Chapter 3 - Distributions of Random Variables

CUNY MSDA - DATA606 - Homework 3

Completed by: Duubar Villalobos Jimenez mydvtech@gmail.co

2017-02-25

Graded: 3.2 (see normalPlot), 3.4, 3.18 (use qqnormsim from lab 3), 3.22, 3.38, 3.42

3.2 Area under the curve

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

$$Z = \frac{x - \mu}{\sigma}$$

```
mu <- 0
sd <- 1
Z <- -1.13
# finding value for 'x'
x <- Z * sd + mu
x
```

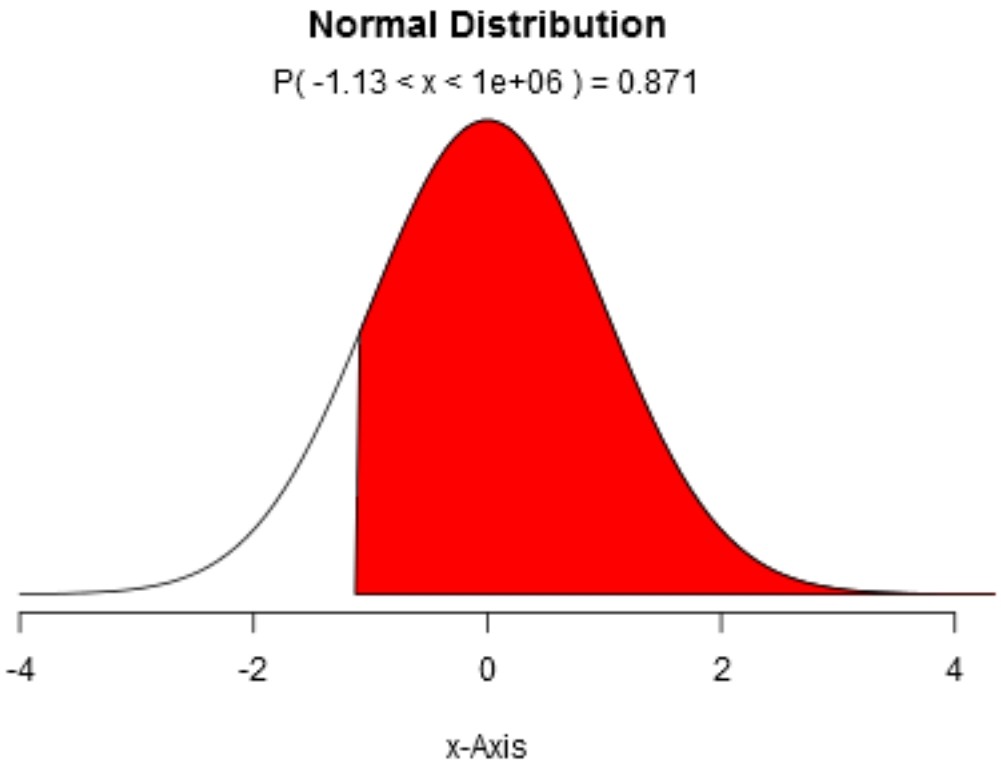
```
## [1] -1.13
```

Since we have that $x > -1.13$

```
# Finding probability fo x > -1.13
1 - pnorm(x, mean = 0, sd = 1)
```

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

```
# Probability curve plot
normalPlot(mean = 0, sd = 1, bounds = c(x, 1e+06), tails = FALSE)
```



Answer: The percentage represented on the region is: 87.08%

(b) Z < 0.18

$$Z = \frac{x - \mu}{\sigma}$$

```
mu <- 0
sd <- 1
Z <- 0.18
# finding value for 'x'
x <- Z * sd + mu
x
```

