Chapter 4 - Distributions of Random Variables

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) $|Z| > 2$

```
## 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: bitops

## 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
```

Answer Area under the curve, PArt I

Answer (a)

```
za <- -1.35 #Z score needs to be below -1.35
sd <- 1 # standard deviation is 1
m <- 0 # mean is 0

# from this we can find the observeation for za

oa <- (sd * za)+m
oa
```

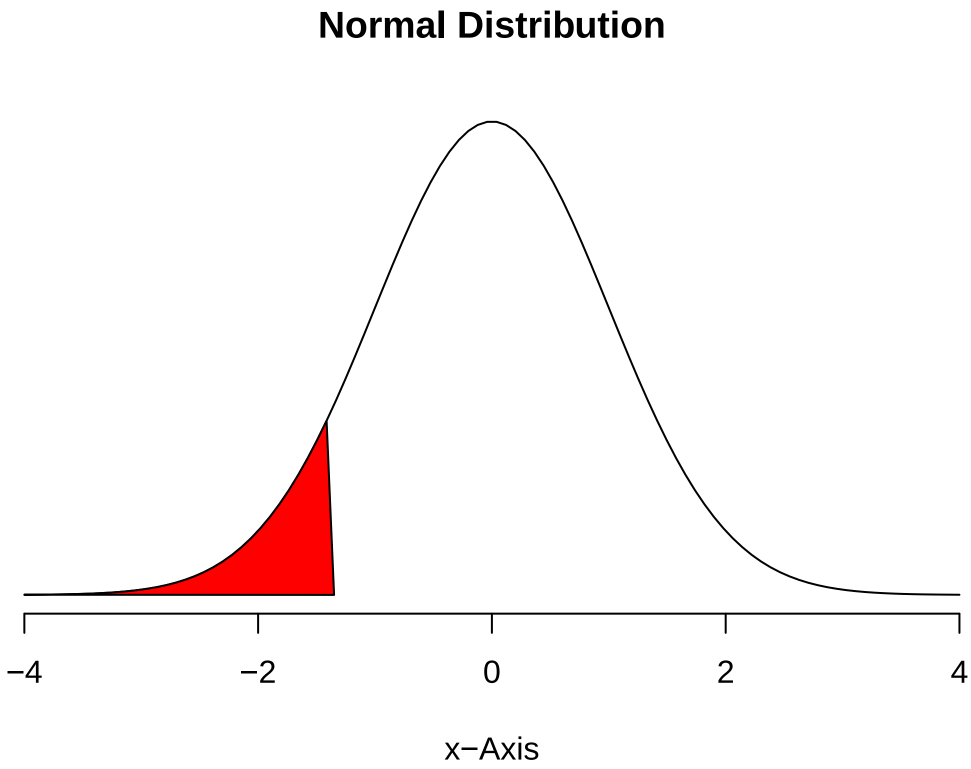
```
## [1] -1.35
```

```
# using pnorm to calculate the percentile

pnorm(oa, mean = 0, sd=1, lower.tail = TRUE)
```

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

Area under the curve for -135 is 0.088

```
# using normalPlat function to plot the normal distribution
normalPlot(mean=0, sd=1, bounds = c(oa), tails = TRUE)
```



Answer b

```
zb <- 1.48 #Z score needs to be above 1.48

# from this we can find the observeation for zb

ob <- (sd * zb)+m
ob
```

