HW3

Dhananjay Kumar

September 16, 2016

```
library(IS606)
```

```
##
## Welcome to CUNY IS606 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='IS606') will list the demos that are available.

##
## Attaching package: 'IS606'

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

3.2 Area under the curve, Part II

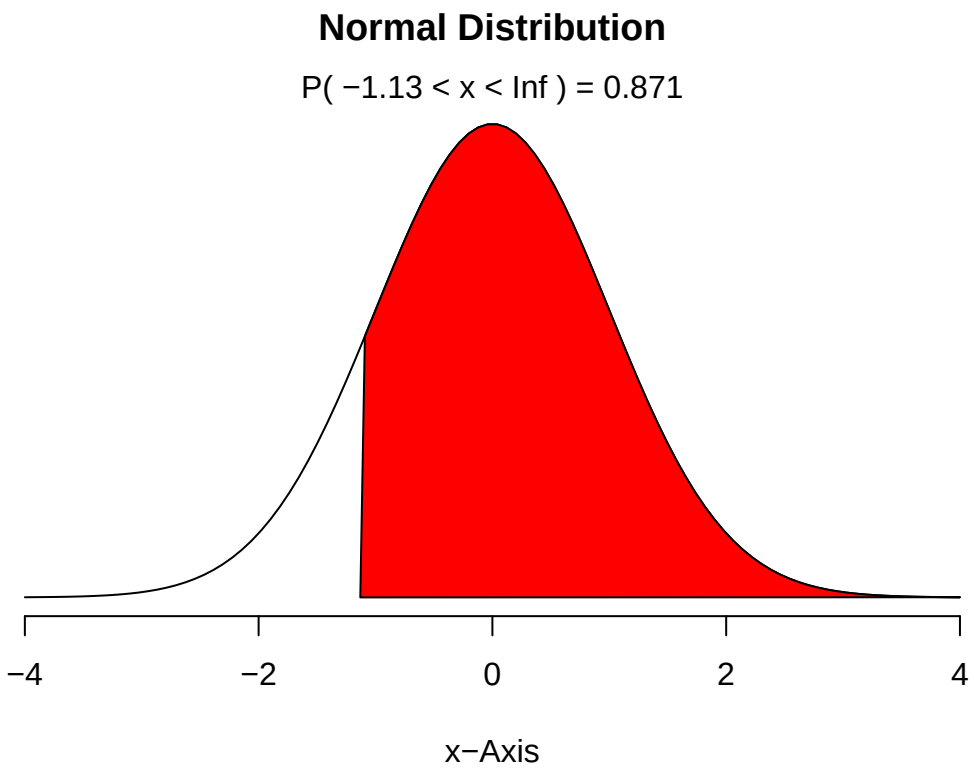
What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph. (a) Z > ????1:13 (b) Z < 0:18 (c) Z > 8 (d) jZj < 0:5

Solution: (a)

