Chapter 4 - Distributions of Random Variables

Subhalaxmi Rout

Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z < -1.35$
(b) $Z > 1.48$
(c) $-0.4 < Z < 1.5$
(d) $|Z| > 2$

** **Answer** **

(a) $Z < -1.35$

```
library("DATA606")
```

```
## Loading required package: shiny

## Loading required package: openintro

## Please visit openintro.org for free statistics materials

##
## Attaching package: 'openintro'

## The following objects are masked from 'package:datasets':
##
##     cars, trees

## Loading required package: OIdata

## Loading required package: RCurl

## Loading required package: maps

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:openintro':
##
##     diamonds
## 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.
```

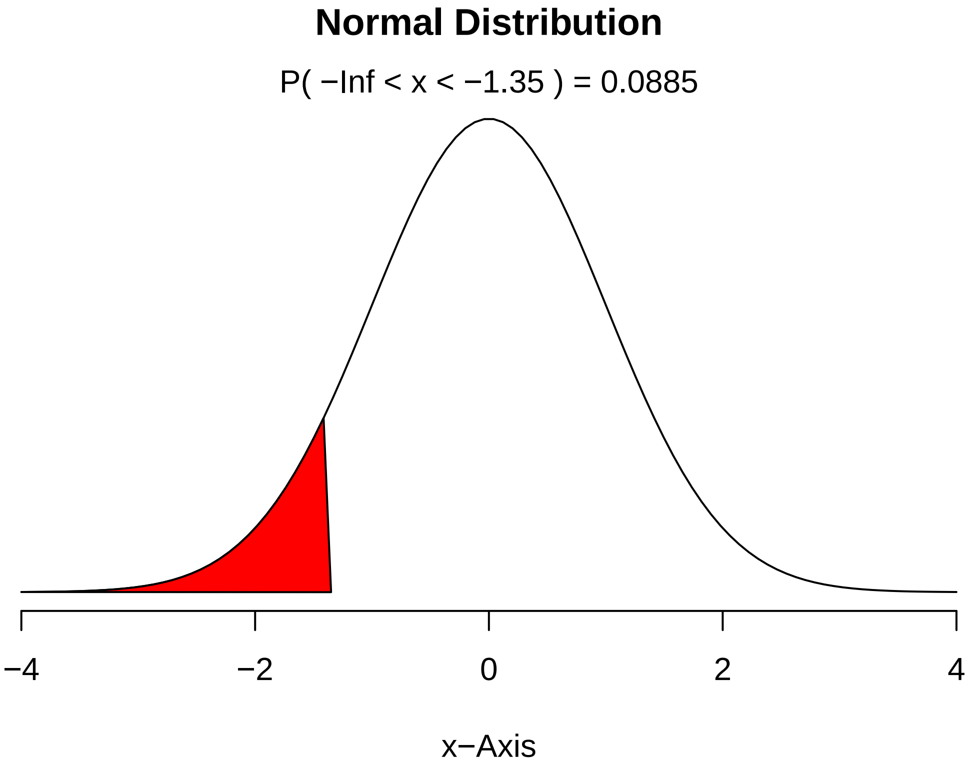
```
##
## Attaching package: 'DATA606'
```

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

```
mean = 0
sd = 1
area <- pnorm(-1.35, mean=0, sd=1)
area
```

```
## [1] 0.08850799
```

```
normalPlot(mean = 0, sd = 1,bounds = c(-Inf,-1.35), tails = F)
```



```r
paste("Z < -1.35 area under curve is 8.85%")
```

