Chapter 4 - Distributions of Random Variables

Raghed Mirza

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.

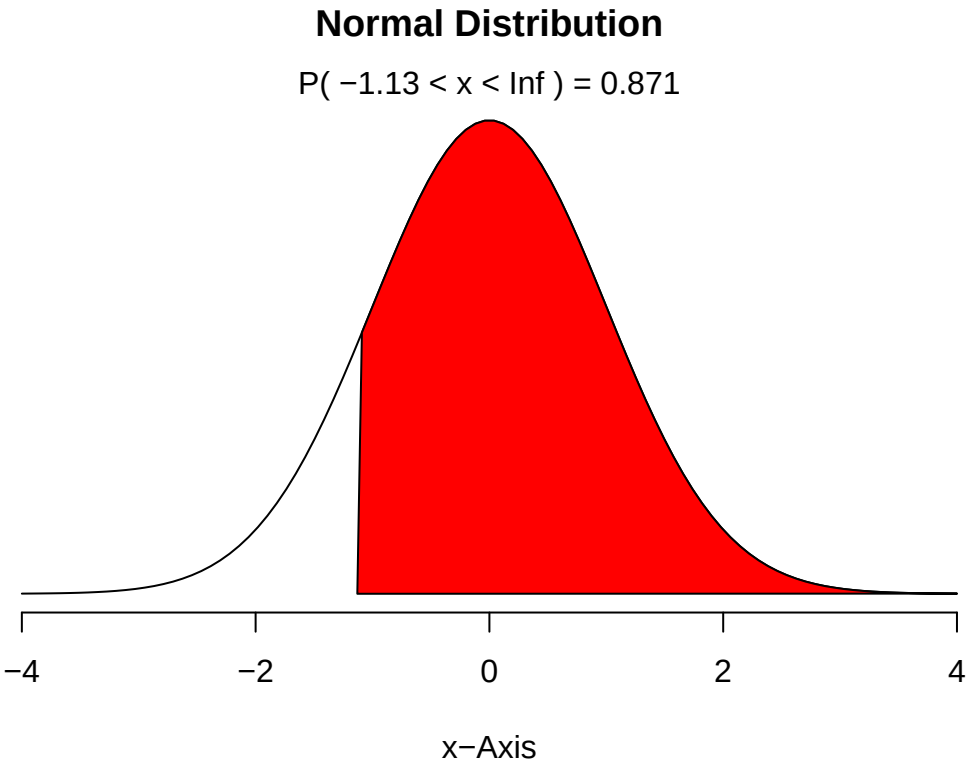
(a) $Z < -1.35$
(b) $Z > 1.48$
(c) $-0.4 < Z < 1.5$
(d) $|Z| > 2$

```
## 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.
```

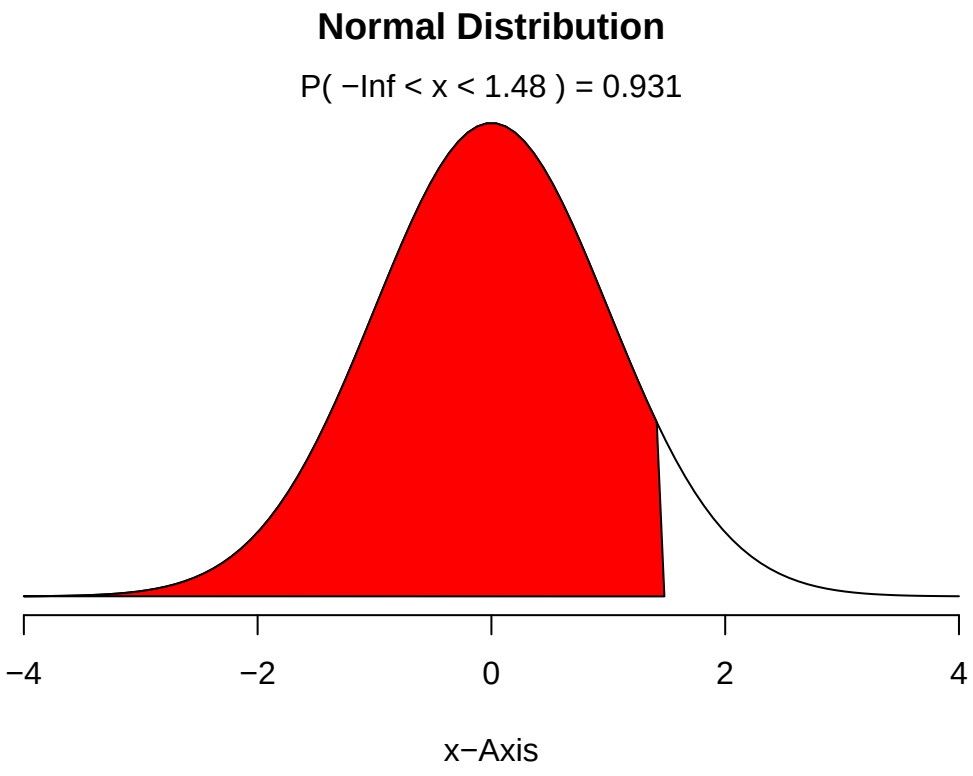
```
##
## Attaching package: 'DATA606'
```

