Illapani_Homework_HW3

Srini Illapani

September 16, 2015

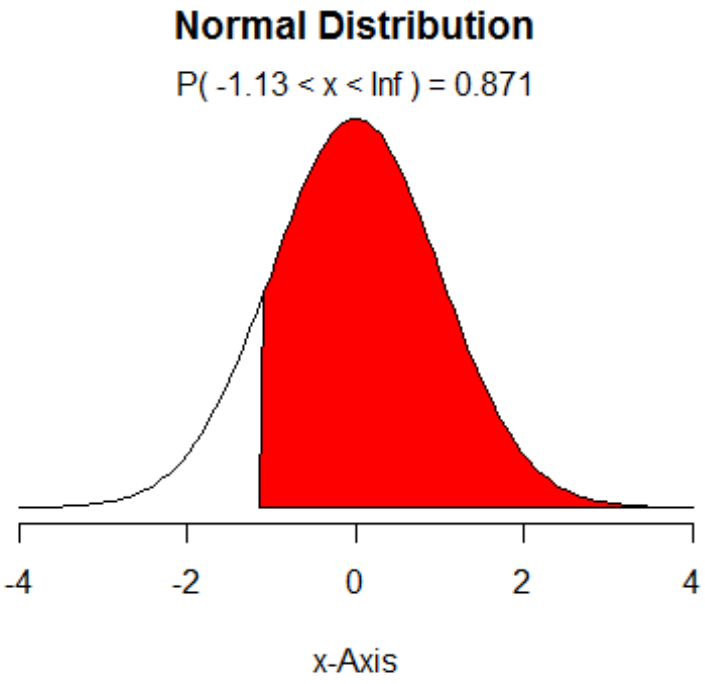
```
library(IS606)

##
## 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.

library(Rcpp)
library(scales)
```

3.2 Area under the curve, Part II

What percent of a standard normal distribution N(?? = 0,! = 1) is found in each region? Be sure to draw a graph.

(a) Z > ???1.13 (b) Z < 0.18 (c) Z > 8 (d) |Z| < 0.5

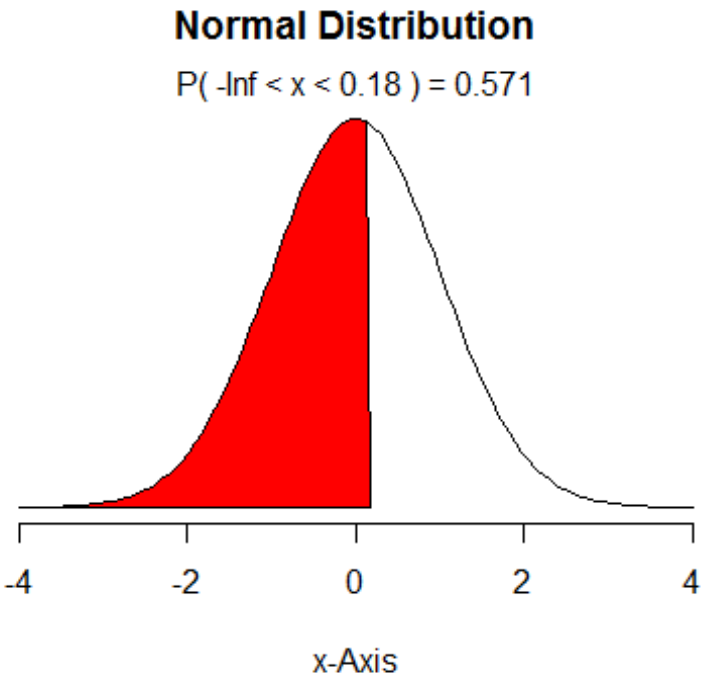
```
#a. Z > ???1.13
IS606::normalPlot(0, 1, c(-1.13, Inf))
```




```
1 - pnorm(-1.13, mean = 0, sd = 1)
```

