Fatemeh Nadi
810101285
fatemehnadi@ut.ac.ir

Homework 3
Statistical Inference , Fall 1401

December 1, 2022

1

$n = 81,\ \bar{x} = \$800000,\ \sigma = \$90000$

Checking conditions:

1. Independence: random sample and $n < 10\%$ of all houses in New York.
2. Sample size/skew: $n \geq 30$.

So, we can assume that the sampling distribution of average house price in New York from samples of size 81 will be nearly normal.

Central Limit Theorem (CLT)

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right) = N(800000, 10000)$$

General Form of Confidence Interval

$$point\ estimate\ \pm\ margin\ of\ error$$
$$\bar{x} \pm z^*\ SE$$

- a
  CI = 98% $\rightarrow\ (1 - 0.98)/2 = 0.01$

  $> qnorm\,(0.01) = 2.32$

  approximate 98% CI for μ: 800000 $\pm$2.32 SE

  $$\boxed{(776800, 823200)}$$

  We are 98% confidence that the mean point of house price in New York is between \$776800 and \$823200.

- b
  CI = 95% $\rightarrow\ (1 - 0.95)/2 = 0.025$

  $> qnorm\,(0.025) = 1.96$

  approximate 95% CI for μ: 800000 $\pm$1.96 SE

  $$\boxed{(780400,\ 819600)}$$

  We are 95% confidence that the mean point of house price in New York is between \$780400 and \$819600.

- c
  CI = 90% $\rightarrow$ $(1-0.9)/2 = 0.05$

  $> qnorm\,(0.05) = 1.64$

  approximate 90% CI for μ: 800000 $\pm$1.64 SE

  $$\boxed{(783600,\ 816400)}$$

  We are 90% confidence that the mean point of house price in New York is between \$783600 and \$816400.

- d
  CI = 50% $\rightarrow$ $(1-0.5)/2 = 0.25$

  $> qnorm\,(0.25) = .67$

  approximate 50% CI for μ: 800000 $\pm$0.67 SE

  $$\boxed{(793300,\ 806700)}$$

  We are 50% confidence that the mean point of house price in New York is between \$793300 and \$806700.

- e
  A range of values so defined that there is a specified probability that the value of a parameter lies within it.

- f
  Decreasing the level of confidence can narrow your confidence interval.

- g
  $margin\ of\ error = \$5000$
  CI = 99% $\rightarrow$ $(1-0.99)/2 = 0.005$

  $> qnorm\,(0.005) = 2.57$

  $2.57 \times \frac{90000}{\sqrt{n}} = 5000 \rightarrow n \geq \boxed{2140}$

- h
  $margin\ of\ error = \$2500$
  The margin of error is halved. As a result, n must multiply by 4.

  $$\begin{aligned} \frac{1}{2}ME &= \frac{1}{2}z^*\frac{s}{\sqrt{n}} \\ &= z^*\frac{s}{\sqrt{4n}} \\ &\rightarrow n \geq \boxed{8560} \end{aligned}$$

2

$n = 70$ *high school teens*, $\bar{x} = 10$ *hours*, $\mu = 7$ *hours*

$$H_0 : \boxed{x <} 7 \quad \textit{We never ever test a sample statistic. We KNOW everything about the sample,}$$
$$\textit{so we don't need to test it. Thus they should be } \mu, \textit{ not } x.$$
$$\textit{On average, high school teens spend almost seven hours each weekday}-$$
$$\textit{on educational activities so must use } = \textit{ not } < .$$
$$H_a : \boxed{x > 10} \quad \textit{We never ever test a sample statistic. We should use } \mu \textit{instead of } x \textit{ valus.}$$

Let's define null hypothesis (H0) and alternative hypothesis (HA):

$$H_0 : \mu = 7 \quad \textit{on average, high school teens spend seven hours each weekday.}$$
$$H_a : \mu > 7 \quad \textit{on average, high school teens have spent more than 7 hours each weekday.}$$

and p-value like:

$$\begin{aligned} p - value &= P\left(\textit{observed or more extreme outcome} \mid H_0 \textit{ true}\right) \\ &= P\left(\bar{x} \geq 10 \mid H_0 : \ \mu = 7\right) \\ &= P\left(\bar{z} \geq \frac{10 - 7}{\frac{s}{\sqrt{70}}}\right) \end{aligned}$$

According to the value of the p-value, we can decide about the result of this test reject H0 or not.