```
1-pnorm(-1.13)
```

```
## [1] 0.8707619
```

```
normalPlot(bounds = c(-1.13,Inf))
```

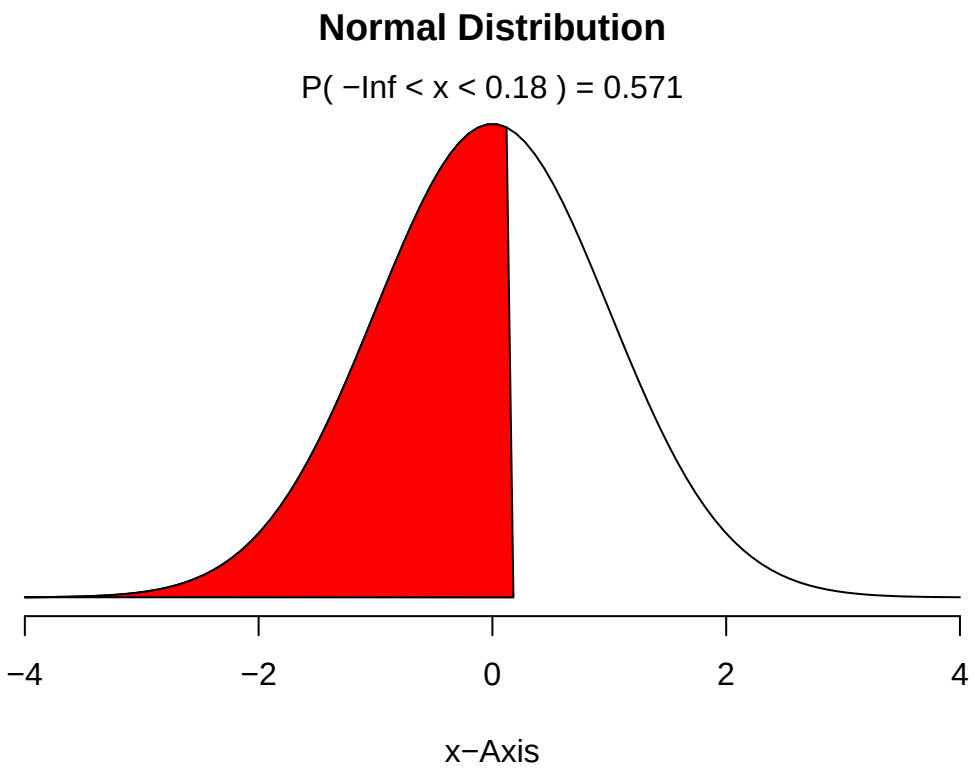


(b)

```
pnorm(0.18)
```

```
## [1] 0.5714237
```

```
normalPlot(bounds = c(-Inf, 0.18))
```

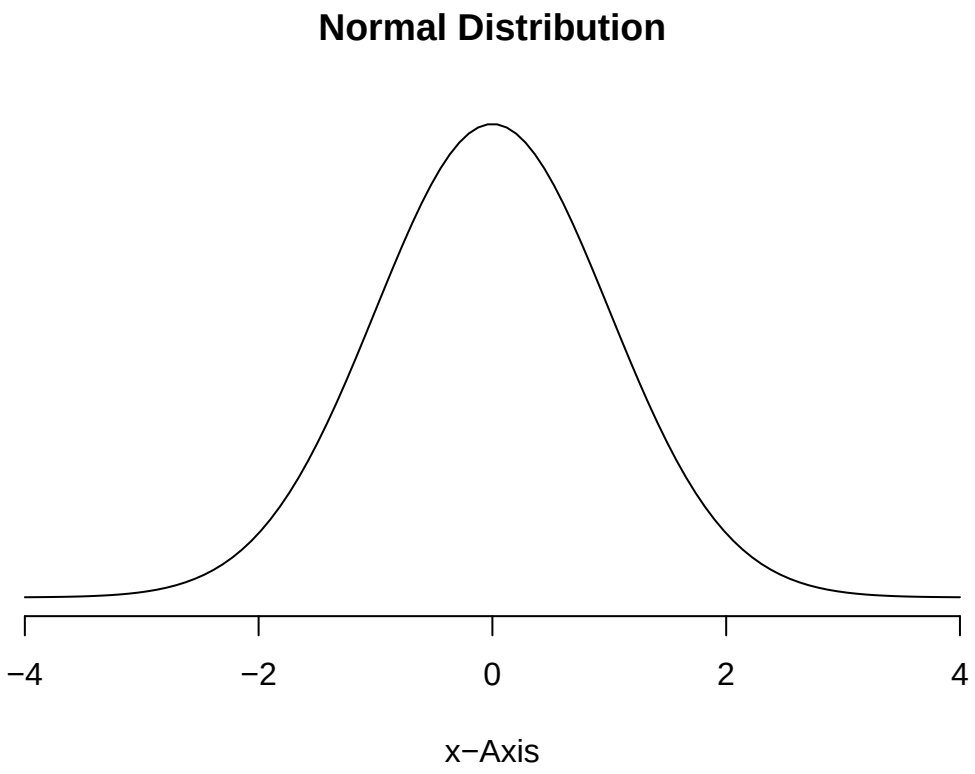


(c)