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

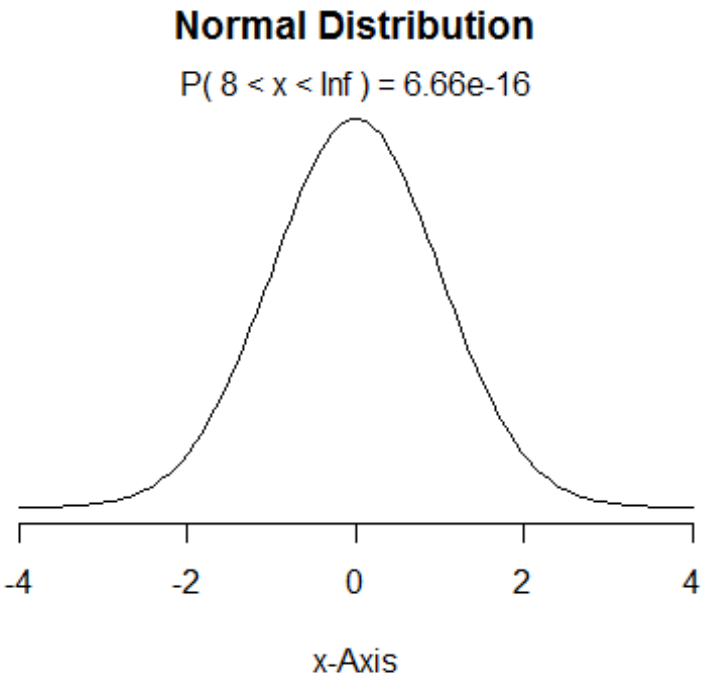
```
#b. Z < 0.18
IS606::normalPlot(0, 1, c(-Inf, .18))
```




```
pnorm(.18, mean = 0, sd = 1)
```

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

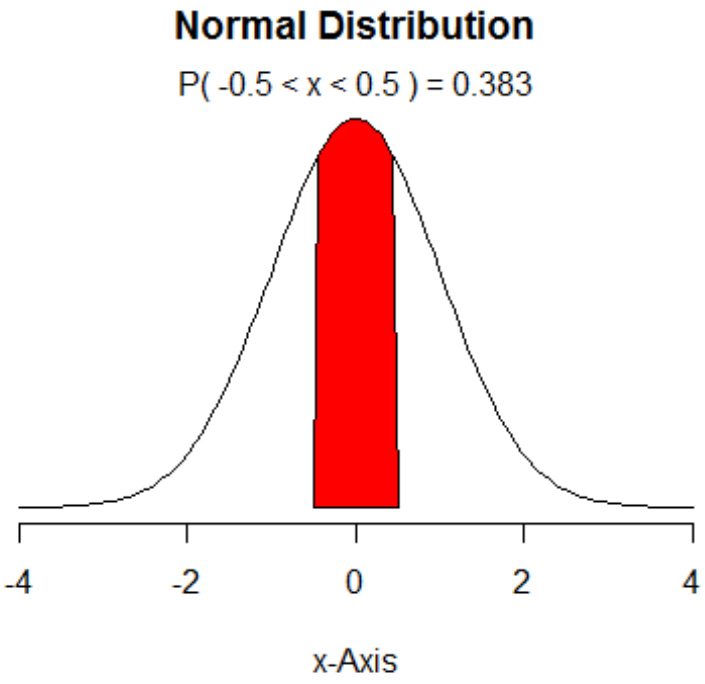
```
#c. Z > 8
IS606::normalPlot(0, 1, c(8, Inf))
```




```
1 - pnorm(8, mean = 0, sd = 1)
```

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

```
#d.|Z| < 0.5
IS606::normalPlot(0, 1, c(-.5, .5))
```




```
pnorm(.5, mean = 0, sd = 1) - pnorm(-.5, mean = 0, sd = 1)
```

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

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.

- Distribution_Men = N(= 4313, = 583)
- Distribution_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?

```
# Leo
Leo_finishtime <- 4948
men_mean_finishtime <- 4313
men_sd_finishtime <- 583
z_Leo <- (Leo_finishtime - men_mean_finishtime) / (men_sd_finishtime)
z_Leo <- round(z_Leo,2)
z_Leo
```

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

```
# Mary
Mary_finishtime <- 5513
women_mean_finishtime <- 5261
women_sd_finishtime <- 807
z_Mary <- (Mary_finishtime - women_mean_finishtime) / (women_sd_finishtime)
z_Mary <- round(z_Mary,2)
z_Mary
```

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

- Leo's z score is higher compared to Mary's. Meaning he is further away from the mean timing compared to Mary. The lesser the finish time the better, so Mary's performance is better compared to Leo's.

---

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

- Mary would rank higher in her respective group than Leo would, the reason being that Mary is closer to the mean for her respective group than Leo is. He has a better time in his group but when we accout for the distribution within each of their groups, Mary does better than Leo.

---

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

```
library(scales)
Leo_faster <- (1 - pnorm(Leo_finishtime, mean = men_mean_finishtime, sd =
men_sd_finishtime))
Leo_faster
```

```
## [1] 0.1380342
```

- Leo was faster than 13.8% of triathletes in his group.

e. What percent of the triathletes did Mary finish faster than in her group?

```
Mary_faster <- (1 - pnorm(Mary_finishtime, mean = women_mean_finishtime, sd = women_sd_finishtime))
Mary_faster
```

```
## [1] 0.3774186
```

- Mary was faster than 37.75% of triathletes in her group.

---

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

- When we look at part b the z scores would not change so b would be same. But for c - e, Z assumes normality so when there is a non-normal distribution, the accuracy is no more there.