3

$n = 20,\ \bar{x} = 4.6,\ sd = 2.2,\ \alpha = 0.05, \mu = 5\ years$

Checking conditions:

1. Independence: random sample and $n < 10\%$ of all children from the city which is renowned for its music school. it is a small city!
2. Sample size/skew: $n \geq 30$.

So, we can't assume that the sampling distribution of average year from learn music in this school from samples of size 20 will be nearly normal.
But we can use Student's t-distribution for this problem.

Estimating the Mean

$$point\ estimate\ \pm\ margin\ of\ error$$
$$\bar{x} \pm t^*_{df}\ SE$$

$df = n - 1 = 19$

$SE = \frac{s}{\sqrt{n}} = \frac{2.2}{\sqrt{20}} \simeq 0.491$

$$\begin{aligned} H_0 &: \mu = 5 \quad on\ average,\ child\ from\ this\ city\ takes\ at\ least\ 5\ years\ of\ piano. \\ H_a &: \mu < 5 \quad on\ average,\ child\ from\ this\ city\ takes\ less\ than\ 5\ years\ of\ piano. \end{aligned}$$

$t = \frac{observation - null}{SE}$

and p-value like:

$$\begin{aligned} p - value &= P\left(observed\ or\ more\ extreme\ outcome \mid H_0\ true\right) \\ &= P\left(\bar{x} < 4.6 \mid H_0:\ \mu = 5\right) \\ &= P\left(T < \frac{4.6 - 5}{0.491}\right) \\ &= P\left(T < -0.814\right) \end{aligned}$$

- a

  $> pt\left(-0.814, df = 19\right) = 0.21 \not< \alpha = 0.05 \rightarrow$ Can't reject H_0 in favor of H_a.

- b

  CI = 95% $\rightarrow\ \left(1 - 0.95\right)/2 = 0.025$

  $> qt\left(0.025, df = 19\right) = -2.093$

  approximate 95% CI for μ: 4.6 $\pm$2.093$\times$SE

$$\boxed{(3.57, 5.62)}$$

- c

  Yes, because μ is this range we can't reject H_0 hypothesis.

4

$n = 52,\ \bar{x} = 98.2846,\ s = 0.6824, \mu = 98.6$

- a

  Checking conditions:

  1. Independence: random sample and $n < 10\%$ of all healthy adults.
  2. Sample size/skew: $n \geq 30$.

  So, we can assume that the sampling distribution of average the body temperature from samples of size 52 healthy adults will be nearly normal.

  $SE = \frac{s}{\sqrt{n}} = \frac{0.6824}{\sqrt{52}} = 0.0946$

  Central Limit Theorem (CLT)

  $$\bar{x} \sim N\,(mean = \mu, SE) = N(98.6, 0.0946)$$

- b

  We have a random variable with a Normal Distribution. We should standardize it with Z score:

  $Z = \frac{observation - \mu}{SE} = \frac{98.2846 - 98.6}{0.0946} = -3.334$

  Setting the hypothesis:

  $H_0 : \mu = 98.6 \quad on\ average,\ the\ normal\ body\ temperature\ in\ degrees\ Fahrenheit\ is\ 98.6.$
  $H_a : \mu > 98.6 \quad on\ average,\ the\ normal\ body\ temperature\ in\ degrees\ Fahrenheit\ is\ more\ than\ 98.6.$

  and p-value like:

  $$\begin{aligned} p - value &= P\,(observed\ or\ more\ extreme\ outcome \mid H_0\ true) \\ &= P\,(\bar{x} \geq 98.2846 \mid H_0 :\ \mu = 98.6) \\ &= P\,(Z \geq -3.334) \\ &\simeq 0 < \alpha = 0.05 \end{aligned}$$

  Null Hypothesis is rejected in favor of H_a. we can believes that this temperature is an underestimate and the normal body temperature in degrees Fahrenheit is more than 98.6.