```
1- pnorm(8)
```

```
## [1] 6.661338e-16
```

```
normalPlot(bounds = c(-Inf, Inf), tails=TRUE)
```

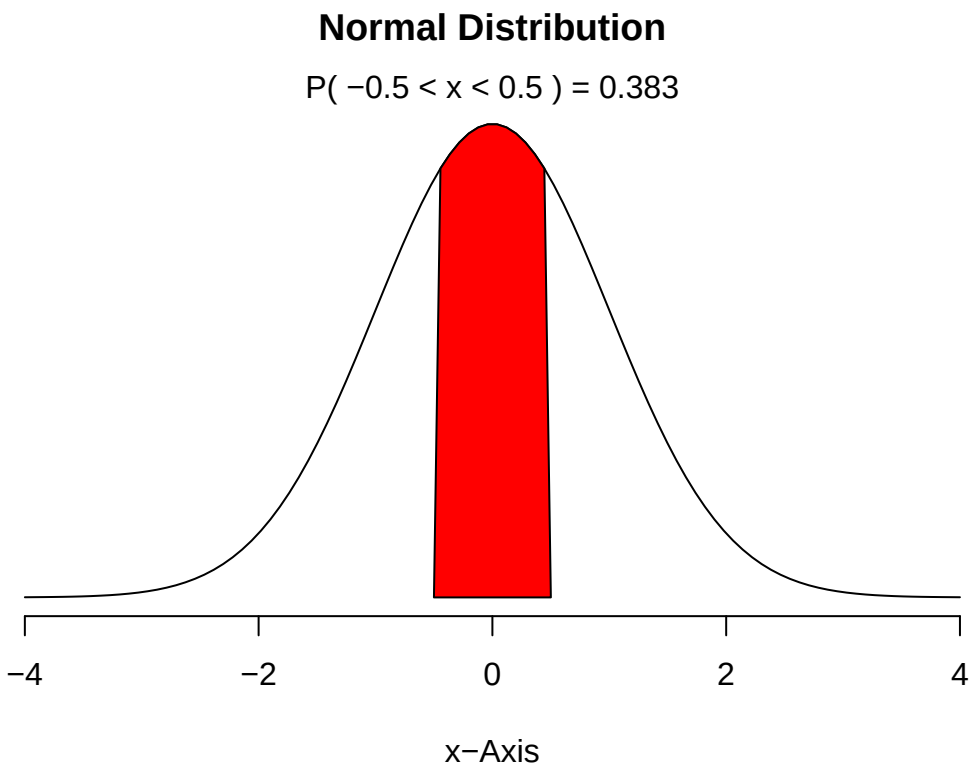


(d)

```
pnorm(0.5) - pnorm(-0.5)
```

```
## [1] 0.3829249
```

```
normalPlot(bounds=c(-0.5, 0.5))
```



3.4 Triathlon times, Part I.

In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the Men, Ages 30 - 34 group while Mary competed in the Women, Ages 25 - 29 group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

The finishing times of the Men, Ages 30 - 34 group has a mean of 4313 seconds with a standard deviation of 583 seconds. The finishing times of the Women, Ages 25 - 29 group has a mean of 5261 seconds with a standard deviation of 807 seconds. The distributions of finishing times for both groups are approximately Normal. Remember: a better performance corresponds to a faster finish.

(a) Write down the short-hand for these two normal distributions.

Solution: Distribution for Men : N(4313, 583) using N(mu/sigma) Distribution for Women : N(5261, 807)

(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

Solution: Z-Score = (X - Mean)/Sigma

```
# Z-Score = (X - Mean)/Sigma
leoZscore <- round((4948 - 4313)/583, 2)
leoZscore
```

```
## [1] 1.09
```

```r
maryZscore <- round((5513 - 5261)/807, 2)
maryZscore
```

```
## [1] 0.31
```

The Z-Score of Leo is significantly greater than Mary's which means that Leo is at a farther distance than Mean which means he ran slower when compared to his peers.

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

Solution In comparison to Leo, Mary did a lot better in her respective group. Leo's Zscore is 1.09 which tell us the how far he was from the Mean whereas Mary's Zscore 0.31 is a lot closer to the Mean.

(d) What percent of the triathletes did Leo finish faster than in his group?

Solution:

