempt to assess the level of overstatement of this database, the insurance company selects at random 400 items from the database (each item represents a dollar amount). Suppose that the population mean overstatement of the entire database is \$8, with population standard deviation \$20.

a. Find the probability that the sample mean of the 400 would be less than \$6.50.

b. The population from where the sample of 400 was selected does not follow the normal distribution. Why?

c. Why can we use the normal distribution in obtaining an answer to part (a)?

d. For what value of $\omega$ can we say that $P(\mu - \omega < \bar{X} < \mu + \omega)$ is equal to 80%?

e. Let $T$ be the total overstatement for the 400 randomly selected items. Find the number $b$ so that $P(T > b) = 0.975$.

**Example 7**

A telephone company has determined that during nonholidays the number of phone calls that pass through the main branch office each hour follows the normal distribution with mean $\mu = 80000$ and standard deviation $\sigma = 35000$. Suppose that a random sample of 60 nonholiday hours is selected and the sample mean $\bar{x}$ of the incoming phone calls is computed.

a. Describe the distribution of $\bar{x}$.

b. Find the probability that the sample mean $\bar{x}$ of the incoming phone calls for these 60 hours is larger than 91970.

c. Is it more likely that the sample average $\bar{x}$ will be greater than 75000 hours, or that one hour's incoming calls will be?

**Example 8**

Assume that the daily $S\&P$ return follows the normal distribution with mean $\mu = 0.00032$ and standard deviation $\sigma = 0.00859$.

a. Find the $75_{th}$ percentile of this distribution.

b. What is the probability that in 2 of the following 5 days, the daily $S\&P$ return will be larger than 0.01?

c. Consider the sample average $S\&P$ of a random sample of 20 days.

   i. What is the distribution of the sample mean?

   ii. What is the probability that the sample mean will be larger than 0.005?

   iii. Is it more likely that the sample average $S\&P$ will be greater than 0.007, or that one day's $S\&P$ return will be?

## The central limit theorem
## The distribution of the sample mean
## The distribution of the sum
## Summary

Suppose a population has mean $\mu$ and standard deviation $\sigma$. Let $X_1, X_2, \cdots, X_n$ be an i.i.d. (independent and identically distributed) sample from this population. This means that $E(X_i) = \mu$, and $Var(X_i) = \sigma^2$. Define the following random variables:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}.$$

and

$$T = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^n X_i.$$

Then, for large $n$ (usually $n \geq 30$) the following statements are *approximately* true regardless of the shape of the population:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