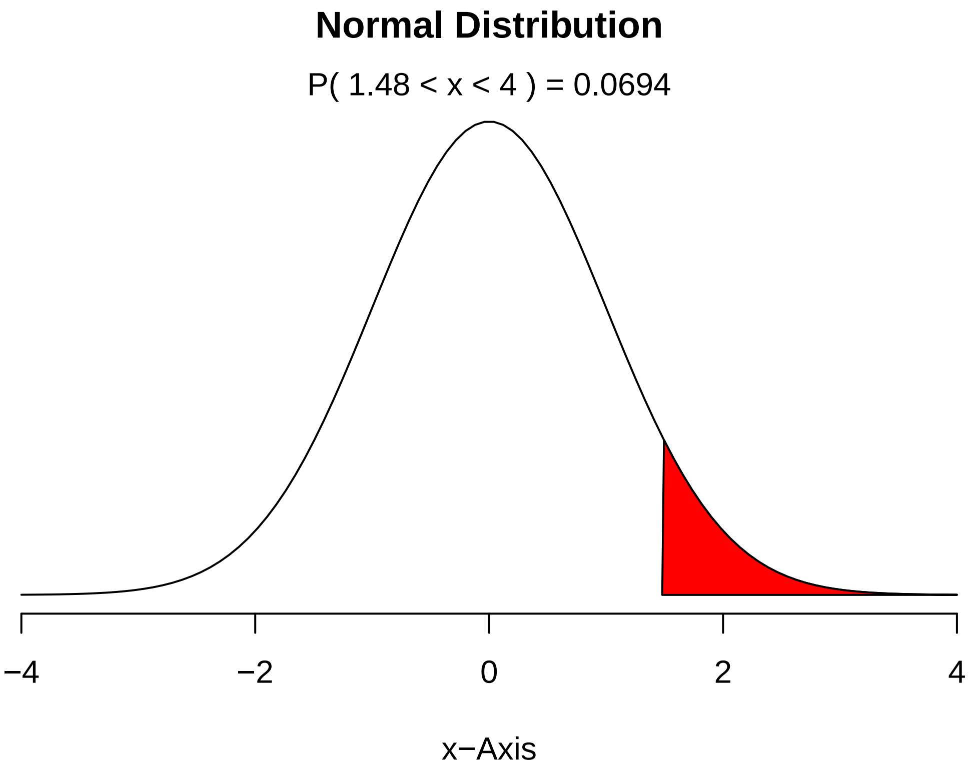
```
## [1] 1.48
```

```
# using pnorm to calculate the percentile

pnorm(ob, mean = 0, sd=1, lower.tail = FALSE)
```

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

Area under the curve for 148 and above is 0.069

```
# using normalPlat function to plot the normal distribution
normalPlot(mean=0, sd=1, bounds = c(ob, 4))
```



Answer (c)

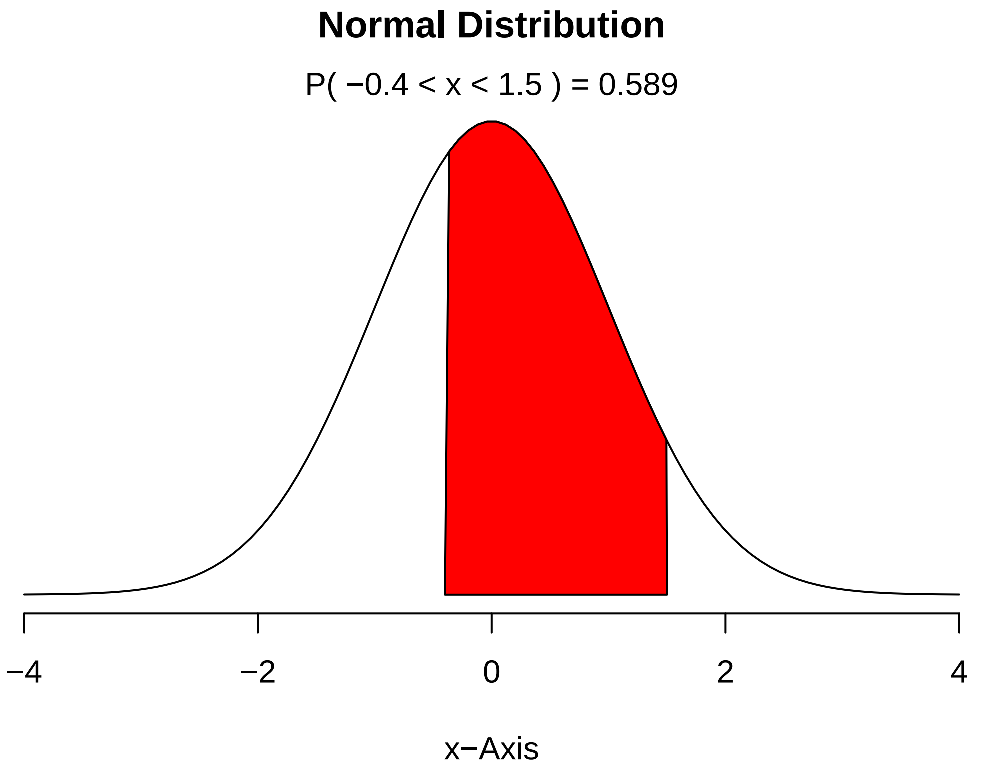
```
zclower <- -0.4 #Z score needs to be between -0.4 and 1.5
zcupper <- 1.5 #Z score needs to be between -0.4 and 1.5

# from this we can find the observeation for zc

oclower <- (sd * zclower)+m
ocupper <- (sd * zcupper)+m
oclower
```

```
## [1] -0.4
```

```
ocupper
```

```
## [1] 1.5
```

```
# using normalPlat function to plot the normal distribution
normalPlot(mean=0, sd=1, bounds = c(oclower, ocupper))
```