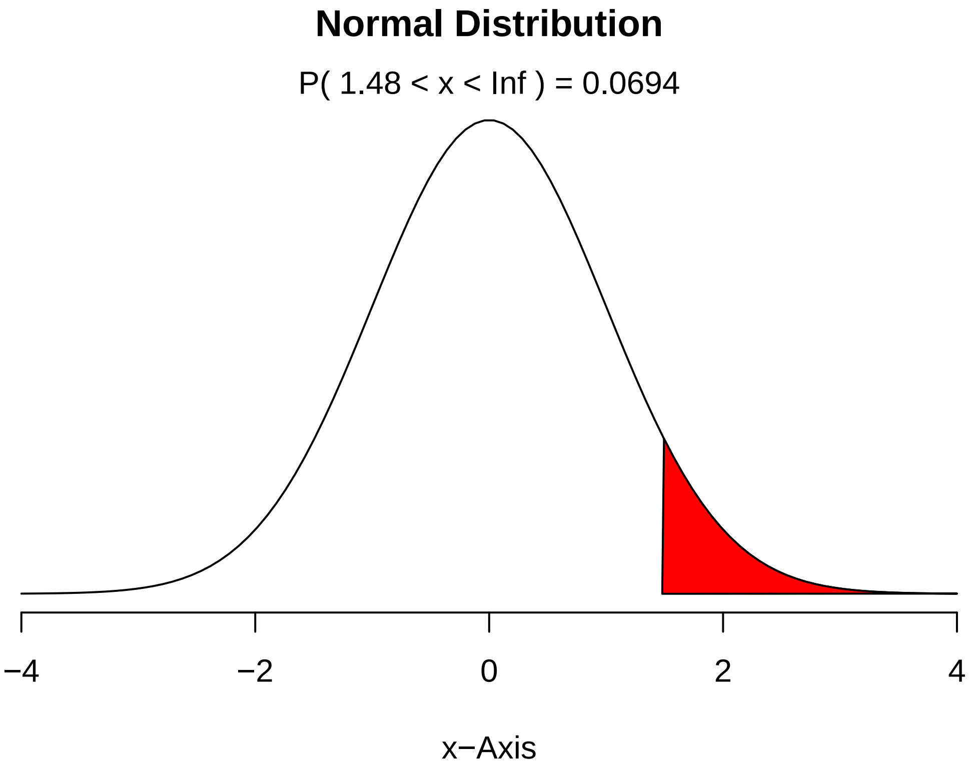
```
## [1] "Z < -1.35 area under curve is 8.85%"
```

(b) $Z > 1.48$

```r
area <- 1 - pnorm(1.48, mean=0, sd=1)
area
```

```
## [1] 0.06943662
```

```r
normalPlot(mean = 0, sd = 1,bounds = c(1.48, Inf), tails = F)
```



```r
paste("Z > 1.48 area under the curve is 6.94%")
```

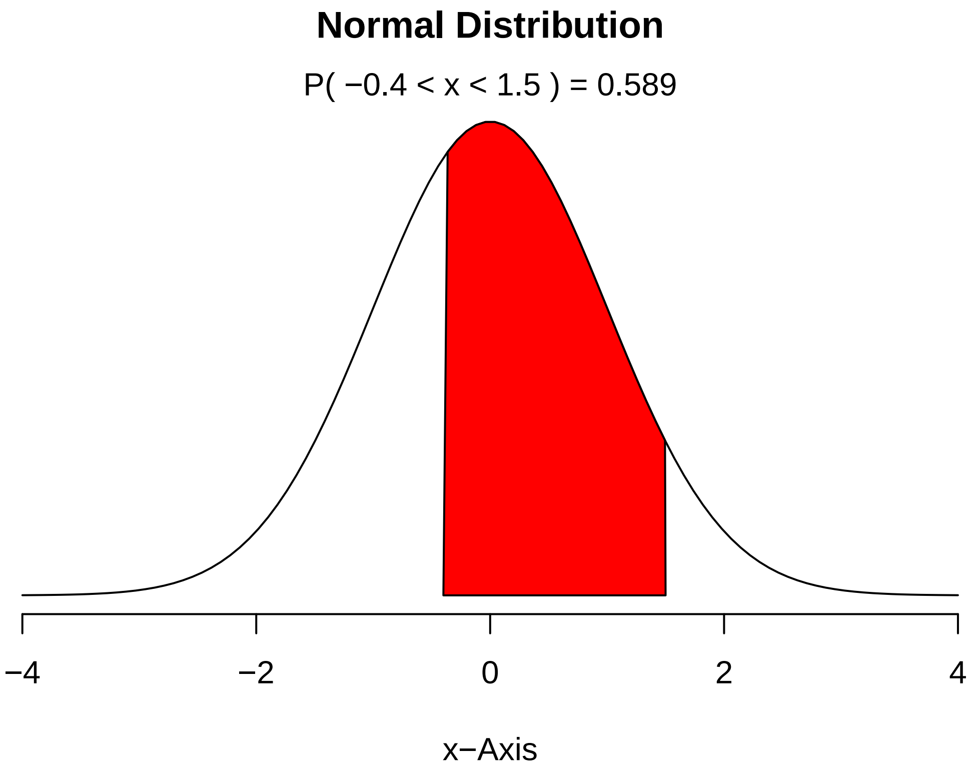
```
## [1] "Z > 1.48 area under the curve is 6.94%"
```