- c

  General Form of Confidence Interval

$$\begin{gathered} point\ estimate\ \pm\ margin\ of\ error \\ \bar{x} \pm z^*\ SE \end{gathered}$$

  CI = 98% $\rightarrow\ (1 - 0.98)/2 = 0.01$

  $> qnorm\,(0.01) = 2.32$

  approximate 98% CI for μ: 98.2846 $\pm$2.32SE

$$\boxed{(98.0646,\ 98.2846)}$$

  We are 98% confidence that the mean point of the normal body temperature in degrees Fahrenheit is between 98.0646 and 98.2846.

- d

  Setting two-sided the hypothesis:

  $H_0 : \mu = 98.6 \quad on\ average,\ the\ normal\ body\ temperature\ in\ degrees\ Fahrenheit\ is\ 98.6.$
  $H_a : \mu \neq 98.6 \quad on\ average,\ the\ normal\ body\ temperature\ in\ degrees\ Fahrenheit\ is\ notequal\ 98.6.$

  and p-value like:

$$\begin{aligned} p - value &= P\left(-98.2846 < \bar{x} < 98.2846 \mid H_0 :\ \mu = 98.6\right) \\ &= P\left(Z > -3.334\right) + P\left(Z < -3.334\right) \\ &\simeq 0 < \alpha = 0.05 \end{aligned}$$

  Null Hypothesis is rejected in favor of H_a. we can believes that the normal body temperature in degrees Fahrenheit is not 98.6.

5

Suppose that $X \sim Binomial\,(100, p)$, so $n = 100$ and according to the hypothesis $p = 0.5$ therefore $q = 1 - p = 0.5$.

- a

  Using the normal to approximate this we have that $\mu = np = 50$ and $\sigma^2 = npq = 25$ and so we have:

  $$X \sim N\left(X|\mu, \sigma^2\right) = N(100, 25)$$

  We know that α is the probability of rejecting the null when it is true. Since we have that H_0 is true then $X \sim Binomial\,(100, \hat{p} = 0.5)$, and so we can write the following:

  $$\begin{aligned}
  \alpha &= P\left(|X - 50| > 10\right) \\
  &= P\left(\frac{|X - 50|}{5} > 2\right) \\
  &= P\left(\frac{|X - 50|}{\sqrt{25}} > 2\right) \\
  &= P\left(\frac{|X - \mu|}{\sigma} > 2\right) \\
  &= P\left(Z > 2 \;or\; Z < -2\right) \\
  &= P\left(Z > 2\ \right) + P\left(Z < -2\ \right) \\
  &= 2 \times pnorm(-2) = 0.0455
  \end{aligned}$$

  so $\alpha \geq 0.0455$ for rejecting H_0 in this situation.

- b

The power is the probability of correctly rejecting the null given H_A is true. We write this as:

$$\beta = 1 - P(|X - 50| > 10) = 1 - P(40 < X < 60)$$

```
n = 100
x <- seq(-1, 1, by = .01)
curve(1 - ( pnorm(60, mean=n*x, sd=sqrt(n*x*(1-x))) +
               pnorm(40, mean=n*x, sd=sqrt(n*x*(1-x)),lower.tail = FALSE )))
```