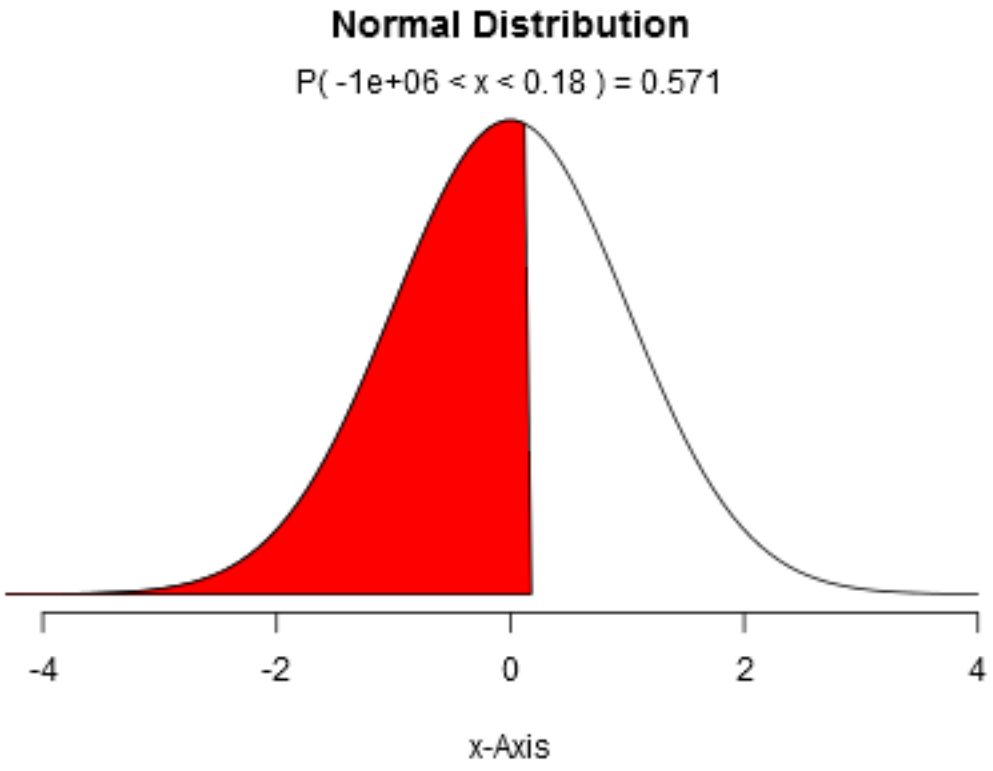
```
## [1] 0.18
```

Since we have that $x < 0.18$

```
# Finding probability fo x < 0.18
pnorm(x, mean = 0, sd = 1)
```

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

```
# Probability curve plot
normalPlot(mean = 0, sd = 1, bounds = c(-1e+06, x), tails = FALSE)
```



Answer: The percentage represented on the region is: 57.14%

(c) Z > 8

$$Z = \frac{x - \mu}{\sigma}$$

```
mu <- 0
sd <- 1
Z <- 8
# finding value for 'x'
x <- Z * sd + mu
x
```

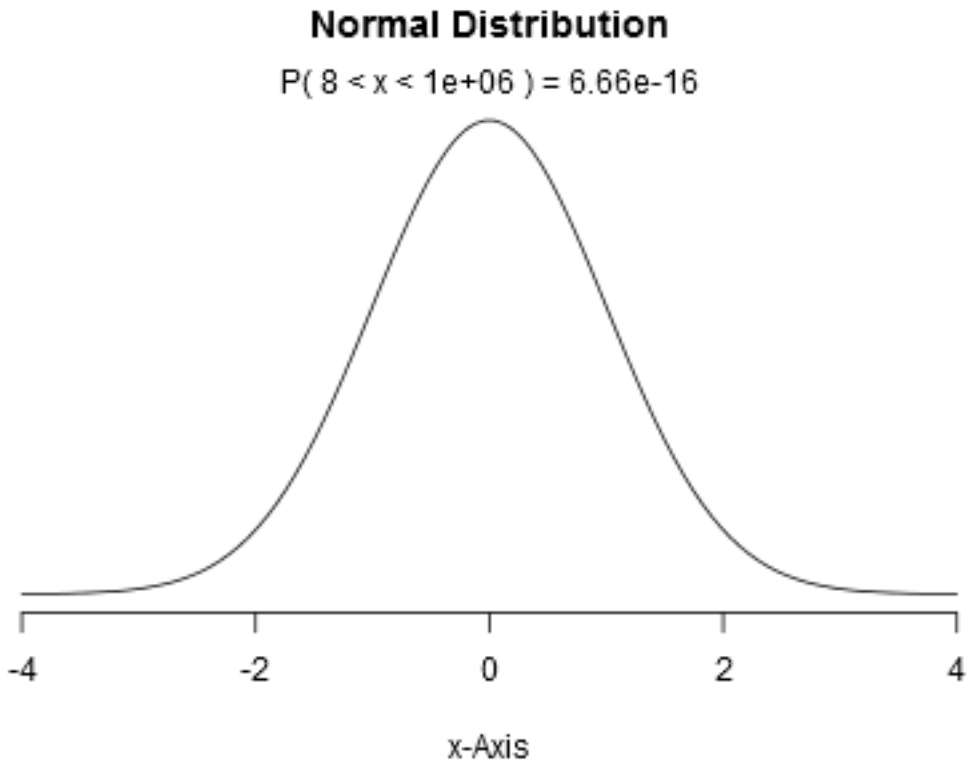
```
## [1] 8
```

Since we have that $x > 8$

```
# Finding probability fo x > 8
1 - pnorm(x, mean = 0, sd = 1)
```