** The area under the curve is 0.589**

Answer (d)

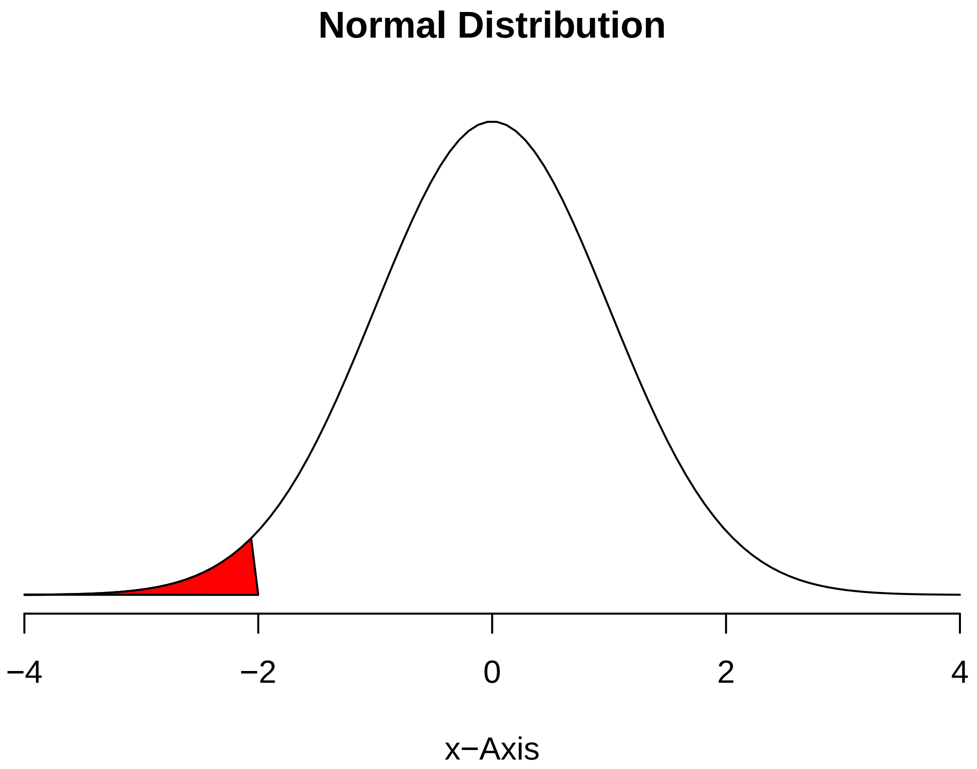
```
zdlower <- -2 #Z needs to be greater than 2 and lower than -2
zdupper <- 2

odlower <- (sd * zdlower)+m
odupper <- (sd * zdupper)+m
odupper
```

```
## [1] 2
```