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

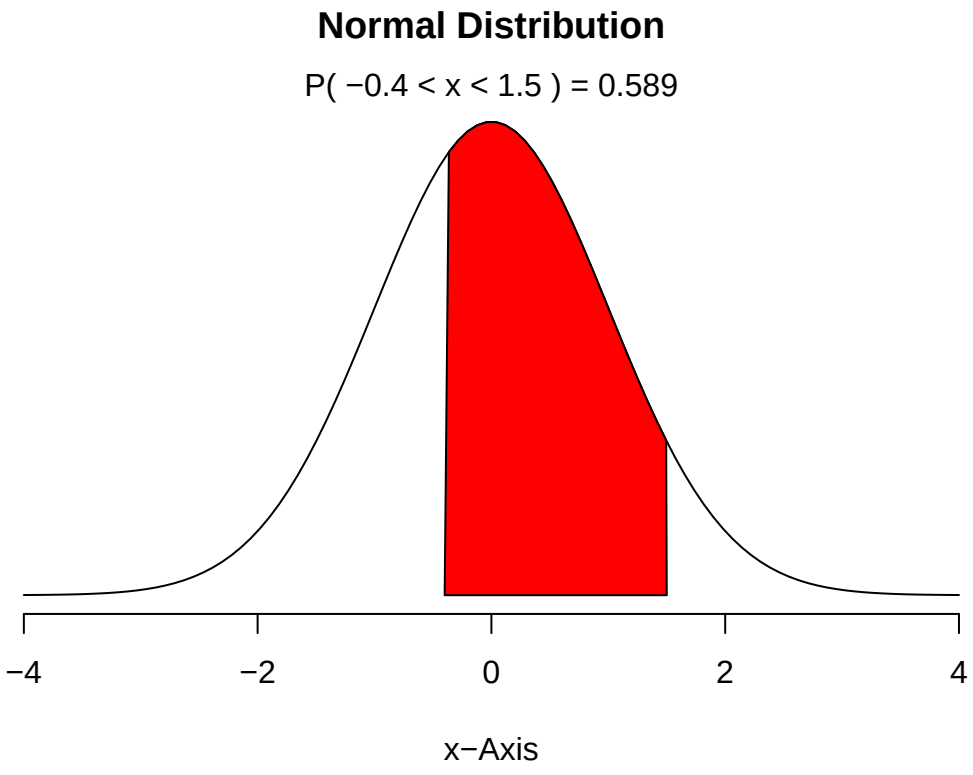
```
## [1] 0.1292381
```



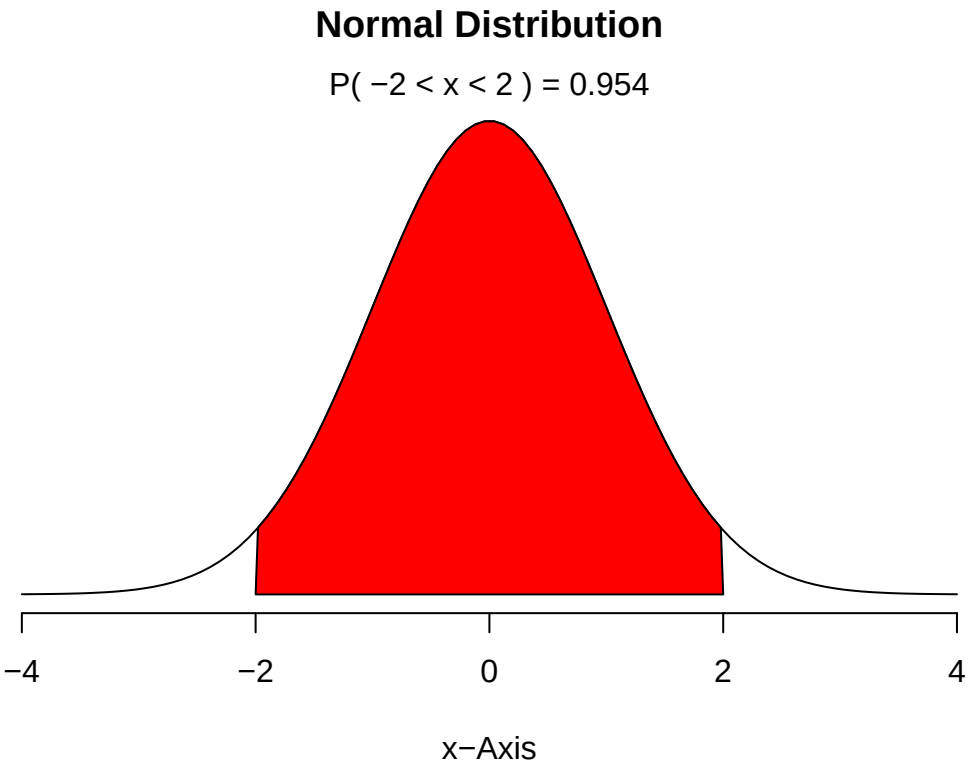

```
## [1] 0.06943662
```




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



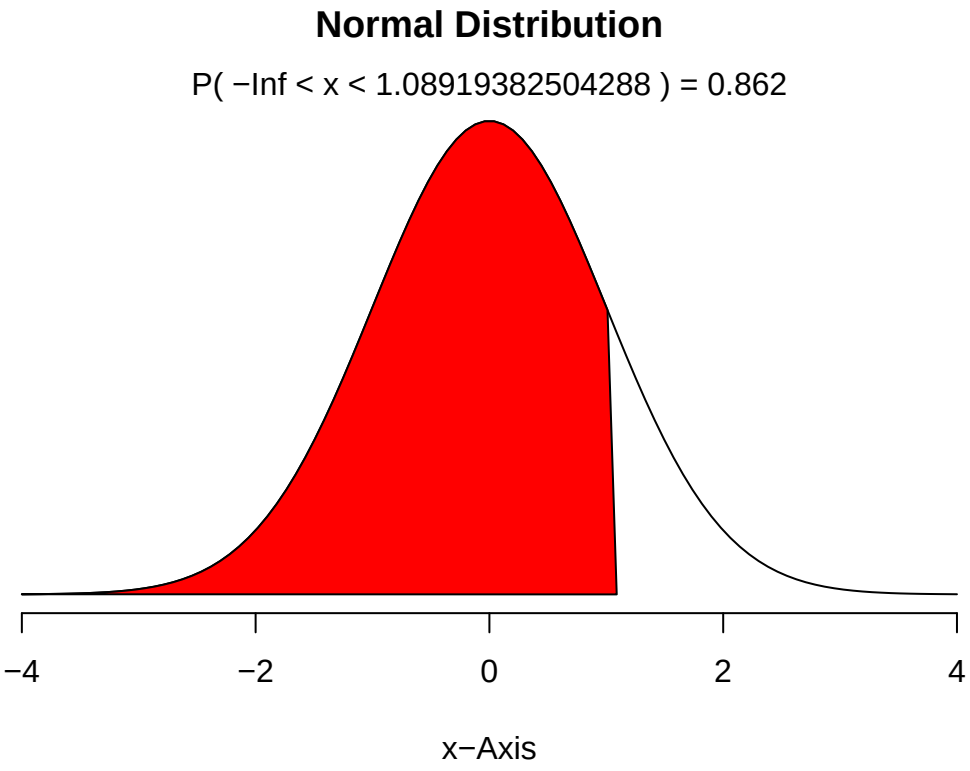

```
## [1] 0.02275013
```



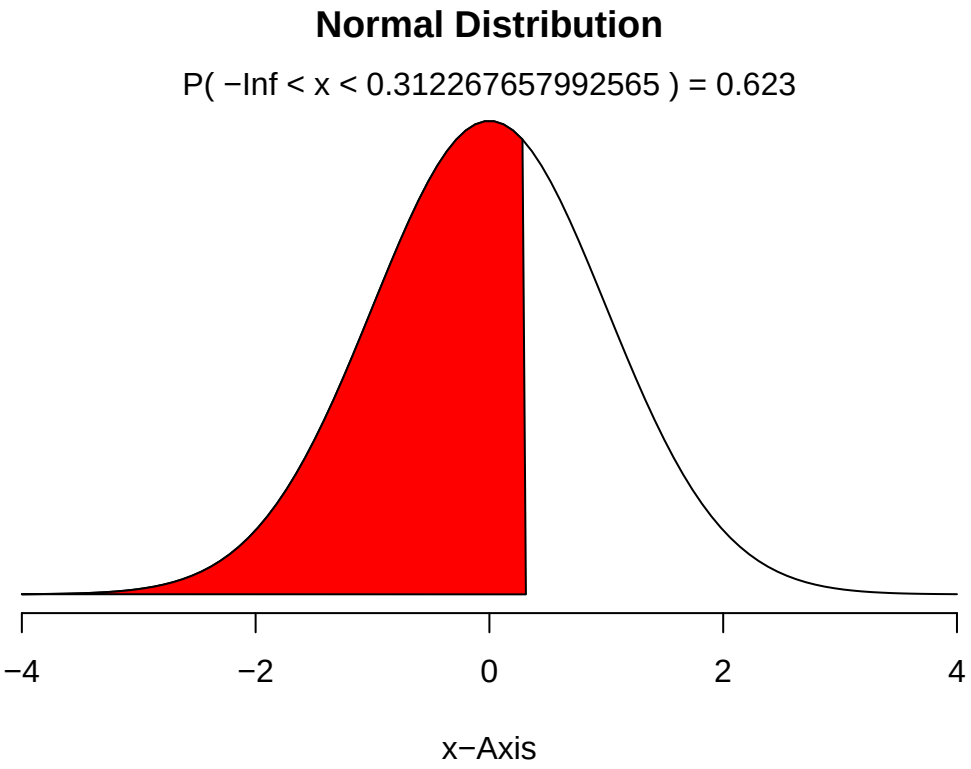

```
## [1] 1.089194
```

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

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



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


Normal Distribution
P(−Inf < x < 0.312267657992565) = 0.623
−4
−2
0
2
4
x−Axis

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.

```
# 4.4 (a)

#Distribution for Men's group (Ages 30-34) is N(mean = 4313, std. dev. = 583)
#Distribution for Women's group (Ages 25-29) is N(mean = 5261, std. dev. = 807)

# 4.4 (b)

# men's team's mean and std dev
m_mean <- 4313
m_sd <- 583


# women's team's mean and std dev
w_mean <- 5261
w_sd <- 807

# Leo's z-score
z_leo <- (4948 - m_mean) / m_sd
z_leo
```