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

```
# Probability curve plot
normalPlot(mean = 0, sd = 1, bounds = c(x, 1e+06), tails = FALSE)
```



Answer: The percentage represented on the region is: 0.00%

(d) $|Z| < 0.5$

$$Z = \frac{x - \mu}{\sigma}$$

```
mu <- 0
sd <- 1
Z <- 0.5
# finding value for 'x'
x <- Z * sd + mu
x
```

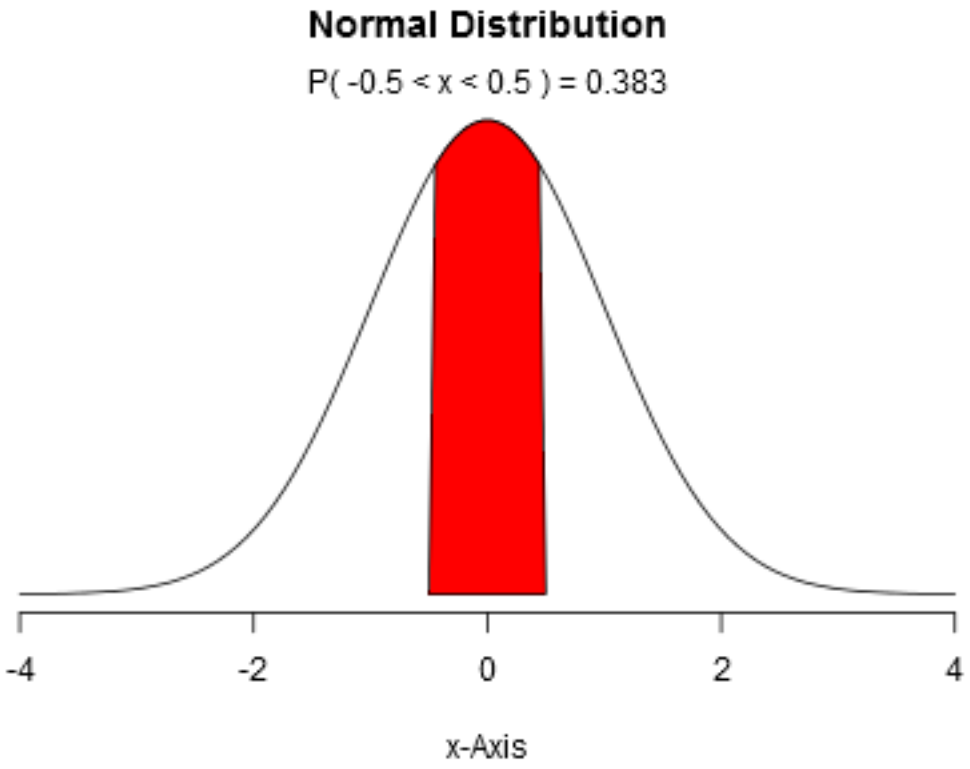
```
## [1] 0.5
```

Since we have that $|x| < 0.5$

```
# Finding probability fo |x| < 0.5 = -x < 0.5 < x
x1 <- pnorm(-x, mean = 0, sd = 1)
x2 <- pnorm(x, mean = 0, sd = 1)
x2 - x1
```

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

```
# Probability curve plot
normalPlot(mean = 0, sd = 1, bounds = c(-x, x), tails = FALSE)
```



Answer: The percentage represented on the region is: 38.29%

3.4 Triathlon times

Leo's group: Men, Ages 30 - 34.

Leo's race time: 1:22:28 (4948 seconds).

Men, ages 30 - 34 mean: 4313 seconds.

Men, ages 30 - 34 standard deviation: 583 seconds.

Mary's Group: Women, Ages 25 - 29.

Mary's race time: 1:31:53 (5513 seconds).

Women, ages 25 - 29 mean: 5261 seconds.