therefore

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

and

$$T \sim N(n\mu, \sigma\sqrt{n})$$

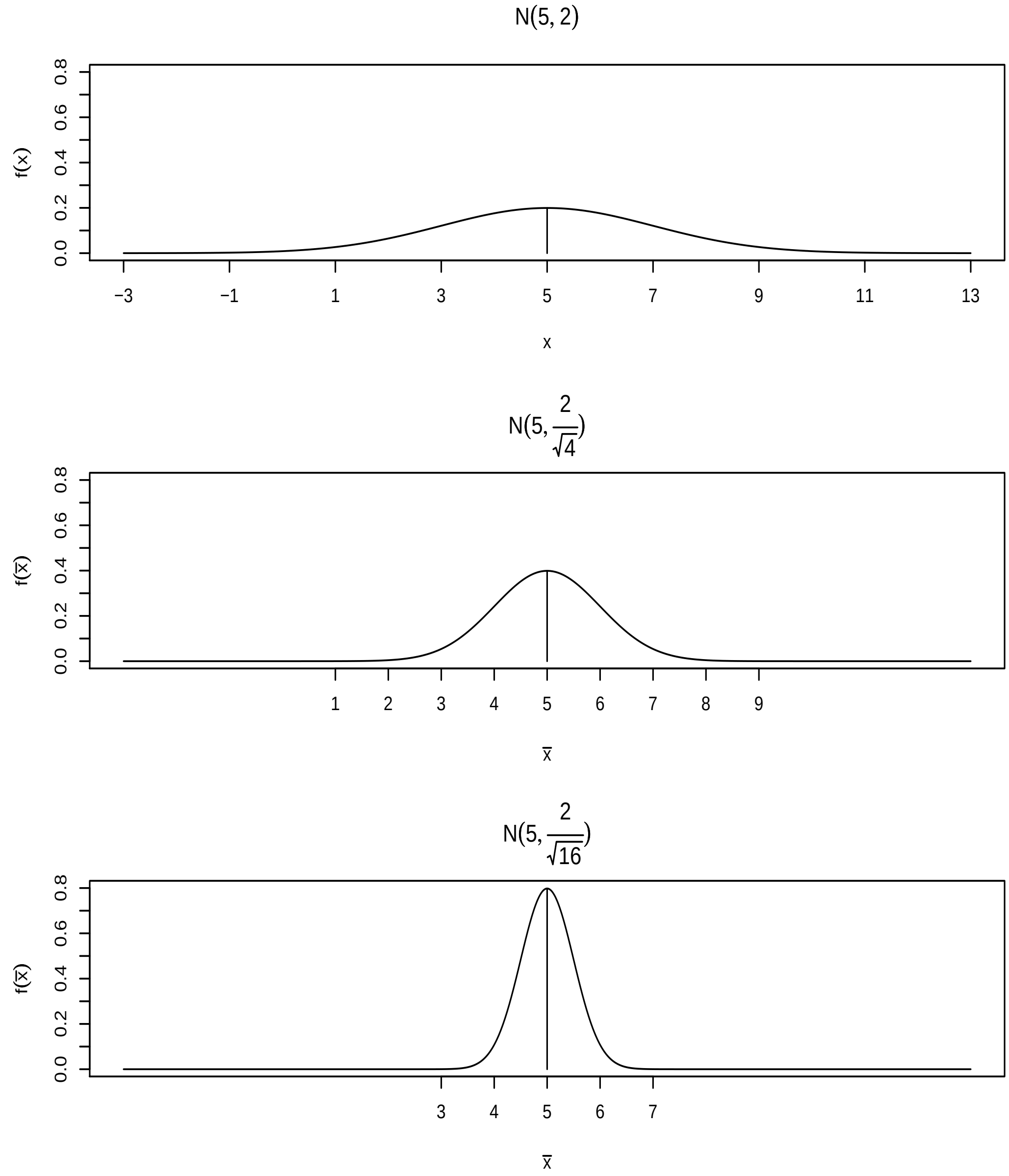
therefore,

$$Z = \frac{T - n\mu}{\sigma\sqrt{n}}$$

**Note:** If the population from where the sample is selected follows the normal distribution, the above statements are *exactly* true regardless of the sample size. In this case $n$ can be small or large (see next page).

## Distribution of the sample mean - Sampling from normal distribution

If we sample from normal distribution $N(\mu, \sigma)$ then $\bar{X}$ follows exactly the normal distribution with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$ regardless of the sample size $n$. In the next figure we see the effect of the sample size on the shape of the distribution of $\bar{X}$. The first figure is the $N(5, 2)$ distribution. The second figure represents the distribution of $\bar{X}$ when $n = 4$. The second figure represents the distribution of $\bar{X}$ when $n = 16$.



## Central limit theorem - More problems

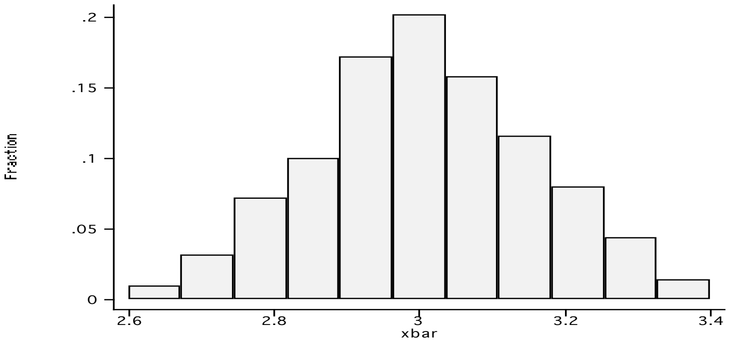
**Problem 1**
**Part A:**
It is claimed that the histogram below shows the distribution of the sample mean $\bar{x}$, when repeated samples of size $n = 36$ are selected with replacement from the population $(2.6, 2.8, 3.0, 3.2, 3.4)$. Clearly explain if there is anything wrong with this histogram.

**Part B:**

a. What distribution does the sum (total) of 36 observations selected from the same population as above follow?

b. Sketch the histogram (roughly) of the total of repeated samples (with repcalcement) of size 36 selected from the above population. Make sure that you mark off some important values on the horizontal axis.