Figure 1: Q5-b code

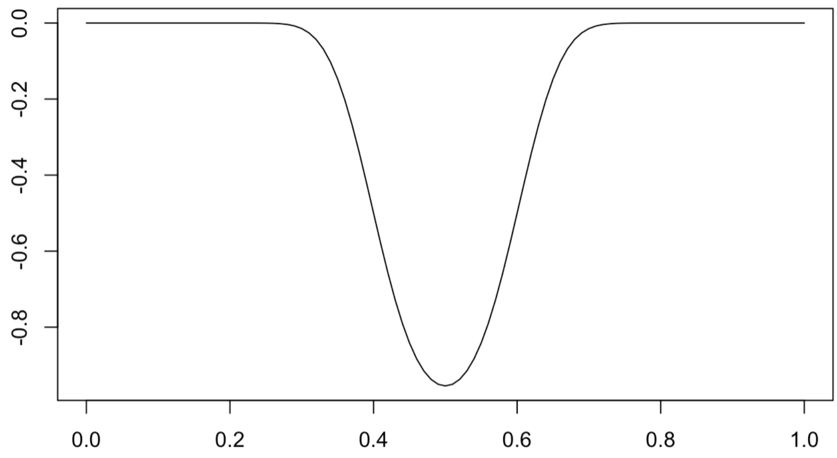

Figure 2: Q5-b curve of the power function of p

6

$n = 50,\ \bar{y} = 25.9,\ s = 5.6$

Checking conditions:

1. Independence: random sample and $n < 10\%$ of all population of interest.
2. Sample size/skew: $n \geq 30$.

So, we can assume that the sampling distribution of size 50 will be nearly normal.

Central Limit Theorem (CLT)

$$\bar{y} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right) = N(28, 0.79)$$

$$H_0 : \mu \geq 28$$
$$H_a : \mu < 28$$

- a

  $Z = \frac{observation - \mu}{SE} = \frac{25.9-28}{0.79} = -2.65$

  and p-value like:

  $$\begin{aligned} p - value &= P\left(\bar{y} < 25.9 \mid H_0 :\ \mu = 28\right) \\ &= P\left(Z < -2.65\right) \\ &\simeq 0.004 < \alpha = 0.05 \end{aligned}$$

  Null Hypothesis is rejected in favor of H_a.

- b

  Let us suppose that the actual mean number is 27 so:

  $$\mu_a = 27 \rightarrow \bar{y} \sim N\left(\mu = 27, SE = \frac{s}{\sqrt{n}} = 0.79\right)$$

  $\alpha = 0.05,\ one\ sided\ test \rightarrow z_a = ?$

  $P\left(Z < z_a\right) = 0.05 \rightarrow z_a = qnorm(0.05) = -1.644$

  $$\begin{aligned} Type\ \mathbb{II}\ Error = \beta &= P\left(fail\ to\ reject\ H_0 | \mu = \mu_a\right) \\ &= P\left(\frac{\bar{y}-28}{0.79} < -1.644 |\ \bar{y} \sim N\left(\mu = 27,\ SE = 0.79\right)\right) \\ &= P\left(\bar{y} < 26.70 |\ \bar{y} \sim N\left(\mu = 27,\ SE = 0.79\right)\right) \\ &= P\left(Z < \frac{26.7-27}{0.79}\right) \\ &= pnorm(-0.379) = 0.352 \end{aligned}$$

  $\boxed{Power = 1 - \beta = 0.647}$

- c

  No, because a type II error will be made if we fail to reject H_0, which means that confirms an idea that should have been rejected, such as, for instance, claiming that two observances are the same, despite them being different. A type II error does not reject the null hypothesis, even though the alternative hypothesis is the true state of nature. In other words, a false finding is accepted as true. as you see in part (a), the p-value is too small and we reject H_0.

7

- a

```
n = 50
alpha = 0.05
x <- seq(22,27,1)

mu = 28
s = 5.6

se = s/sqrt(n)
z_a = qnorm(alpha, 0, 1)

plot(x, pnorm( z_a - ((x-mu)/(se)) ), type="l", xlab = "Means", ylab = "Error")
```