Women, ages 25 - 29 standard deviation: 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.

Answer: Group Men, Ages 30 - 34: N(μ =4313, ?? =583), Group Women, Ages 25-29: N(μ =5261, ?? =807).

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

$$Z_{Runner} = \frac{Runner_{Time} - \mu}{\sigma}$$

```
# Leo's Z Score
RunnerLTime <- 4948
GroupLMu <- 4313
GroupLSD <- 583
ZLeo <- (RunnerLTime - GroupLMu)/GroupLSD
ZLeo
```

```
## [1] 1.089194
```

```
# Mary's Z Score
RunnerMTime <- 5261
GroupMMu <- 4313
GroupMSD <- 807
ZMary <- (RunnerMTime - GroupMMu)/GroupMSD
ZMary
```

```
## [1] 1.174721
```

Answer: Mary had a better performance since her Z score is higher than Leo's.

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

```
# Leo's Rank
LeoP <- 1 - pnorm(RunnerLTime, mean = GroupLMu, sd = GroupLSD)
LeoP
```

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

```
# Mary's Rank
MaryP <- 1 - pnorm(RunnerMTime, mean = GroupMMu, sd = GroupMSD)
MaryP
```

```
## [1] 0.1200531
```

Answer: Mary performed better in her group than Leo did on his group, this is due to Mary is in the top 12% of best times for her group, while Leo is in top 13.80% for best times on his group.

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

```
pnorm(RunnerLTime, mean = GroupLMu, sd = GroupLSD)
```

```
## [1] 0.8619658
```

Answer: Leo finished faster than 86.20% triathletes in his group.

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

```
pnorm(RunnerMTime, mean = GroupMMu, sd = GroupMSD)
```

```
## [1] 0.8799469
```

Answer: Mary finished faster than 87.99% 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.

Answer: The answer to part (b) would not change as Z-scores can be calculated for distributions that are not normal. However, we could not answer parte since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

3.18 Heights of female college students.

Read data from GitHub file

```
url <- "https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from%20openintro.org
height <- read.table(url, header = TRUE, stringsAsFactors = FALSE)
head(height)
```

```
##   heighs
## 1     54
## 2     55
## 3     56
## 4     56
## 5     57
## 6     58
```

(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.

```
hmean <- 61.52
hsd <- 4.58
height$z <- (height$heighs - hmean)/hsd
height$nearest_sd <- round(height$z, 0)
kable(height)
```

heighs	z	nearest_sd
54	-1.6419214	-2
55	-1.4235808	-1
56	-1.2052402	-1
56	-1.2052402	-1
57	-0.9868996	-1
58	-0.7685590	-1
58	-0.7685590	-1
59	-0.5502183	-1
60	-0.3318777	0
60	-0.3318777	0
60	-0.3318777	0
61	-0.1135371	0
61	-0.1135371	0
62	0.1048035	0
62	0.1048035	0

heighs	z	nearest_sd
63	0.3231441	0
63	0.3231441	0
63	0.3231441	0
64	0.5414847	1
65	0.7598253	1
67	1.1965066	1
67	1.1965066	1
69	1.6331878	2
73	2.5065502	3

$$Z = \frac{x - \mu}{\sigma}$$

```
# x found by using Z scores within 1 Standard deviation
mu <- hmean
sd <- hsd
Z <- 1
x1 <- Z * sd + mu
x1
```

```
## [1] 66.1
```

```
# Probaility found by using data
sum(height$heighs < x1)/length(height$heighs)
```

```
## [1] 0.8333333
```

```
# Comapring Probability found by using 'pnorm'
pnorm(q = x1, mean = hmean, sd = hsd)
```

```
## [1] 0.8413447
```

```
# x found by using Z scores within 2 Standard deviation
mu <- hmean
sd <- hsd
Z <- 2
x2 <- Z * sd + mu
x2
```

```
## [1] 70.68
```

```
# Probaility found by using data
sum(height$heighs < x2)/length(height$heighs)
```

```
## [1] 0.9583333
```

```
# Comapring Probability found by using 'pnorm'
pnorm(q = x2, mean = hmean, sd = hsd)
```

```
## [1] 0.9772499
```

```
# x found by using Z scores within 3 Standard deviation
mu <- hmean
sd <- hsd
Z <- 3
x3 <- Z * sd + mu
x3
```

```
## [1] 75.26
```

```
# Probaility found by using data
sum(height$heighs < x3)/length(height$heighs)
```