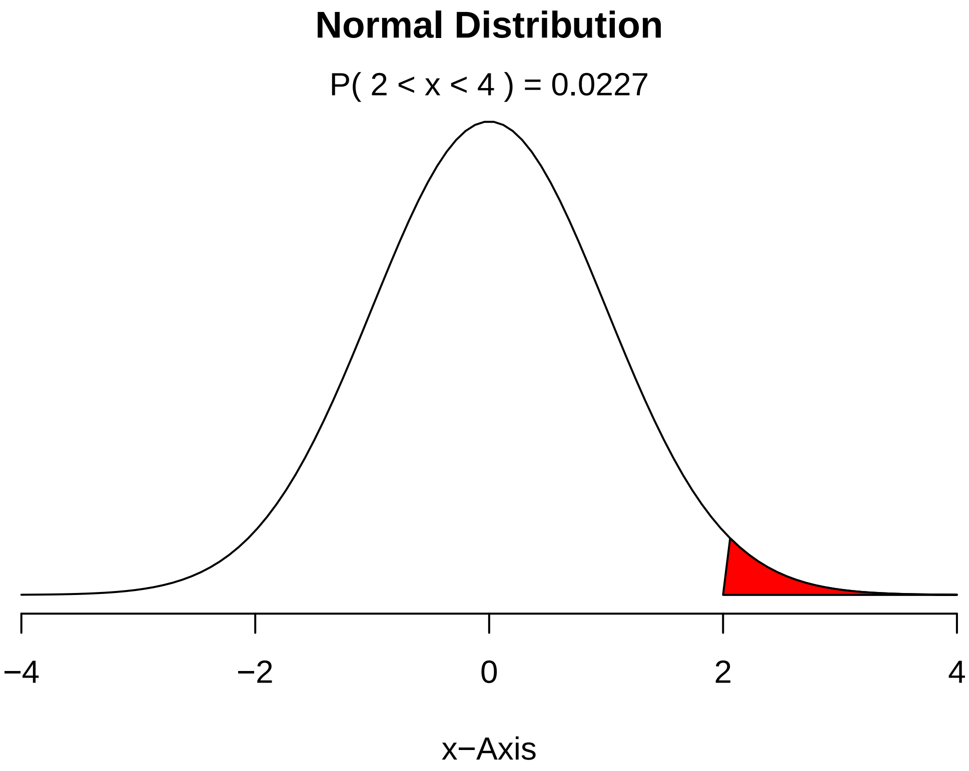
```
odlower
```

```
## [1] -2
```

```
#area below the curve that observation is less than -2 and area below the curve that the observation is

# using normalPlat function to plot the normal distribution
normalPlot(mean=0, sd=1, bounds = c(odlower), tails = TRUE)
```



```
normalPlot(mean=0, sd=1, bounds = c(odupper, 4))
```



```
pnorm(odlower, mean = 0, sd=1, lower.tail = TRUE)
```

```
## [1] 0.02275013
```

```
pnorm(odupper, mean=0, sd=1, lower.tail = FALSE)
```

```
## [1] 0.02275013
```

Area under the curve for each plot is 0.0227= Total area= 0.0454

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 Triathlon times, Part I

Answer (a)