Figure 3: Q7-a code

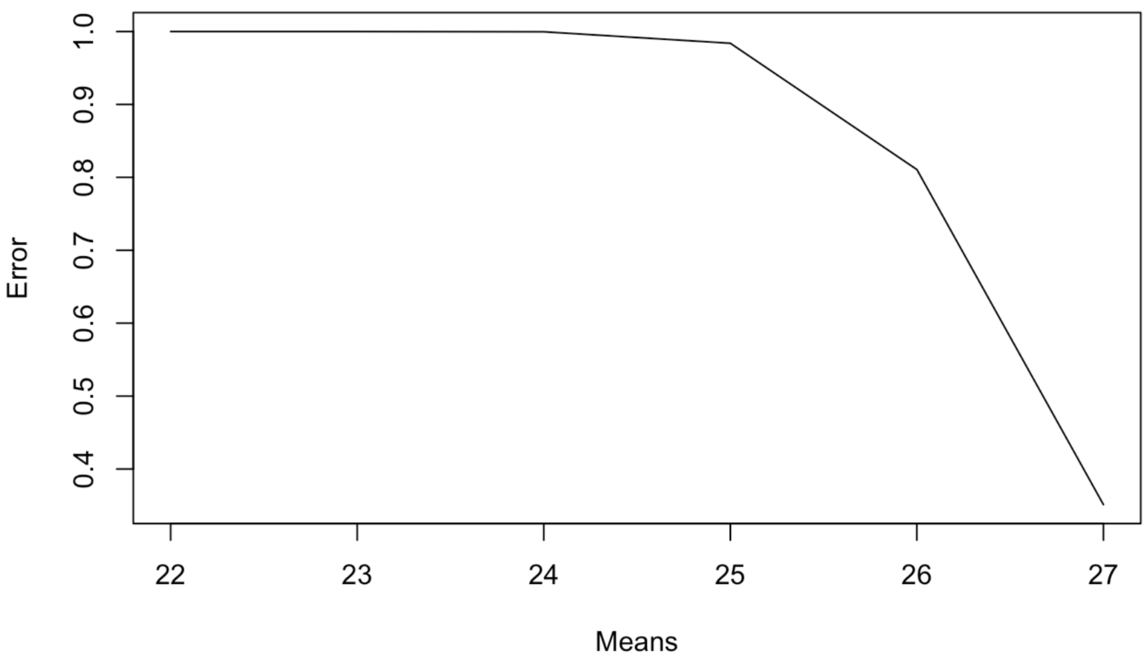

Figure 4: Q7-Error Type II curve

- b

```
n = 50
alpha = 0.05

mu = 28
s = 5.6
x <- seq(22,27,1)

se = s/sqrt(n)
z_a = qnorm(alpha, 0, 1)

plot(x, pnorm( z_a - ((x-mu)/(se)) ), type="l", xlab = "Means", ylab = "Error")

alpha<-0.01
n<-50

z_a_new = qnorm(alpha, 0, 1)
se_new = s/sqrt(n)

lines(x, pnorm( z_a_new - ((x-mu)/(se_new)) ), type="l", xlab = "Means", ylab = "Error", col='red')
```