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

```
# Comapring Probability found by using 'pnorm'
pnorm(q = x3, mean = hmean, sd = hsd)
```

```
## [1] 0.9986501
```

Answer: Yes, these heights approximately follow the 68-95-99.7% Rule since:

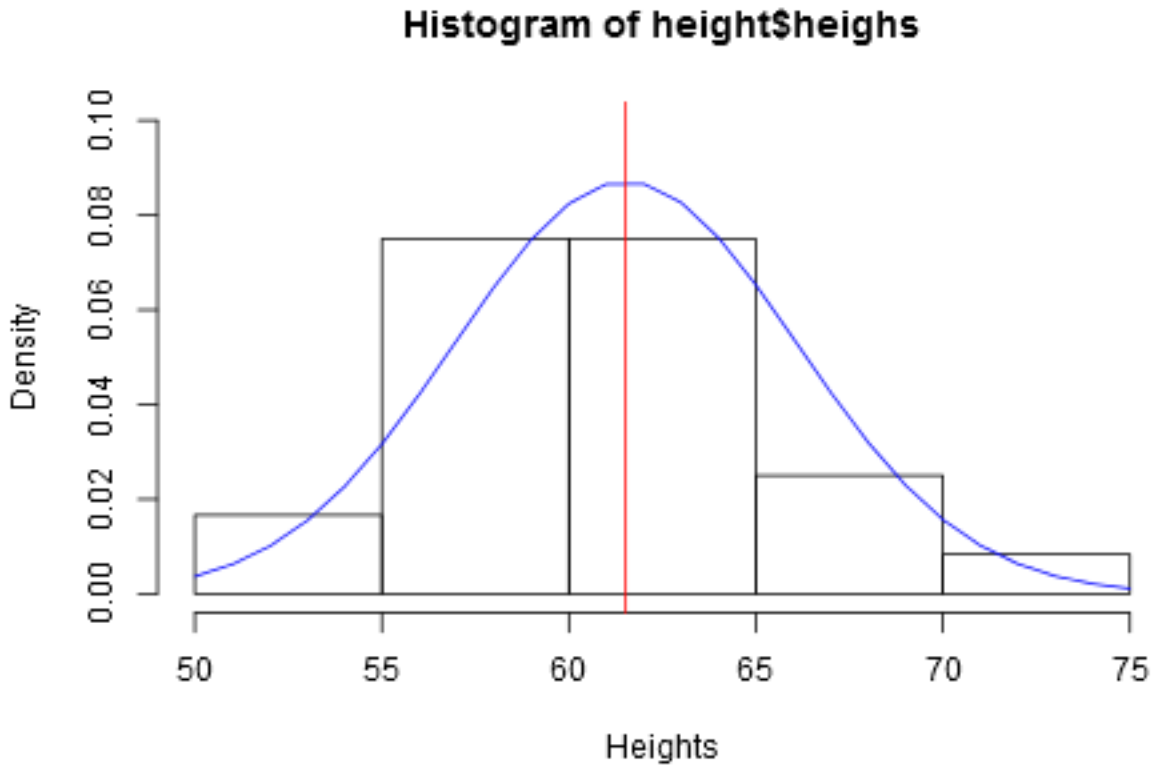
a) 83.33% of the data are within 1 standard deviation of the mean.

b) 95.83% of the data are within 2 standard deviation of the mean.