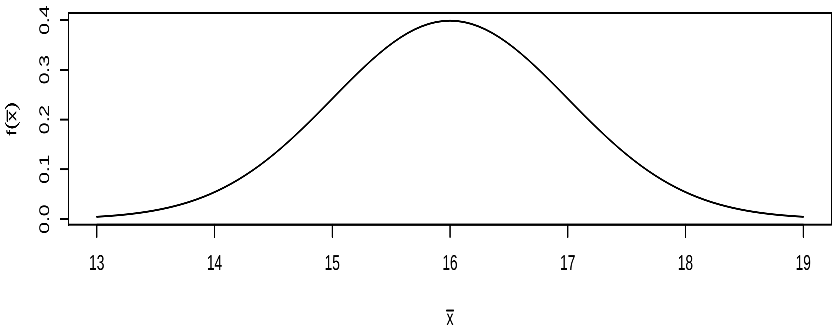
c. Find the $5_{th}$ percentile of the distribution of $T$.



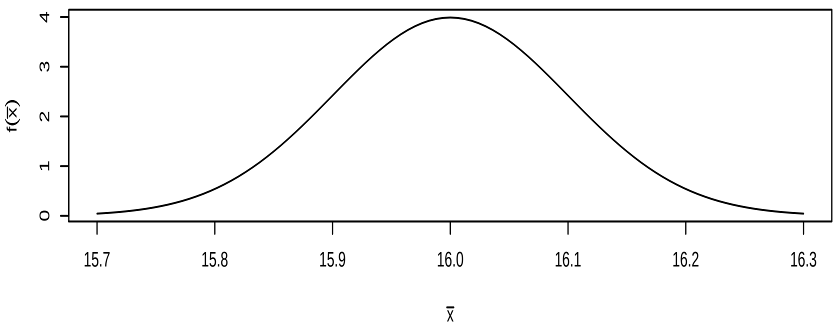
**Problem 2**

A random sample of size $n = 100$ is selected form a distribution with mean $\mu = 16$ and standard deviation $\sigma = 4$. Which one of the graphs below represents the distribution of the sample mean. Please explain your answer.