Figure 5: Q7-b code

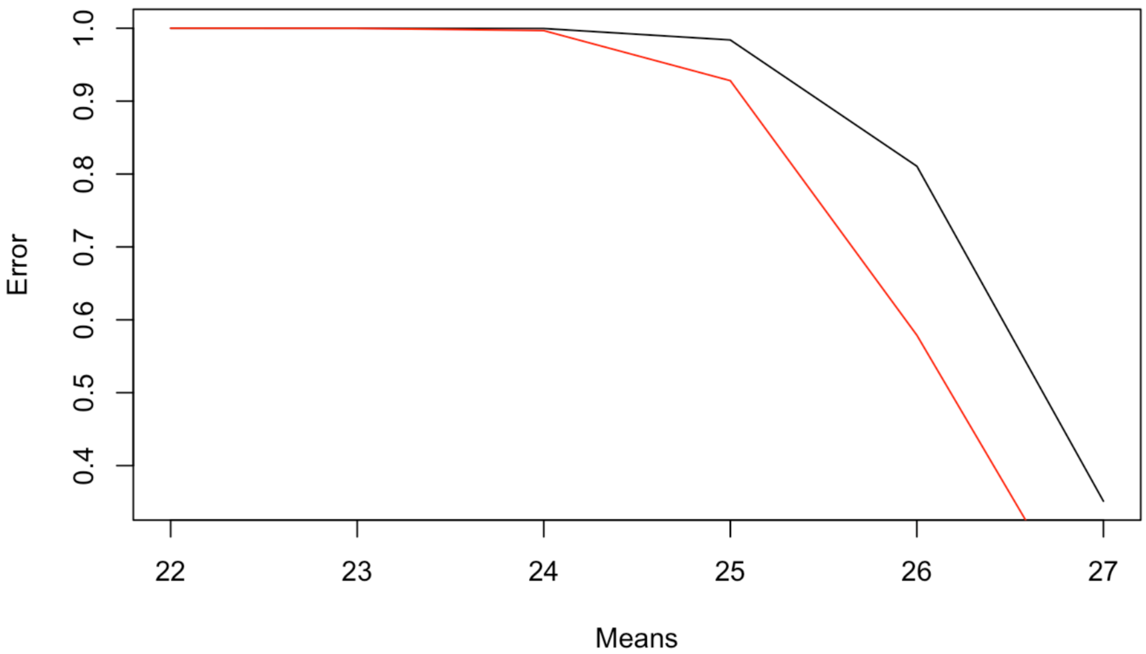

Figure 6: Q7-b Error Type II curve

- c

```r
n = 50
alpha = 0.05

mu = 28
s = 5.6
x <- seq(22,27,1)

se = s/sqrt(n)
z_a = qnorm(alpha, 0, 1)

plot(x, pnorm( z_a - ((x-mu)/(se)) ), type="l", xlab = "Means", ylab = "Error")

alpha<-0.05
n<-10

z_a_new = qnorm(alpha, 0, 1)
se_new = s/sqrt(n)

lines(x, pnorm( z_a_new - ((x-mu)/(se_new)) ), type="l", xlab = "Means", ylab = "Error", col='red')
```

Figure 7: Q7-c code

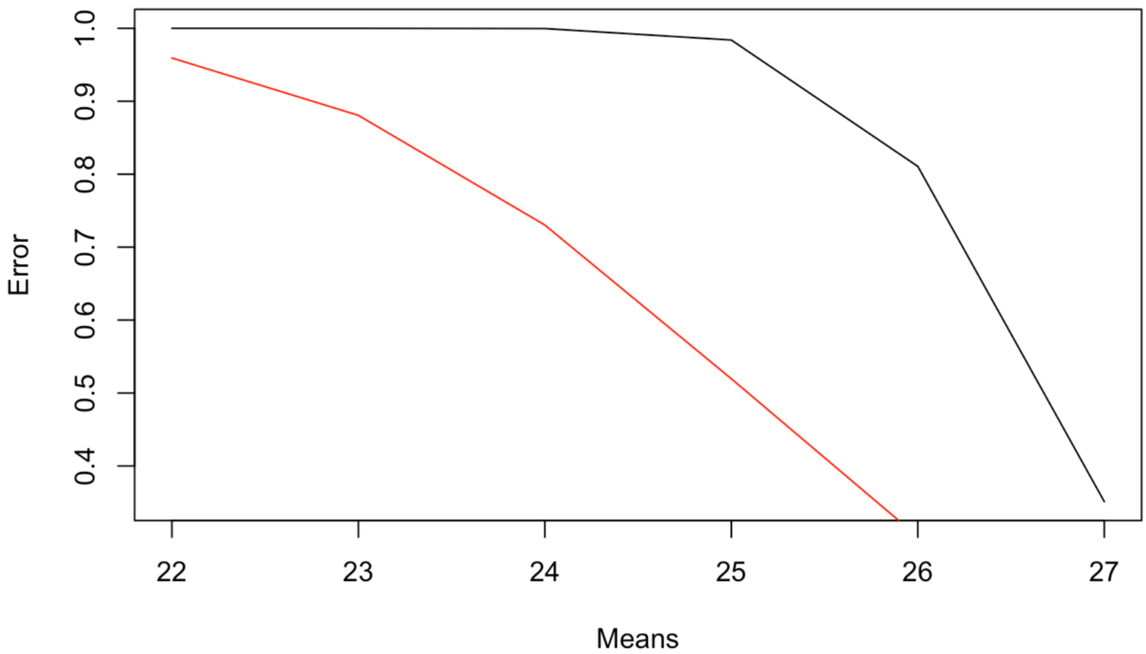

Figure 8: Q7-c Error Type II curve

8 R

Please see this file: "Q8-R.Rmd" and "Q8-R.html"
Code and explanation are provided.