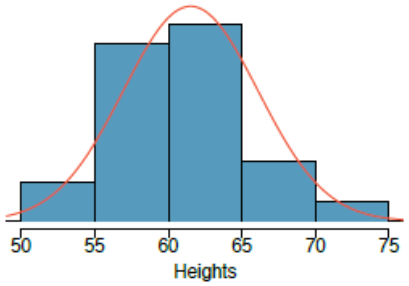
3.18 Heights of female college students

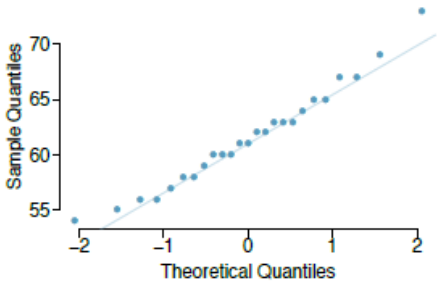
3.18 Heights of female college students. Below are heights of 25 female college students.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
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.

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





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.

```
height <-
c(54,55,56,56,57,58,58,59,60,60,60,61,61,62,62,63,53,53,64,65,65,67,67,69,76)

one_sd <- subset(height, height < (mean(height) + sd(height)) & height > (mean(height) - sd(height)))
one_sd <- percent(length(one_sd)/length(height))
one_sd
```

```
## [1] "68%"

two_sd  <- subset(height, height < (mean(height) + (2* sd(height))) & height
> (mean(height) - (2 * sd(height))))
two_sd <- percent(length(two_sd)/length(height))
two_sd

## [1] "96%"

three_sd  <- subset(height, height < (mean(height) + (3 * sd(height))) &
height > (mean(height) - (3 * sd(height))))
three_sd <- percent(length(three_sd)/length(height))
three_sd

## [1] "100%"
```

- The values are very close to the 68-95-99.7% rule, so the heights of female college studdents does follow the rule.

---

b. Does the data appear to follow a normal distribution? Explain your reasoning using the graphs provided above.

- The distribution is normal,it curve is symmetrical.
  The points along the normal probability plot are closely aligned to the line and looks like a 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?

```
defect_rate <- 0.02
prob_10th_defect <- ((1 - defect_rate)^(10 - 1)) * defect_rate
```

- The probability that the 10th transistor produced is the first with a defect is 1.67%

---

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

```
prob_100_no_defects <- ((1 - defect_rate)^(100 - 1)) * (1 - defect_rate)
```

- The probability that the machine produces no defective transistors in a batch of 100 is 13.26%

---

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

```
first_defect <- (1/defect_rate)
sd <- sqrt((1 - defect_rate)/(defect_rate ^ 2))
```

- On average 50 transistors would be produced before the first defect is found.
- The Standard Devistion is 49.5

---

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?

```
defect1_rate <- 0.05
first_defect1 <- 1/defect1_rate
sd1 <- sqrt((1 - defect1_rate)/(defect1_rate ^ 2))
```

- One would expect on average 20 transistors would be produced before the first defect would appear.
  The standard deviation now is 19.5

---

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?

- By increasing the probability of a defect occuring, the defects will appear sooner and more frequently. The standard deviation is smaller as a result of the defects occuring more often.

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.

```
n <- 3
k <- 2
p <- 0.51

two_boys_a <- choose(3,2) * ((p^k)*((1-p)^(n-k)))
two_boys_a

## [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.

```
B <- 0.51
G <- 0.49

a <- c(B,B,G)
b <- c(B,G,B)
c <- c(G,B,B)

X <- rbind(a)
X <- rbind(X,b)
X <- rbind(X,c)
X <- matrix(X, nrow = 3)
X <- cbind(X, seq(1:3))

X[1,4] <- X[1,1] * X[1,2] * X[1,3]
X[2,4] <- X[2,1] * X[2,2] * X[2,3]
X[3,4] <- X[3,1] * X[3,2] * X[3,3]
two_boys_b <- sum(X[,4])
two_boys_b
```

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

- The probabbility using the above binomial model and the adddition rules methods is the same.

---

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).

- The approach in b includes creating and addding the matrix manually, and it takes more time compared to the approach a where we used the formula and can be easily extended do a large number of subjects or sample and it is also less error prone.

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?

```
n <- 10
k <- 3
p <- 0.15
```

```
prob_10th_3rd <- choose(9,2) * ((p^k)*((1-p)^(n-k)))
prob_10th_3rd

## [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?

- It will be 15% as her serves are independent of each other.

---

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?

- In question a, the ask was to get a success on her 10th serve and it would be her 3rd successful serve. In question b, the ask is just the probability of success in her 10th attempt. Hence the prob is lower for question a.