Mens, Ages 30-34

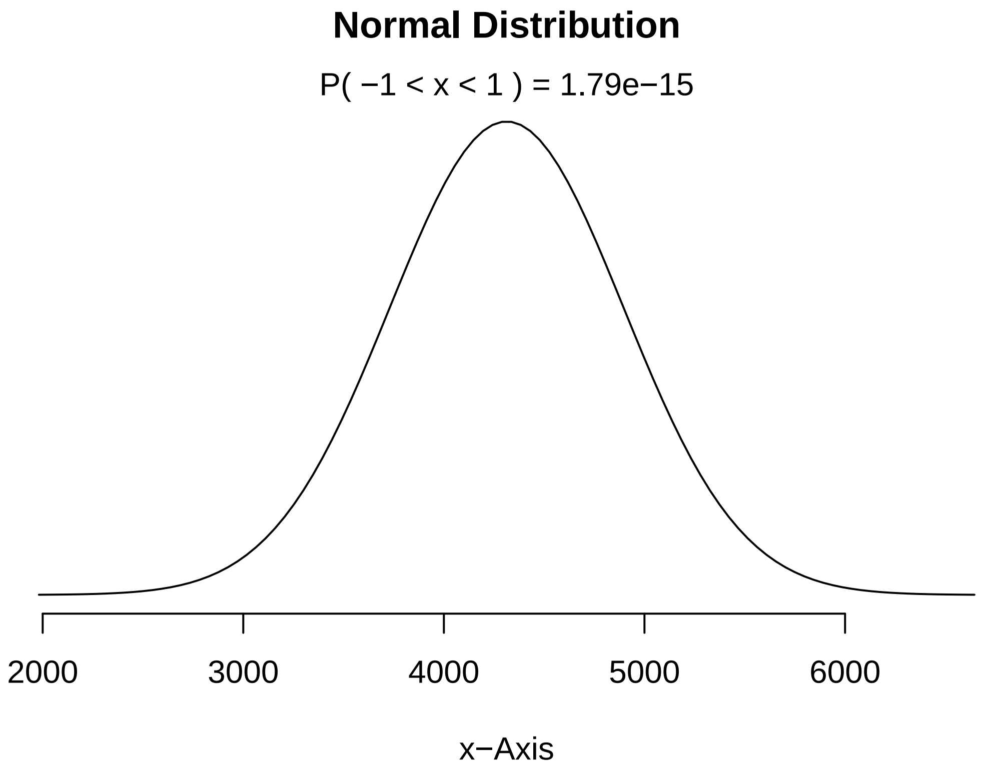
$$N(\mu = 4313, \sigma = 583)$$

Womens, Ages 25-29

$$N(\mu = 5261, \sigma = 807)$$

Answer (b)

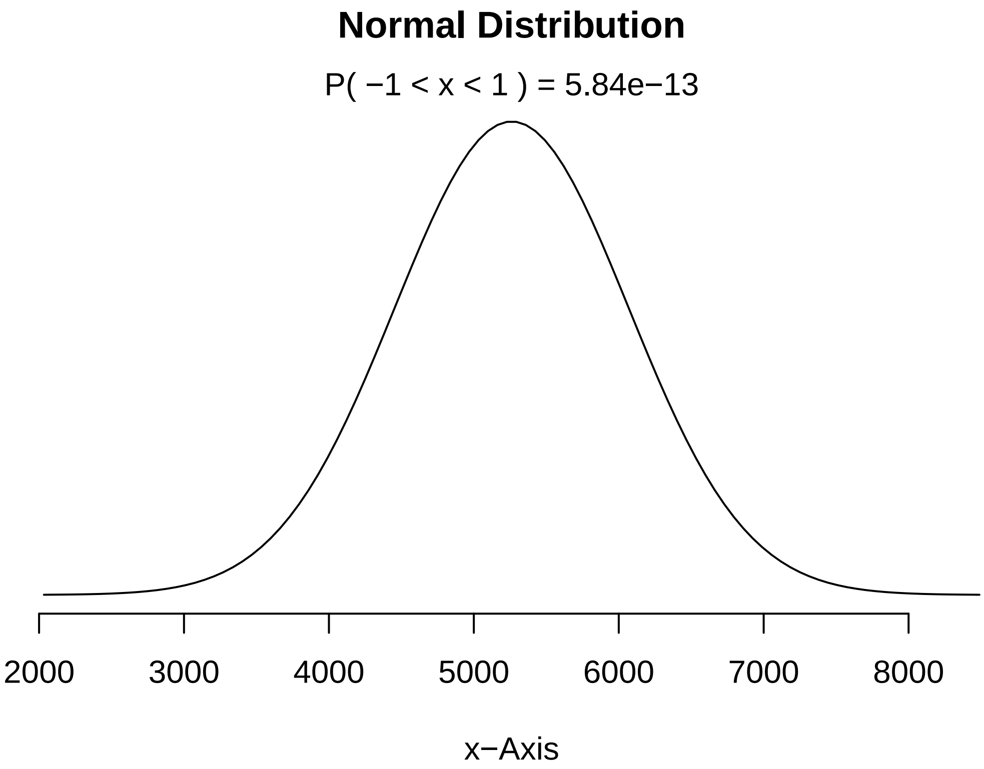
```
# looking at the normal distribution plot for Mens

normalPlot(mean=4313, sd=583)
```



```
# Z score for Leo

zl <- (4948-4313)/583
zl
```

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

```
# looking at the normal distribution for women

normalPlot(mean=5261, sd=807)
```



```
# calculating z score for Mary
zm <- (5513-5261)/807
zm
```

```
## [1] 0.3122677
```

Answer (c)

The z score for Leo is 1.089 ; meaning Leo's time is over 1 standard deviation away from the mean. Which means 1.089 standard deviation above the mean. On the other hand, Mary's z score is 0.312 which is less than half standard deviation away from the mean(above the mean). In this case, Mary did better than Leo within her group because it is closer to the mean which is time in seconds to finish the triatlhon. (considering having high seconds in terms of finishing a triathlon is not a good thing)

Answer(d)

```
# calculating Leo's percentile

pnorm(4948, mean = 4313, sd=583, lower.tail = TRUE)
```

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

Percentile for Leo is 86%. He finish time was higher than 86% of the group. His finish time was lower(better) than 100-86= 14% of the group.