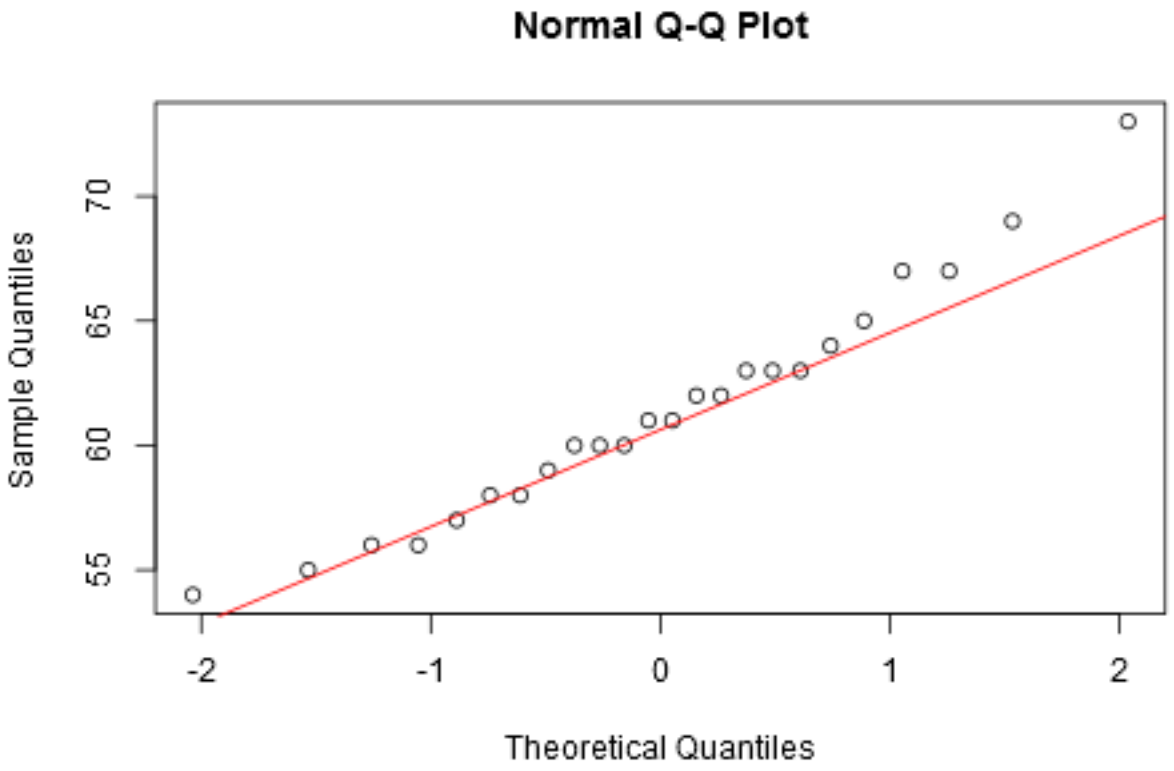
c) 100% of the data are within 3 standard deviation of the mean.

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

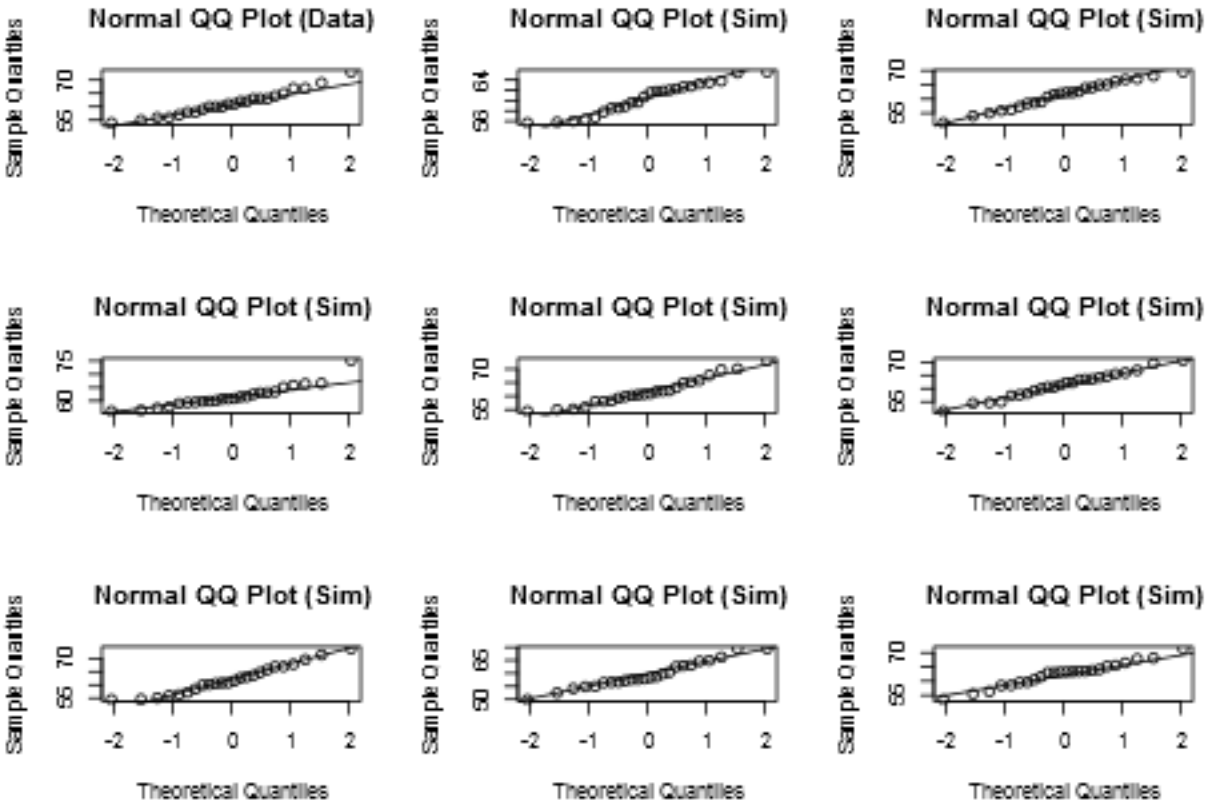
```
hist(height$heighs, probability = TRUE, xlab = "Heights", ylim = c(0, 0.1))
x <- 50:75
y <- dnorm(x = x, mean = mu, sd = sd)
lines(x = x, y = y, col = "blue")
abline(v = mu, col = "red")
```



```
qqnorm(height$heighs)
qqline(height$heighs, col = 2)
```




```
qqnormsim(height$heighs)
```