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

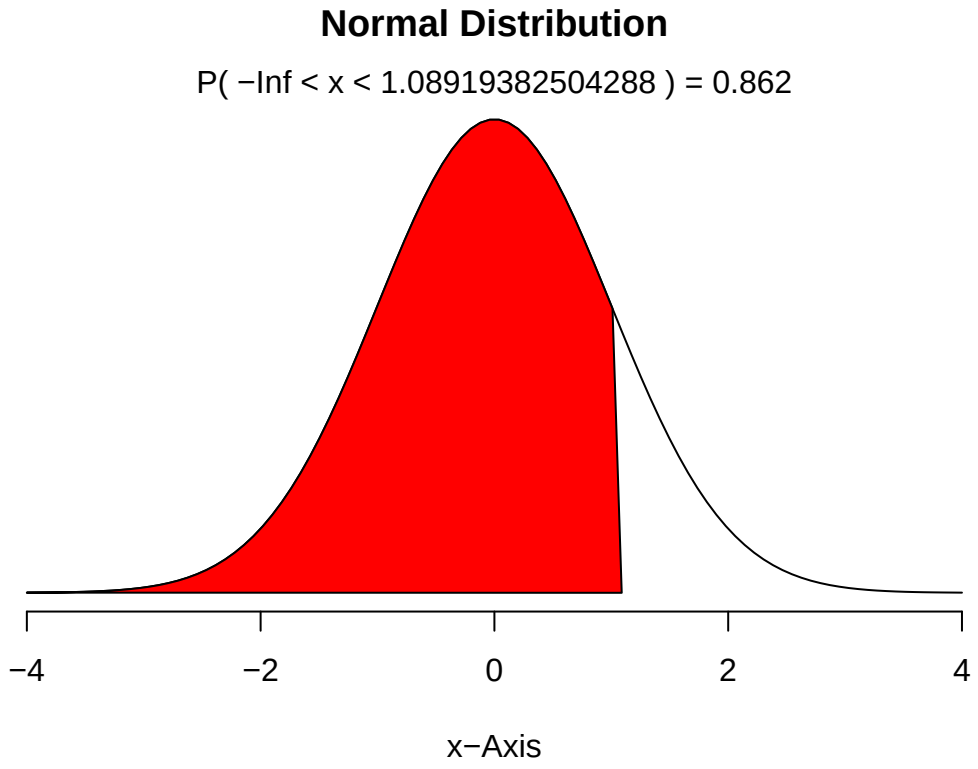
```
# Mary's z-score
z_mary <- (5513 - w_mean) / w_sd
z_mary
```

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

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

# Solution: Lower z-score for Mary indicates that she performed better compared to Leo. A -ve z-score w

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

# From Leo's z-score

1 - pnorm(z_leo, 0, 1)
```

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

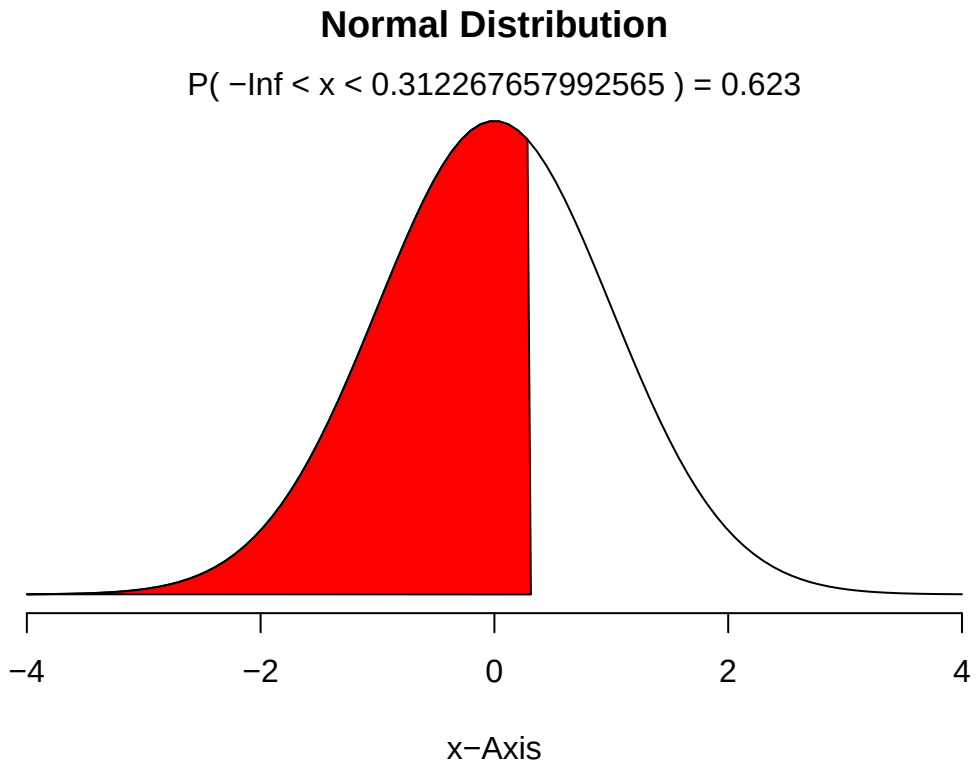
```r
normalPlot ( bounds = c(-Inf, z_leo))
```