- a

```r
k = 0
mu = mean(galton$child)
times = 2000

# seq(1, times, by=1)
for (i in 1:times) {

  n = 60
  #reads the dataset 'galton' and take the 60 rows as sample
  sdf<- sample(1:nrow(galton), n)

  #sample 60 rows
  sub_galton <- galton[sdf,]

  sub_galton.mean = mean(sub_galton$child)

  ci = 0.97
  z = qnorm((1-ci)/2)

  se <- sd(sub_galton$child) / sqrt(n) #SE = s/sqrt(n)

  lower <- sub_galton.mean + z * se
  upper <- sub_galton.mean - z * se

  if (mu>= lower &mu<=upper){
    k = k+1
  }
}
print((k/times)*100)
```

Figure 9: Q8-a code

```
## [1] 97.25
```

Figure 10: percentage of the intervals include the real mean of the society

- b

```
k = 0
mu = mean(galton$child)
times = 1000
# seq(1, times, by=1)
for (i in 1:times) {

  n = 10
  #reads the dataset 'galton' and take the 10 rows as sample
  sdf<- sample(1:nrow(galton), n)

  #sample 10 rows
  sub_galton <- galton[sdf,]

  sub_galton.mean = mean(sub_galton$child)

  ci = 0.9
  z = qnorm((1-ci)/2)

  se <- sd(sub_galton$child) / sqrt(n) #SE = s/sqrt(n)

  lower <- sub_galton.mean + z * se
  upper <- sub_galton.mean - z * se

  if (mu>= lower &mu<=upper){
    k = k+1
  }
}
print((k/times)*100)
```

Figure 11: Q8-b code

```
## [1] 88.3
```

Figure 12: percentage of the intervals include the real mean of the society

Conclusion: the percentage of the intervals includes the actual mean of the society close to the confidence interval.

- c

  Use normal distribution because it satisfies the condition of CLT.

```
n = 70
sdf<- sample(1:nrow(galton), n)
sub_galton <- galton[sdf,]

z <- (mean(sub_galton$parent)-60)/(sd(sub_galton$parent)/sqrt(length(sub_galton$parent)))
pvalue = 2*pnorm(-abs(z))

ci = 0.95
z = qnorm((1-ci)/2)

se <- sd(sub_galton$parent) / sqrt(n) #SE = s/sqrt(n)

lower <- sub_galton.mean + z * se
upper <- sub_galton.mean - z * se

muactual = mean(galton$parent)
s = sd(sub_galton$parent)

Zleft <- (lower-muactual)/(s/sqrt(n))
Zright <-(upper-muactual)/(s/sqrt(n))
b <- pnorm(Zright)-pnorm(Zleft)
power = 1-b
power
```

Figure 13: Q8-c code

```
## [1] 0.9949705
```

Figure 14: power

- d

  Use t'Student distribution because it does not satisfy the condition of CLT.

```
n = 10
sdf<- sample(1:nrow(galton), n)
sub_galton <- galton[sdf,]

t <- (mean(sub_galton$parent)-60)/(sd(sub_galton$parent)/sqrt(length(sub_galton$parent)))
pvalue = 2*pt(-abs(z),df=n-1)

ci = 0.95
t = qnorm((1-ci)/2)

se <- sd(sub_galton$parent) / sqrt(n) #SE = s/sqrt(n)

lower <- sub_galton.mean + t * se
upper <- sub_galton.mean - t * se

muactual = mean(galton$parent)
s = sd(sub_galton$parent)

Zleft <- (lower-muactual)/(s/sqrt(n))
Zright <-(upper-muactual)/(s/sqrt(n))
b <- pt(Zright, df = n-1)-pt(Zleft, df = n-1)

power = 1-b
power
```

Figure 15: Q8-d code

```
## [1] 0.5041667
```

Figure 16: power

- e
  It follows that the sample size decreases as the power decreases, which is a positive correlation(in this interval).
  For our test, ten samples do not fit the minimum requirement because we need more than ten samples to achieve statistical power(at least 80%) and effect size for our analysis.