Answer: The distribution is unimodal and symmetric. The superimposed normal curve seems to approximate the distribution pretty well. The points on the normal probability plot also seem to follow a straight line. There is one possible outlier on the lower end that is apparent in both graphs, but it is not too extreme. We can say that the distribution is nearly normal.

3.22 Defective rate

Defective rate = 2%.

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?

```
# Rate of success and failure definition
pf <- 0.02
ps <- 1 - pf
n <- 10
# This is a geometric distribution
round(dgeom(n, ps), 4)
```

```
## [1] 0
```

```
round(ps * (1 - ps)^(n - 1), 4)
```

```
## [1] 0
```

Answer: The probability that the 10th transistor produced is the first with a defect is almost 0%.

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

```
# Rate of success and failure definition
pf <- 0.02
ps <- 1 - pf
n <- 100
# This is a geometric distribution
round(ps^n, 4)
```

```
## [1] 0.1326
```

Answer: 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?

$$E(X) = \frac{1}{p}$$

```
# Expected value of a geometric distribution
pf <- 0.02
Ex <- 1/pf
Ex
```

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

```
# Standard deviation f a geometric distribution
sd <- ((1 - pf)/pf^2)^(1/2)
sd
```

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

Answer: On average, I would expect to produce 50 transistors before the first one comes with a defect, with a standard deviation of 49.50.

(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?

```
# Expected value of a geometric distribution
pf <- 0.05
Ex <- 1/pf
Ex
```

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

```
# Standard deviation f a geometric distribution
sd <- ((1 - pf)/pf^2)^(1/2)
sd
```

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

Answer: On average, I would expect to produce 20 transistors before the first one comes with a defect, with a standard deviation of 19.50.

(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: When the probability of failure is bigger, the event is more common, meaning the expected number of trials before a success and the standard deviation of the waiting time are smaller.

3.38 Male children.

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
pboy <- 0.51
pboy2 <- choose(n, k) * (1 - pboy)^(n - k) * (pboy)^k
pboy2
```

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

Answer: The probability that two of them will be boys is 38.23%

(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.

```
children <- data.frame(c("BBG", "BGB", "GBB"))
children$p <- c(pboy * pboy * (1 - pboy), pboy * (1 - pboy) * pboy, (1 - pboy) *
    pboy * pboy)
names(children) <- c("Kids", "p")

sump <- sum(children$p)
sump
```

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

```
pboy2 - sump
```

```
## [1] 0
```

```
kable(children)
```

Kids	p
BBG	0.127449
BGB	0.127449
GBB	0.127449

Answer: Both results 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: The second method will be more tedious since we will have to create combination of 56 different possibilities making it very tedious to work with.

3.42 Serving in volleyball.

15% chance of making the serve.

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?

```
# This is a Negative Binomial distribution
p <- 0.15
n <- 10
k <- 3
choose(n - 1, k - 1) * (1 - p)^(n - k) * p^k
```

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

Answer: The probability that on the 10th try she will make her 3rd successful serve is 3.9%

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

Answer: The probability that her 10th serve will be successful is 15% since all 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?

Answer: The probabilities are different because in the negative binomial model the last trial is taken as a success by definition.