(c) $-0.4 < Z < 1.5$

```r
area1 <- pnorm(-0.4, mean=0, sd=1)
area2 <- 1 - pnorm(1.5, mean=0, sd=1)
area <- 1-(area1 + area2)
area
```

```
## [1] 0.5886145
```

```r
normalPlot(mean = 0, sd = 1,bounds = c(-0.4, 1.5), tails = F)
```



```r
paste("-0.4 < Z < 1.5 area under the curve is 58.86%")
```

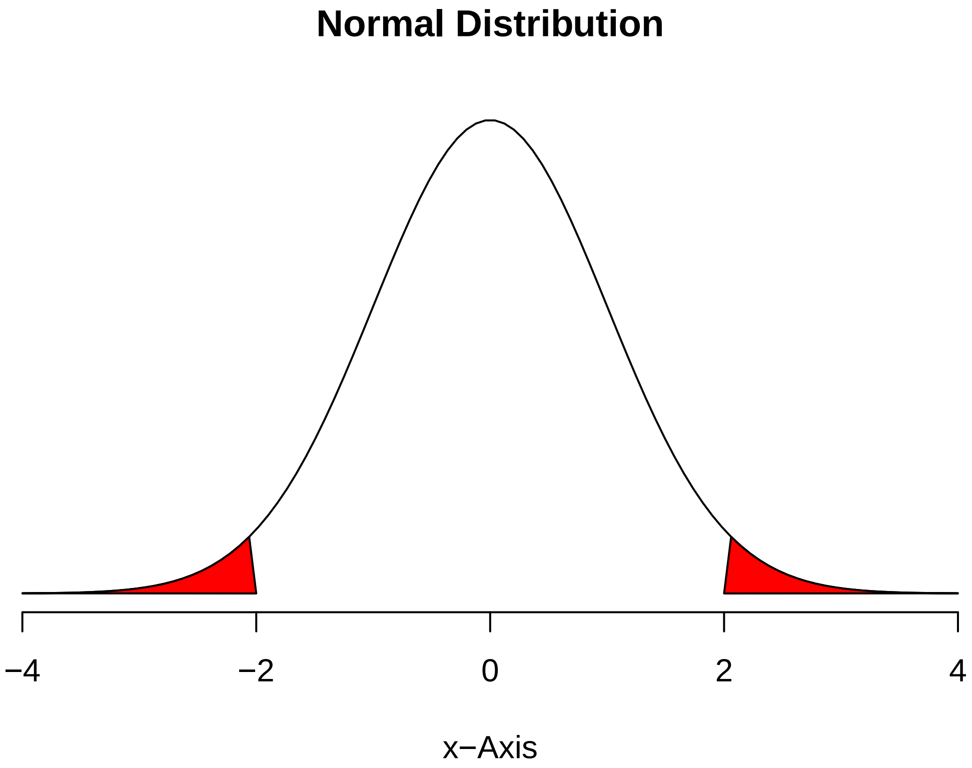
```
## [1] "-0.4 < Z < 1.5 area under the curve is 58.86%"
```

