

Data 606: Chapter-4 Homework

Student Name : Sachid Deshmukh

Date : 11/06/2018

- GitHub Location for rmd file
 - GitHub Location for pdf file
 - RPub's location of published file
-

```
library('DATA606')

## Loading required package: shiny
## Loading required package: openintro
## Please visit openintro.org for free statistics materials
##
## Attaching package: 'openintro'
## The following object is masked from 'package:datasets':
##
##   cars
## Loading required package: OIdata
## Loading required package: RCurl
## Loading required package: bitops
## Loading required package: maps
## Loading required package: ggplot2
## Loading required package: markdown
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
##
## Attaching package: 'DATA606'
```

```
## The following object is masked from 'package:utils':
##
##      demo
```

1. 4.4 Heights of adults

- Average height = 171.1 cm, Median = 170.3 cm
- Standard deviation = 9.4cm, Inter Quartile Range = 14 cm
- Z value of 180 = $(180-171.1)/9.4 = 0.95$ and Z value of 155 = $(155-171.1)/9.4 = -1.71$ The height of 180 is within one SD from mean. 68% of the data lies within one SD from mean. That way height of 180 is not uncommon. Height of 155 is within 2 SD from mean. It is definitely uncommon than 180 cm, however it is not unusually short
- Another sample will provide us different point estimates. They will represent overall population however point estimates will be slightly vary between samples
- Standard Error measures variability of estimates between samples. $SE = 9.4/\text{SQRT}(507) = 0.417$

2. 4.14 Thanksgiving spending Part1

- False. Point estimate of the sample will be always within confidence interval
- False. We have large sample size
- False. Confidence interval for one sample mean is not for another sample mean
- True. Population parameter is being estimated by the point estimate and the confidence interval
- False. General practice is to go with 95% confidence interval
- False. For decreasing margin of error by 3 we need to increase sample by 9
- True. Margin of error is half of the confidence interval. $(89.11-80.31)/2 = 4.4$

3. 4.24 Gifted Children Part-1

- Conditions for inferences are satisfied. Sample is random and sample size is greater than 30. Overall population doesn't have strong skew
- Null Hypothesis : $\mu \geq 32$ and Alternate Hypothesis : $\mu < 32$ $SE = 4.31/\text{sqrt}(36) = 0.72$ $Z \text{ Value} = (30.69-32)/0.72 = -1.81$ Conclusion : At 90% confidence interval critical value = -1.28. Since Z stat is lower than critical value, we reject null hypothesis in favor of alternate hypothesis. We can conclude that true mean is lesser than 32
- P value = 0.034. Assuming null hypothesis is true, probability of observing sample mean lesser than 30 is in only 3%
- 90% confidence interval = $30.69 \pm (1.65*0.72) = 29.50$ to 31.87
- From the confidence interval we can say that mean age for gifted childrens to count to 10 is between 29.5 and 31.87 months. This is inline with alternate hypothesis.

4. 4.26 Gifted Children Part-2

- Null hypothesis is $\mu = 100$ and Alternate hypothesis is $\mu \neq 100$ $SE = 6.5/\text{sqrt}(36) = 1.08$ $Z \text{ Value} = (118.2-100)/1.08 = 16.85$. At 90% confidence interval critical value is 1.64. Observed value 16 is way higher than critical value. We have to reject null hypothesis in favor of alternate hypothesis
- 90% confidence interval = $118.2 \pm (1.65 * 1.08) = 116.41$ to 119.98
- 90% confidence interval for mean IQ of mothers of gifted childrens is between 116.4 to 120. This aligns with the alternate hypothesis that the mean is not 100

5. 4.34 CLT

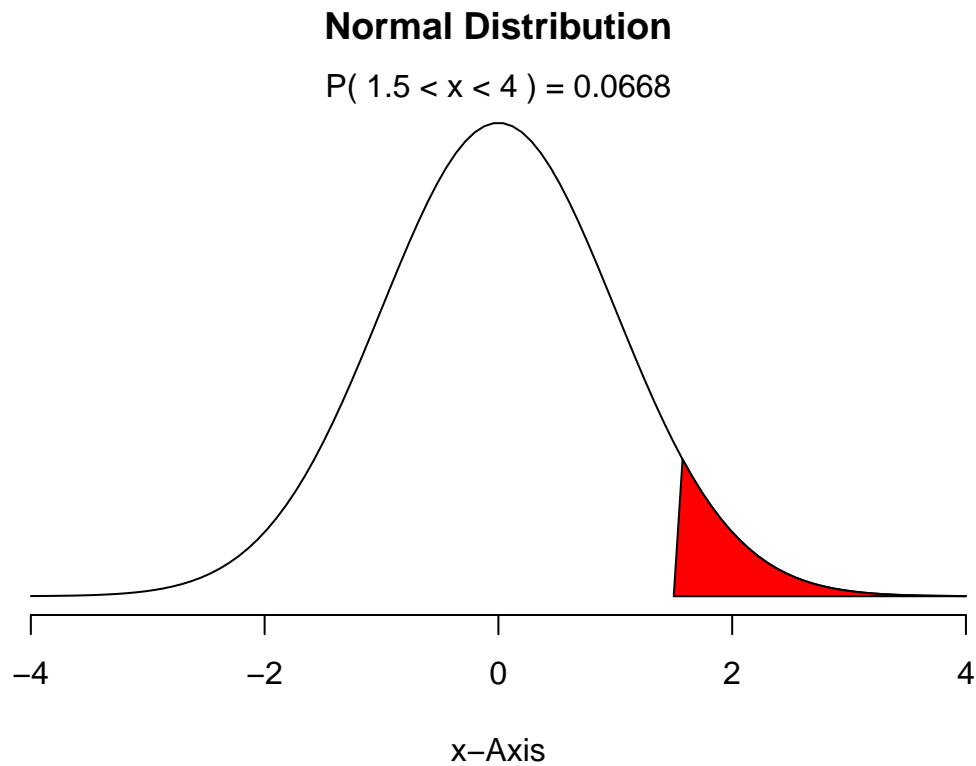
Sampling distribution of the sample means is the distribution by drawing multiple samples from population and plotting sample means. As per central limit theorem this distribution is always normal irrespective of the actual distribution of the population from which samples are drawn. It is recommended to have sample size ≥ 30 to get best results while performing hypothesis testing. As sample size increases the standard error goes down and we can better estimate population parameter.