```r
# Looking at the plot, Leo performed better than the 13.8%(1 - 0.862) of the triathletes in his group.

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

1 - pnorm(z_mary, 0, 1)
```

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

```r
normalPlot ( bounds = c(-Inf, z_mary))
```




```
# Looking at the plot, Mary performed better than the 37.7%(1 - 0.623) of the triathletes in her group.

# 4.4 (e)

# Solution: The answer for (b) could still be reliably calculated and the outcome could be explained in
```

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.

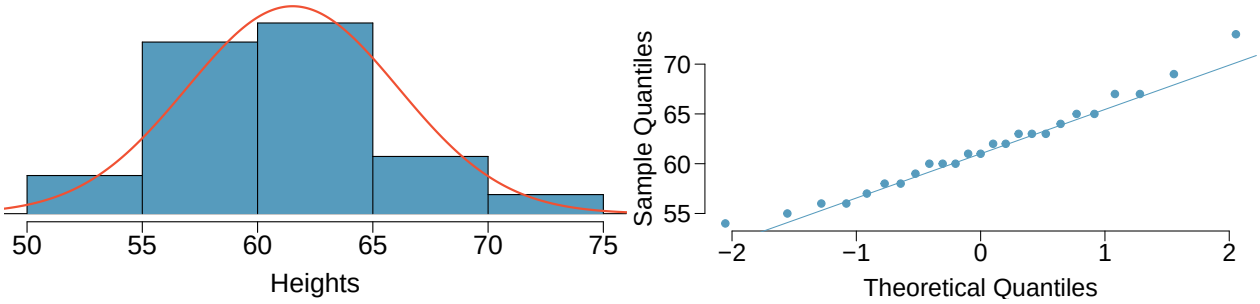
Solution

Yes, it seems to follow a normal distribution and all the data points seem to fall within 68-95-99.7% Rule.

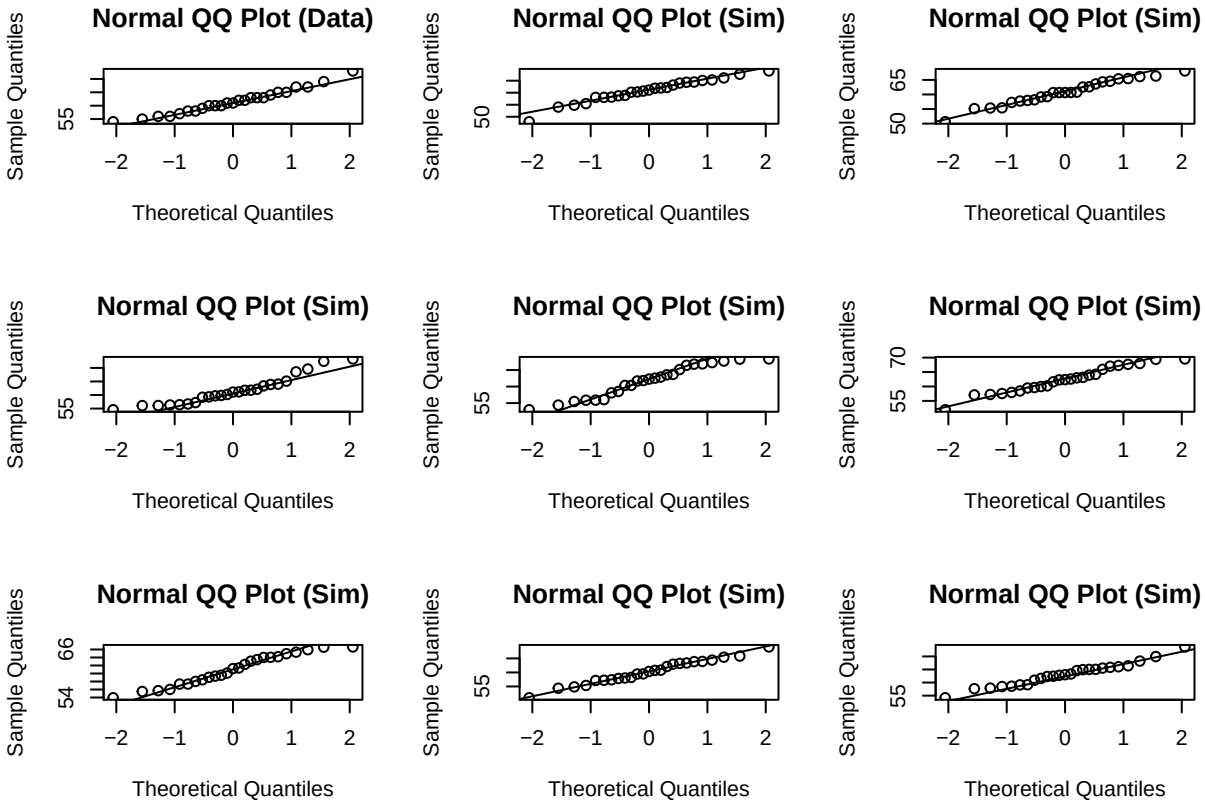
(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.

Solution

Yes, the data is following the normal distribution as majority of the data points fall under the bell curve and are distributed very close to the regression line except for a few that might be outliers.




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


Normal QQ Plot (Data)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
55
50
65
70
66
54

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?