(d) $|Z| > 2$

```r
area1 <- pnorm(-2, mean=0, sd=1)
area2 <- 1 - pnorm(2, mean=0, sd=1)
area <- area1 + area2
area
```

```
## [1] 0.04550026
```

```
normalPlot(mean = 0, sd = 1,bounds = c(-2, 2), tails = T)
```



```
paste("|Z| > 2 area under the curve is 4.55%")
```

```
## [1] "|Z| > 2 area under the curve is 4.55%"
```

Triathlon times, Part I (4.4, p. 142) 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.
(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?
(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.
(d) What percent of the triathletes did Leo finish faster than in his group?
(e) What percent of the triathletes did Mary finish faster than 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.

** Answer **

```
(a) Men = N(mu = 4313, sigma = 583)
    Women = N(mu = 5261, sigma = 807)
```

(b)

```r
leo_X <- 4948
leo_Mu <- 4314
leo_sd <- 583
leo_z <- (leo_X - leo_Mu) / leo_sd
paste0("Leo's z score is: ",leo_z)
```

```
## [1] "Leo's z score is: 1.08747855917667"
```

```r
paste("Leo's finish time is 1.087 sd above the mean.")
```

```
## [1] "Leo's finish time is 1.087 sd above the mean."
```

```r
mary_X <- 5513
mary_Mu <- 5261
mary_sd <- 807
mary_z <- (mary_X - mary_Mu) / mary_sd
paste0("Mary's z score is: ",mary_z)
```

```
## [1] "Mary's z score is: 0.312267657992565"
```

```r
paste("Mary's finish time is 0.312 sd above the mean.")
```

```
## [1] "Mary's finish time is 0.312 sd above the mean."
```

(c) Mary did better in her group because Mary's finish average time is faster than women group. But, finish average time of Leo, slower than men group.

(d)

```r
1 - pnorm(leo_z)
```

```
## [1] 0.1384127
```

```r
paste("13.84 percent of the triathletes finish slower than Leo in his group.")
```

```
## [1] "13.84 percent of the triathletes finish slower than Leo in his group."
```

(e)

```r
1 - pnorm(mary_z)
```

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

```r
paste("37.74 percent of the triathletes finish slower than Mary in her group.")
```

```
## [1] "37.74 percent of the triathletes finish slower than Mary in her group."
```

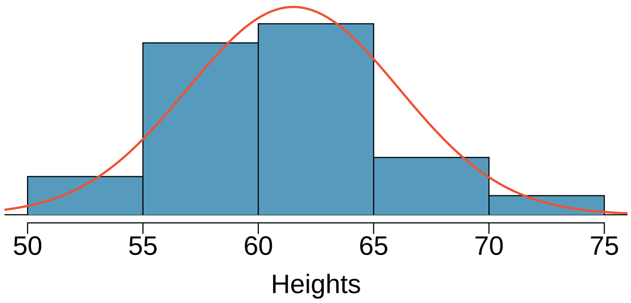
(f) If the distribution of finishing time is not normal, then we cannot use pnorm to calculate (b),(d), and (e). I think (c) wo not change because this is calculate based on average finish time.

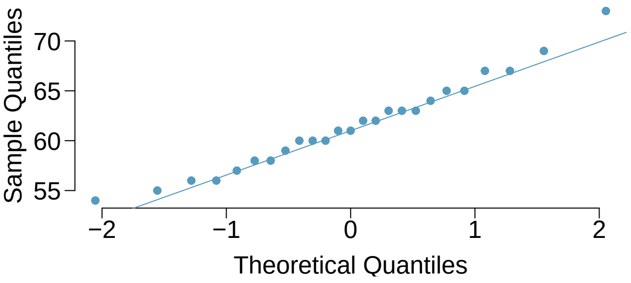
Heights of female college students Below are heights of 25 female college students.

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.