```r
# P_1 <- Probability of Participants faster than Leo
p_1 <- pnorm(round((4948 - 4313)/583, 2))
p_1
```

```
## [1] 0.8621434
```

```r
# p_2 <- Probability of Leo faster than other Participants
p_2 <- 1 - p_1
p_2
```

```
## [1] 0.1378566
```

```r
# In Percentage
p_2 <- p_2 * 100
p_2
```

```
## [1] 13.78566
```

(e) What percent of the triatheletes did Mary finish faster than in her group?

Solution:

```r
# P_1 <- Probability of Participants faster than Mary
p_1 <- pnorm(round((5513 - 5261)/807, 2))
p_1
```

```
## [1] 0.6217195
```

```r
# p_2 <- Probability of Mary faster than other Participants
p_2 <- 1 - p_1
p_2
```

```
## [1] 0.3782805
```

```
# In Percentage
p_2 <- p_2 * 100
p_2
```

```
## [1] 37.82805
```

(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

Solution: Yes, my response would change because if its not nornal distribution then our result would differ.

3.18 Heights of female college students.

Below are heights of 25 female college students. 54,55,56,56,57,58,58,59,60,60,60,61,61,62,62,63,63,63,64,65,65,67,67,69,73

(a) The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule

Solution:

```
heights <- c(54,55,56,56,57,58,58,59,60,60,60,61,61,62,62,63,63,63,64,65,65,67,67,69,73)
fHeights <- data.frame(heights)

meanHt <- mean(heights)
meanHt
```

```
## [1] 61.52
```

```
sdHt <- sd(heights)
sdHt
```

```
## [1] 4.583667
```

```
percentBySd <- function(data, numSd)
{
  m <- mean(data)
  s <- sd(data)

  sd1Lower <- m - (s * numSd)
  sd1Upper <- m + (s * numSd)

  sdData <- data[sd1Lower < data & data < sd1Upper]
  pSdData <- length(sdData) / length(data)
  return (pSdData)
}
# 1 Standard Deviation
sd1 <- percentBySd(heights, 1) * 100
# 2 Standard Deviation
sd2 <- percentBySd(heights, 2) * 100
# 3 Standard Deviation
sd3 <- percentBySd(heights, 3) * 100
sdList <- c(sd1, sd2, sd3)
sdList
```

```
## [1]  68  96 100
```

As seen above, the value(68,96,100) closely adheres to 68-95-99.7% Rule

(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.

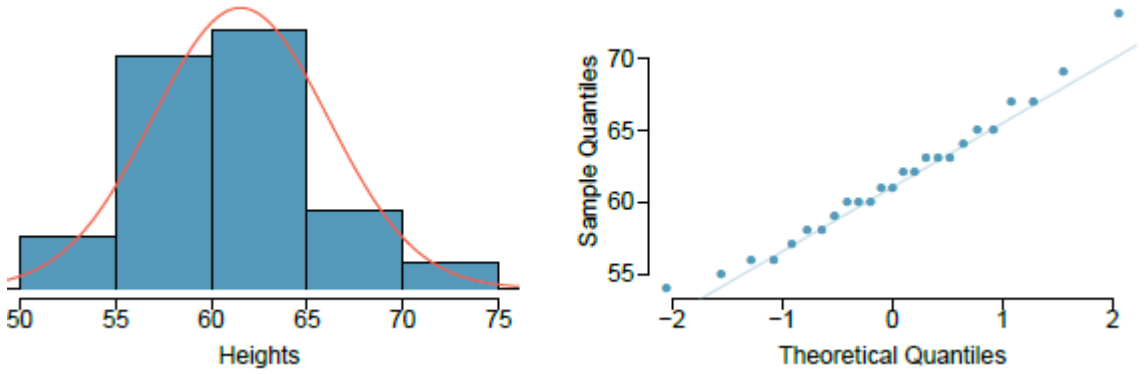


Figure 1: b

Solution As seen in the above Histogram, the data nesrly follows Normal Distribution. It is slighlty skewed towards Right. Similarly , the QQPlot also roughly follows Normal distribution.

3.22 Defective rate.

A machine that produces a special type of transistor (a component of computers) has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

(a) What is the probability that the 10th transistor produced is the first with a defect?

Solution:

```
# As mentioned under Geometric Distribution(3.30) Pg:143
defect <- 0.02
sRate <-  1 - defect
n <- 10
# pDef <- probability that the 10th transistor produced is the first with a defect
pDef <- sRate^(n-1) * defect
pDef
```

```
## [1] 0.01667496
```

(b) What is the probability that the machine produces no defective transistors in a batch of 100?

Solution:

```
pDef <- 0
for(i in 1:100)
{
  pDef <- pDef + (sRate^(i-1) * 0.02)
}
pNoDef <- 1 - pDef
pNoDef
```

```
## [1] 0.1326196
```

(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?

Solution:

```
eVal <- 1 / 0.02
eVal
```

```
## [1] 50
```

```
sDevi <- sqrt( (1 - 0.02) / defect^2 )
sDevi
```

```
## [1] 49.49747
```

(d) Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What is the standard deviation?

Solution:

```
newSRate <- 1 - 0.05
newSRate
```

```
## [1] 0.95
```

```
eVal2 <- 1/0.05
eVal2
```

```
## [1] 20
```

```
sDevi2 <- sqrt( (1 - 0.05) / 0.05^2 )
sDevi2
```

```
## [1] 19.49359
```

(e) Based on your answers to parts (c) and (d), how does increasing the probability of an event affect the mean and standard deviation of the wait time until success?

Solution: Increasing the proability of an event is inversely proportional to the mean and standard deviation i.e. increasing the probability of an decreases the mean and standard deviation of the wait time.

3.38 Male children.

While it is often assumed that the probabilities of having a boy or a girl are the same, the actual probability of having a boy is slightly higher at 0.51. Suppose a couple plans to have 3 kids.

(a) Use the binomial model to calculate the probability that two of them will be boys.

Solution:

```
dbinom(2, 3, prob=0.51)
```

```
## [1] 0.382347
```

```
bdist <- function(n, k, p) {
    choose(n, k) * p^k * (1-p) ^ (n-k)
}
bdist(3, 2, 0.51)
```

```
## [1] 0.382347
```

(b) Write out all possible orderings of 3 children, 2 of whom are boys. Use these scenarios to calculate the same probability from part (a) but using the addition rule for disjoint outcomes. Confirm that your answers from parts (a) and (b) match.

Solution:

```
#1 Boy Girl Boy
#2 Boy Boy  Girl
#3 Girl Boy Boy

#P for Case 1 + Case 2 + Case 3
suM <-( 0.51 * 0.49 * 0.51 ) + ( 0.51 * 0.51 *0.49 ) + ( 0.49 * 0.51 * 0.51 )
suM
```

```
## [1] 0.382347
```

As seen above the answer matches with (b)

(c) If we wanted to calculate the probability that a couple who plans to have 8 kids will have 3 boys, briefly describe why the approach from part (b) would be more tedious than the approach from part (a).

Solution: Since the number of combinationin this case would be significantly larger than that of (b), it would very tedious to do it in the approach followed in (b). the best and convenient way of soint it would be to follow the Binomial distribution model where the formula takes care of all the possible combinations(as seen in (a)).

3.42 Serving in volleyball.

A not-so-skilled volleyball player has a 15% chance of making the serve, which involves hitting the ball so it passes over the net on a trajectory such that it will land in the opposing team's court. Suppose that her serves are independent of each other.

(a) What is the probability that on the 10th try she will make her 3rd successful serve?

Solution:

```
aNs <- choose(10-1, 3-1) * 0.15^3 * ((1-0.15)^(10-3))
aNs
```

```
## [1] 0.03895012
```

(b) Suppose she has made two successful serves in nine attempts. What is the probability that her 10th serve will be successful?

Solution: 0.15, since evey trial is independent of each other so the probability remains the same

(c) Even though parts (a) and (b) discuss the same scenario, the probabilities you calculated should be different. Can you explain the reason for this discrepancy?

Solution: In Part (a) we calculated the success in 10th trial but it also considers the fact that the 10th trial is her third successfull event, so we had to take care of all the possible combinations whereas in part (b) we were calculating the probability of a single successfull event at 10th trial.