Answer(e)

```r
# calculating Mary's percentile
pnorm(5513, mean = 5261, sd=807, lower.tail = TRUE)
```

```
## [1] 0.6225814
```

Mary's percentile is 62%. Mary's finish time was higher than 62% of the group. Her finish time was lower(better) than 100-62=38 % of the group.

Answer(f)

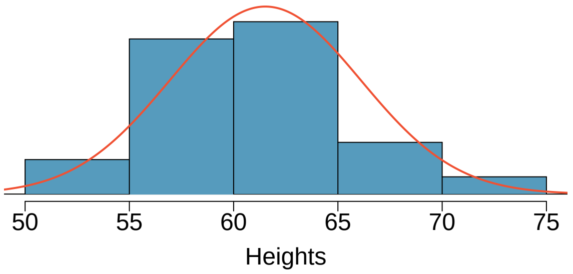
Yes, my responses to those questions would be different. If the distribution is not normal, we can not calculate the percentile. The density of the distribution would be different.

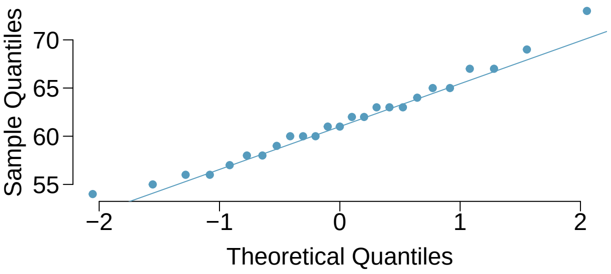
Heights of female college students Below are heights of 25 female college students.

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 Heights of Female Students

Answer (a)

In order to find out if the heights follow the 68-95-99.7% rule, we need to look at the distribution and confirm.

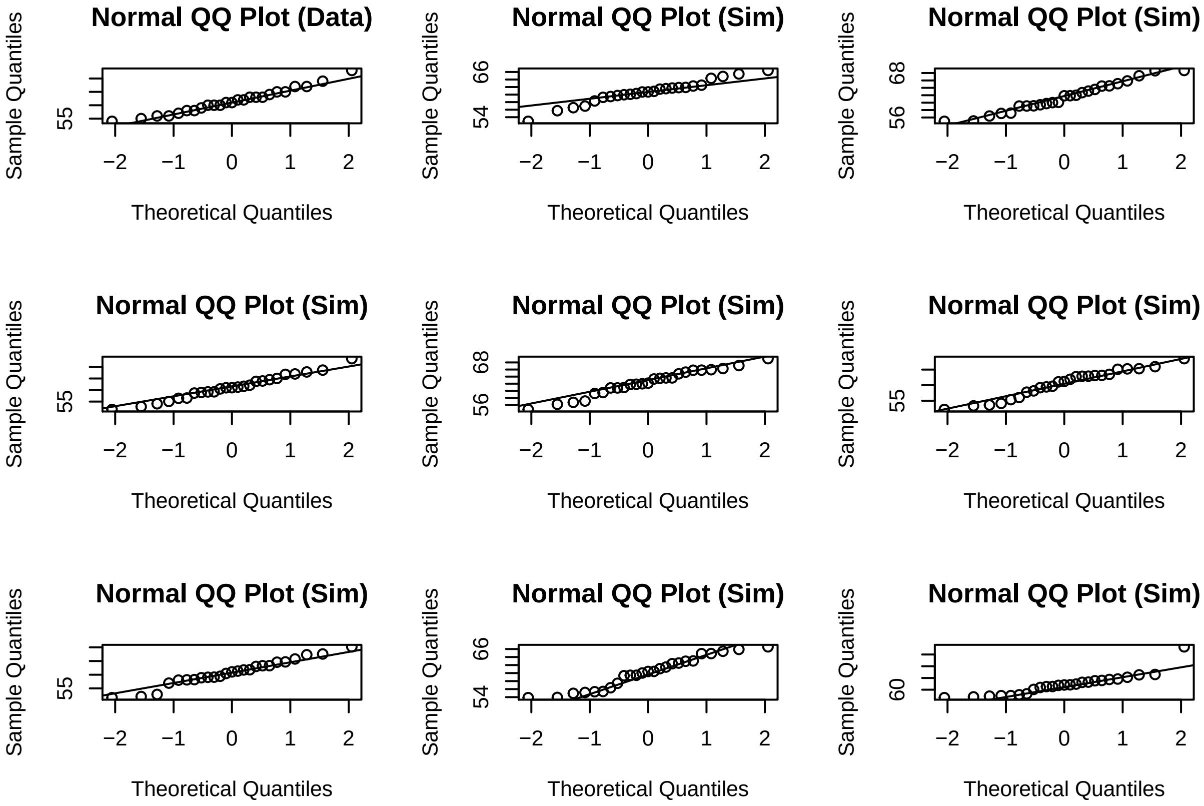
- 68% of the heights fall within 1 standard deviation away from the mean
- 95% of the heights fall within 2 standard deviation away from the mean
- 99.7% of the heights fall within 3 standard deviation away from the mean