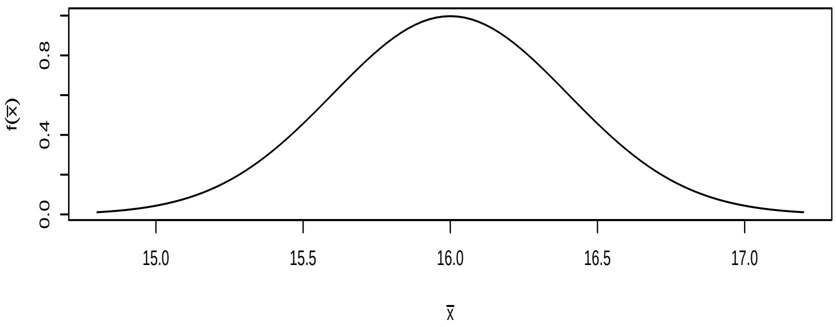
**(a)**



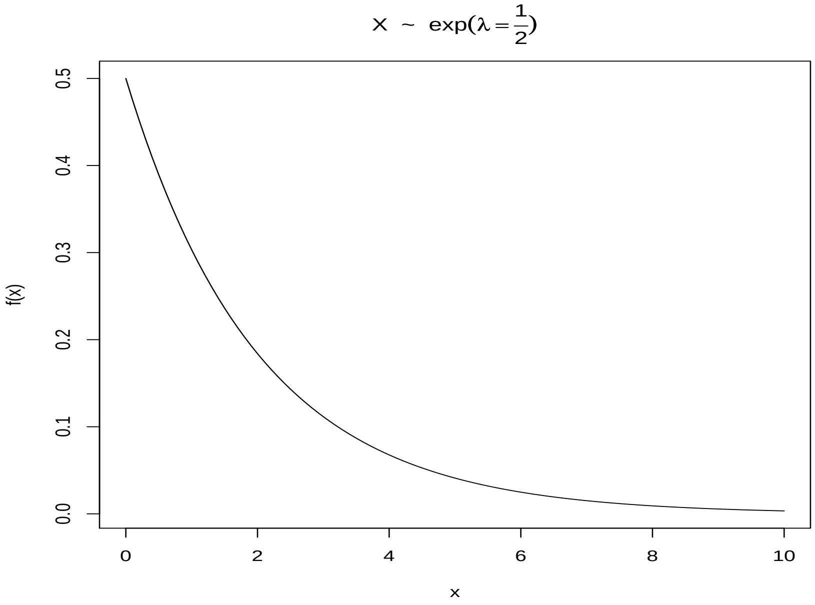
**(b)**

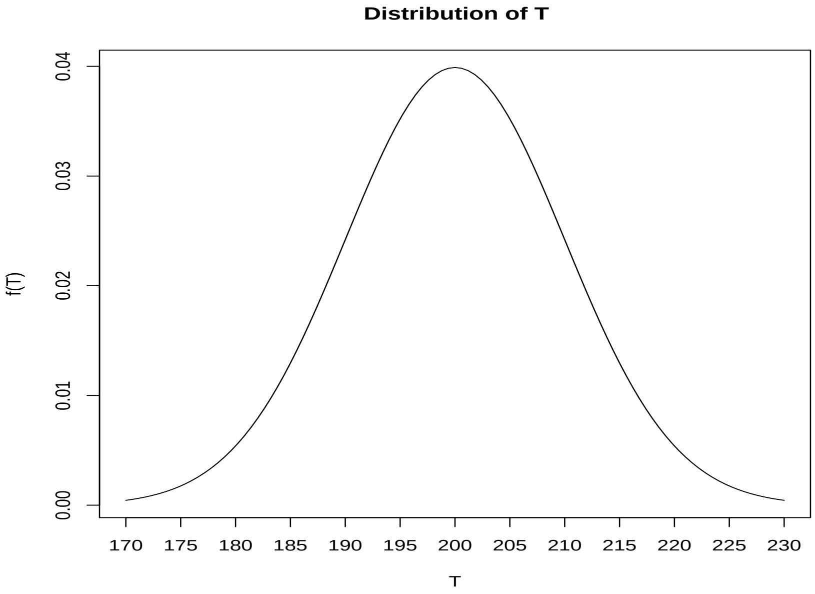


**(c)**



**Problem 3**
Below you see the probability density function of an exponential distribution with parameter $\lambda = \frac{1}{2}$. It is claimed that the second graph is the distribution of the total $T$ when a sample of size $n = 100$ is selected. Is there anything wrong with this graph?





## The distribution of the sample mean and the central limit theorem
## An empirical investigation

The central limit theorem states that if a large sample of size $n$ is selected from a population that hasm mean $\mu$ and standard deviation $\sigma$ then the sample mean $\bar{X}$ follows approximately:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

In this experiment we will roll a die $n = 80$ times. Using the 80 outcomes we will compute the sample mean. We will repeat until we obtain 500 values of $\bar{x}$. At the end we will construct the histogram using the values of $\bar{x}$. Here is the population:

| $X$ | $P(X)$ |
|---|---|
| 1 | $\frac{1}{6}$ |
| 2 | $\frac{1}{6}$ |
| 3 | $\frac{1}{6}$ |
| 4 | $\frac{1}{6}$ |
| 5 | $\frac{1}{6}$ |
| 6 | $\frac{1}{6}$ |

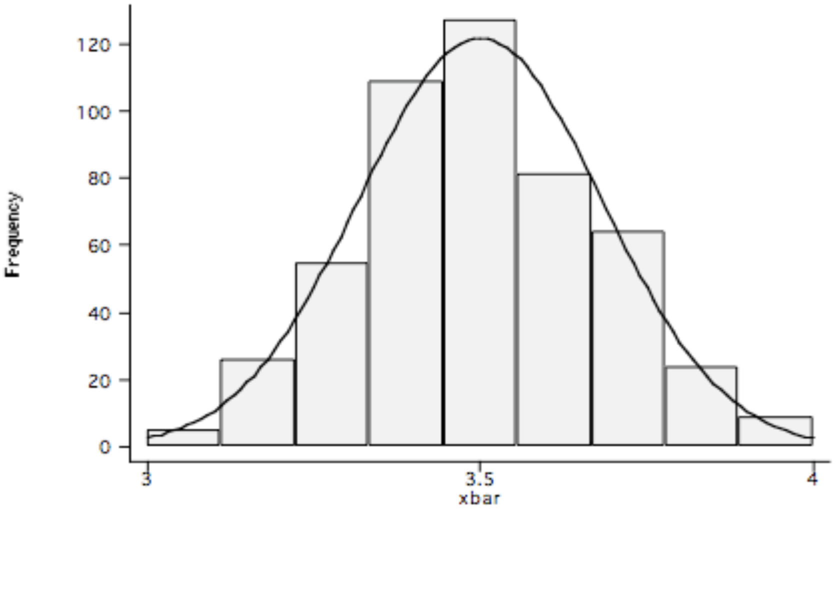
This population has $\mu = 3.5$ and $\sigma = 1.71$. Why?

The summary statistics of the 500 sample means:

```
. summarize xbar

    Variable |      Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        xbar |      500    3.499725    .1822218      3.025          4
```

And below you can see the histogram of these 500 sample means:


Frequency
xbar