6. 4.40 CFLBs

- Z value = $(10500-9000)/1000 = 1.5$

Probability of $x > 10500$ is $1-\text{pnorm}(1.5) = 0.0668$

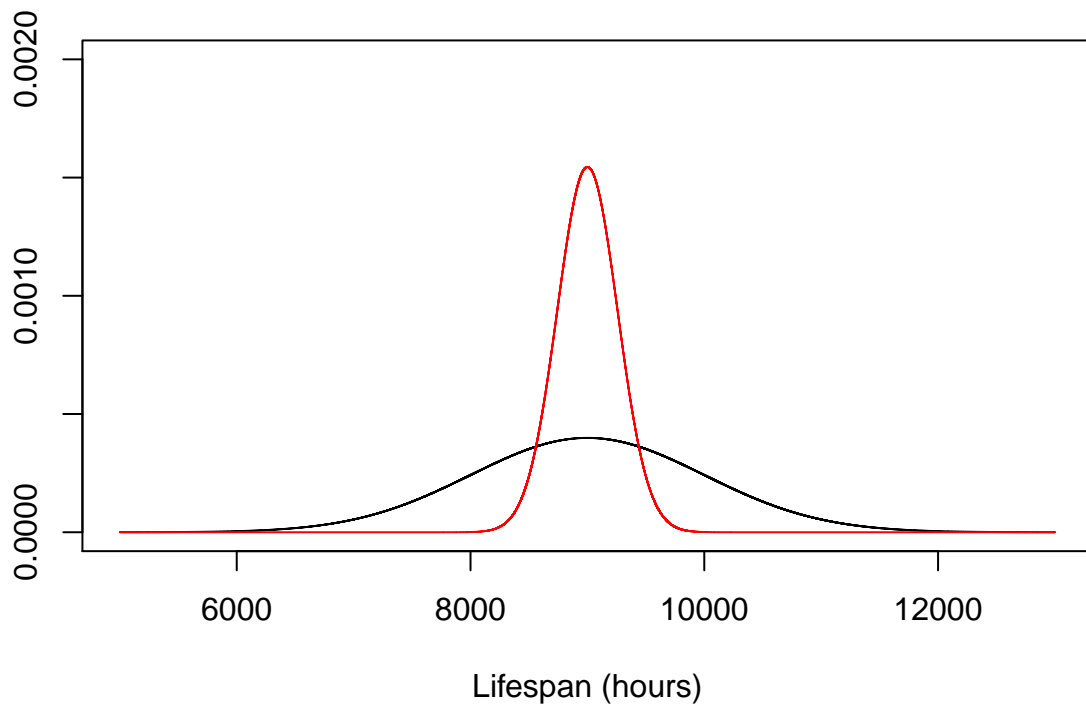
```
normalPlot(mean = 0, sd = 1, bounds=c(1.5,4), tails = FALSE)
```



- Distribution of the mean lifespan of 15 bulbs will be normal with mean of 9000 and SD of 258.19
- Z Value = $(10500-9000)/258.19 = 5.81$, probability is $1-\text{pnorm}(5.81) = 0$
- Population distribution vs Sampling distribution of sample mean

Black : Denotes population distribution Red: Denotes sampling distribution

```
s <- seq(5000,13000,0.01)
plot(s, dnorm(s,9000, 1000), type="l", ylim = c(0,0.002), ylab = "", xlab = "Lifespan (hours)")
lines(s, dnorm(s,9000, 258.1989), col="red")
```



- Part a can't be estimated without a nearly normal distribution. Part c can't be estimated since sample size is not sufficient

7. 4.48 Same observation with different sample size

As sample size increases, standard error decreases and Z value increases for positive Z and decreases for negative Z. This change in Z value causes P value to decrease