This is used for normal distribution.

```
# creating the heights

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)
heights
```

```
##  [1] 54 55 56 56 57 58 58 59 60 60 60 61 61 62 62 63 63 63 64 65 65 67 67
## [24] 69 73
```

```
# Use the DATA606::qqnormsim function
qqnormsim(heights)
```



We can see that there is a linear relationship for each plot outlined here. The Theoretical Quanties are following the normal distribution. We can confirm that the heights approximately follow the 68-95-99.7% rule.

Answer(b)

As per the analysis provided in answer a, the data appear to follow normal distribution. When we look at the sample quantiles and theoretical quantiles we see the linear relationship. On the histrogram plot, we can see the normal distribution curve.

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 Defective Rate

Answer(a)

$$(1-p)^{n-1} * p$$

```
p <- 0.02 # probability of defective
n <- 10 # 10 trials

(1-p)^(n-1)*p
```

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

Answer(b)

```
# no defective probability is 1- defective probability
p2 <- 1-p
n2 <- 100 # 100 trials

(1-p2)^(n2-1)*p2
```

```
## [1] 6.211488e-169
```

Answer(c)

$$\mu = 1/p$$

```
m <- 1/p
m
```

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

The mean is 50. On average 50 transistors would be produced before the first defect.

$$\sigma = \sqrt{(1-p)/p^2}$$

```
sd <- sqrt((1-p)/p^2)
sd
```

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

Standard deviation is 49.4.

Answer (d)

```
p3 <- 0.05
m3 <- 1/p3
m3
```

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

On average 20 transistors would be produced before the first defect.

```
sd3 <- sqrt((1-p3)/p3^2)
sd3
```

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

Standard deviation is 19.4

Answer (e)

Increasing the probability of an event decreases the mean and standard deviation.

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 Male Children

Answer(a)

```r
pb <- 0.51 # probability of success
nb <- 3 # number of trials
kb <- 2 # number of success

dbinom(kb,nb,pb)
```

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

Answer(b)

1- Boy, Boy, Girl 2- Boy, Girl, Boy 3- Girl, Boy, Boy

```r
pg <- 1-pb
p1 <- pb*pb*pg
p2 <- pb*pg*pb
p3 <- pg*pb*pb

p1+p2+p3
```

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

Confirming answer a and answer b matches.

Answer (c)

We increased the trial number from 3 to 8 and success number from 2 to 3. It would be a lot harder for us to outline each possibility option manually. So option b would be much harder. On the other hand, we can use the dbinom function to calculate the probability of k success in n trial.

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 Serving in Volleyball

Answer (a)

```
nv <- 9 # number of trials (10th try)
pv <- 0.15 # probability of success
kv <- 2 # number of success (3rd successfull serve)


dbinom(kv, nv, pv)*0.15 # on her 10th try to get the 3rd success probability is 0.15 again.
```

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

Answer (b)

They are independent events as per the question so the probability of her serving successfully would be same 0.15

Answer (c)

Part b is calculating the probability of success for an event, which is provided by the question. In part a we are looking at all the probabilities for up to 10th try and multiplying them together for exact 3 successes out of 10 attempts.