** **Answer** ** **(a)**

```
# Use the DATA606::qqnormsim function
height <- 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)
#summary(height)

mean <- mean(height)
sd <- sd(height)

#68-95-99.7% Rule

sd_1 <- height[which(height < mean + sd & height > mean - sd)]
sd_1 <- length(sd_1)/length(height)
sd_1
```

```
## [1] 0.68
```

```
paste0(sd_1*100," % of the data fall within 1 standard deviation.")
```

```
## [1] "68 % of the data fall within 1 standard deviation."
```

```
sd_2 <- height[which(height < mean + 2*sd & height > mean - 2*sd)]
sd_2 <- length(sd_2)/length(height)
sd_2
```

```
## [1] 0.96
```

```
paste0(sd_2*100," % of the data fall within 2 standard deviation.")
```

```
## [1] "96 % of the data fall within 2 standard deviation."
```

```
sd_3 <- height[which(height < mean + 3*sd & height > mean - 3*sd)]
sd_3 <- length(sd_3)/length(height)
sd_3
```

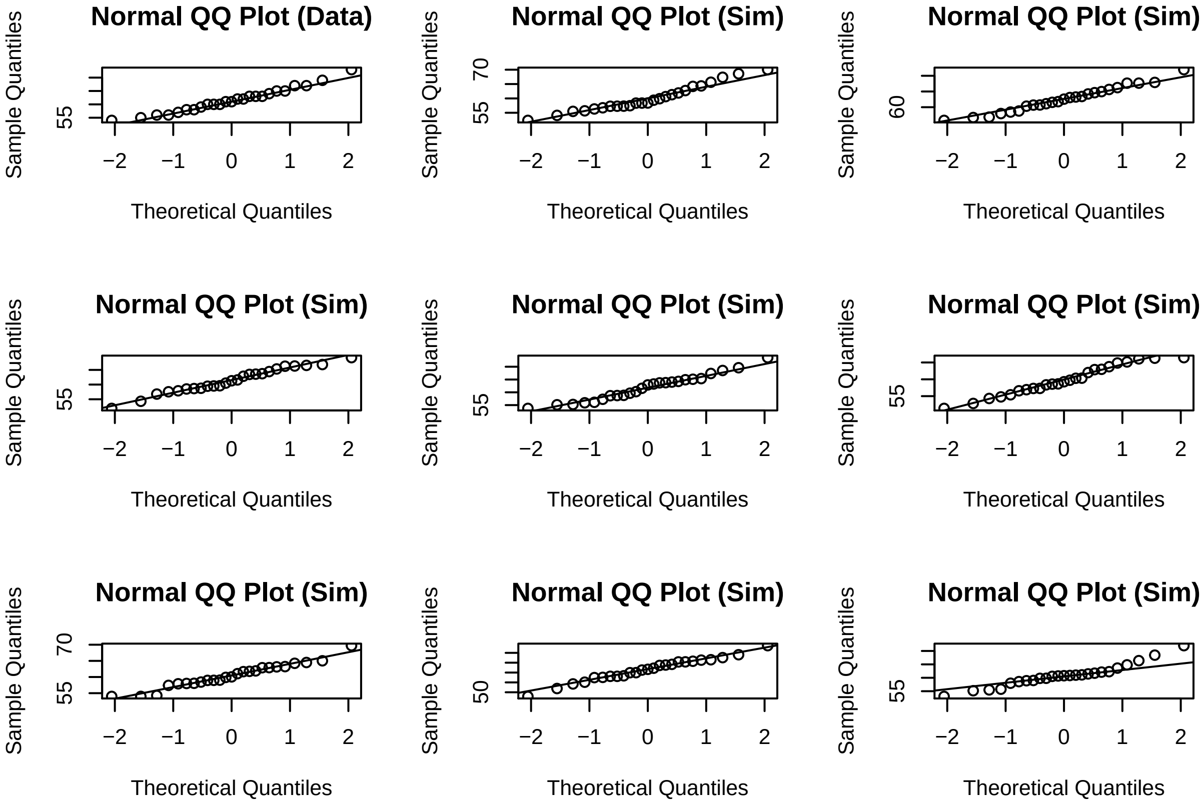
```
## [1] 1
```

```
paste0(sd_3*100," % of the data fall within 3 standard deviation.")
```

```
## [1] "100 % of the data fall within 3 standard deviation."
```