```
# Solution (a)

pgeom(10-1,.02)
```

```
## [1] 0.1829272
```

```
# There is a probability of 0.183 that the 10th transistor is the first one with a defect.
```

```
# Solution (b)

1-pgeom(100,.02)
```

```
## [1] 0.1299672
```

```
# There is a probability of 0.13 that the machine produces no defective transistors in a batch of 100.
```

```
# Solution (c)

# The probability for the sample is:

prob <- 0.02

# The prabability of the first defective piece is:

def_prob <- (1 / prob) / 100

def_prob
```

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

```
# The probability is .5.

sd <- sqrt((1 - prob)/(prob^2))

sd
```

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

```r
# The stardard deviation is 49.497.
```

```r
# Solution (d)
prob2 <- 0.05

def_prob2 <- (1 / prob2) / 100

def_prob2
```

```
## [1] 0.2
```

```r
# The probability in this is case 0.2.

sd2 <- sqrt((1 - prob2)/(prob2^2))

sd2
```

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

```r
# The standard deviation is 19.494.
```

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).

```
# Solution (a)

dbinom(2,3,0.51)
```

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

```
# The probability is 0.382.
```

```
# Solution (b)

# The probability of having a girl is 0.49 (1 - 0.5).

# The possible ordering is ((.49*.51*.51), (.51*.49*.51), (.51*.51*.49))

boys_prob <- ((.49*.51*.51)+(.51*.49*.51)+(.51*.51*.49))
boys_prob
```

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

```
# The probability is 0.382 and the answer matches with (a).
```

```
# Solution (c)

# Amount of possible combinations is 56, which will entail a much longer calculation.

choose(n=8, k=3)
```

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

```
# The answer could be found using a simple formula:

dbinom(3, 8, 0.51)
```

```
## [1] 0.2098355
```

```
# The probability will be 0.21 in both cases.
```

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?

```
n_try <- 10

n_success <- 3

prob_sucess <- 0.15

prob_unsuccess <- (1 - 0.15)

prob3 <- (factorial(n_try - 1)/(factorial(n_success - 1)*factorial(n_try - n_success))*prob_sucessˆn_su

prob3
```

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

```
# The probability that on the 10th try player will make her 3rd successful serve is 0.39.

# Solution (b)

# The probability will still be 0.15 as each attempt is independant.

# Part (a) was based on a combination of events whereas part (b) was an independant event.
```