(b)

```
qqnormsim(height)
```



```
paste("The QQPlot shows normal distribution because most data are on the central line")
```

```
## [1] "The QQPlot shows normal distribution because most data are on the central line"
```

Defective rate. (4.14, p. 148) 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?
(b) What is the probability that the machine produces no defective transistors in a batch of 100?
(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?
(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?
(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?

** **Answer** **

(a)

```
defect <- 0.02
success <- (1-defect)
trials <- 10
prob <- ((success) ^ (10-1)) * defect
paste0("The probability that the 10th transistor produced is the first with a defect is ", prob)
```

```
## [1] "The probability that the 10th transistor produced is the first with a defect is 0.01667495524260
```

(b)

```
trails <- 100
prob <- ((success) ^ (100))
paste0("The probability that the machine produces no defective transistors in a batch of 100 is ",prob)
```

```
## [1] "The probability that the machine produces no defective transistors in a batch of 100 is 0.132619
```

(c)

```
avg <- 1/defect
paste0("Average is ", avg)
```

```
## [1] "Average is 50"
```

```
mean <- avg
sd <- sqrt((success)/((defect)^2))
paste0("standard deviation is ",sd)
```

```
## [1] "standard deviation is 49.4974746830583"
```

(d)

```r
defect <- 0.05
success <- 1 - 0.05
prob <- 1/defect
paste0(prob," transistors would expect to be produced with this machine before the first with a defect.
```

```
## [1] "20 transistors would expect to be produced with this machine before the first with a defect."
```

```r
sd <- sqrt((success)/((defect)^2))
paste0("standard deviation is ",sd)
```

```
## [1] "standard deviation is 19.4935886896179"
```

(e)

```
If the defect rate of probability will increase then standard deviation and mean will decrease.
```

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.
(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.
(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).

** Answer **

(a)

```
p <- 0.51
failure <- 1 - 0.51
n <- 3
k <- 2

success_boy <- choose(3,2) * (p)^2 * failure
paste0("The probability that two of them will be boys in 3 children is ",success_boy*100,"%")
```

```
## [1] "The probability that two of them will be boys in 3 children is 38.2347%"
```

(b)

```
#3 children ordering
# Boy Boy Girl
# Boy Girl Boy
# Girl Boy Boy
boy <- 0.51
girl <- 1 - boy

p1 <- boy * boy * girl
p2 <- boy * girl * boy
p3 <- girl * boy * boy

total <- p1 + p2 +p3
paste0("So probability of boys is ", total * 100, "%")
```

```
## [1] "So probability of boys is 38.2347%"
```

(c)

```
Calculate probability of a couple who plan to have 8 kids will have 3 boys,
will be tedious if we follow (b), because we have to write  56 condition.
So if we do follow (a) then we can apply direct formula.
```

```r
choose(8,3)
```

```
## [1] 56
```

Serving in volleyball. (4.30, p. 162) 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?
(b) Suppose she has made two successful serves in nine attempts. What is the probability that her 10th serve will be successful?
(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?

**** Answer ****

(a)

```
serve_success <- 0.15
serve_failure <- 1 - 0.15
serve_prob <- choose(9,2) * ((serve_success) ^ 3) * ((serve_failure) ^ 7)
paste0("The probability of 10th try she will make her 3rd successful serve ",serve_prob)
```

```
## [1] "The probability of 10th try she will make her 3rd successful serve 0.0389501162261719"
```

(b)

```
Due to all serves are independent, the probability of her 10th serve
will be successful is 0.15.
```

(c)

```
In (a) we calculate marginal probability of success but in (b) we calculate
conditional probability based on previous scinario.
```