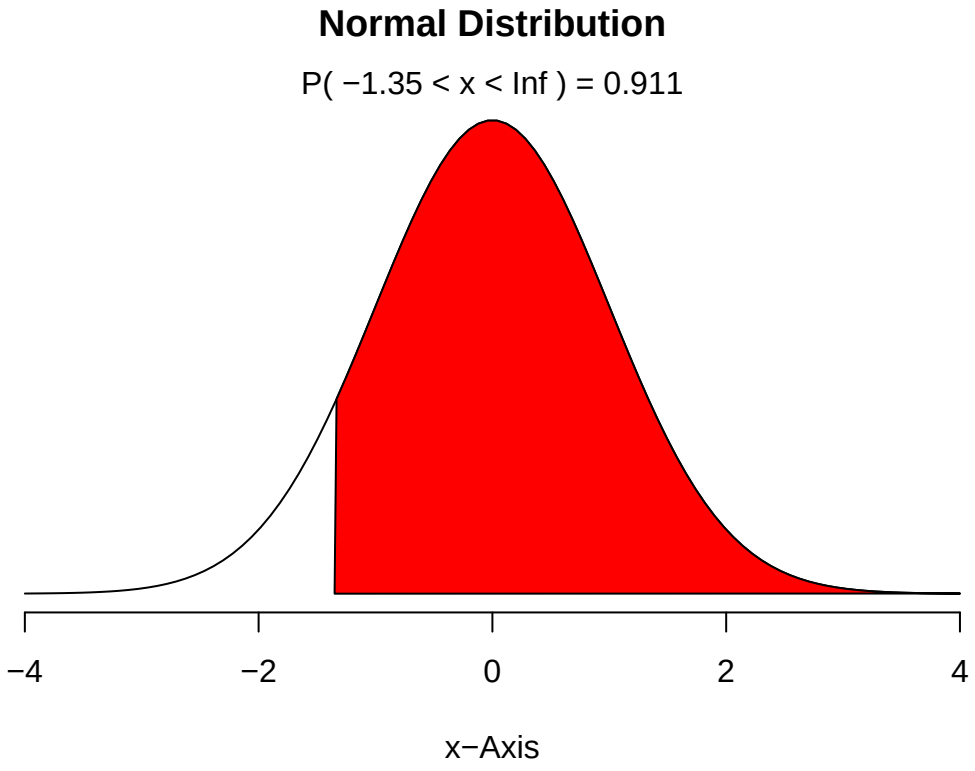
Chapter 4 - Distributions of Random Variables

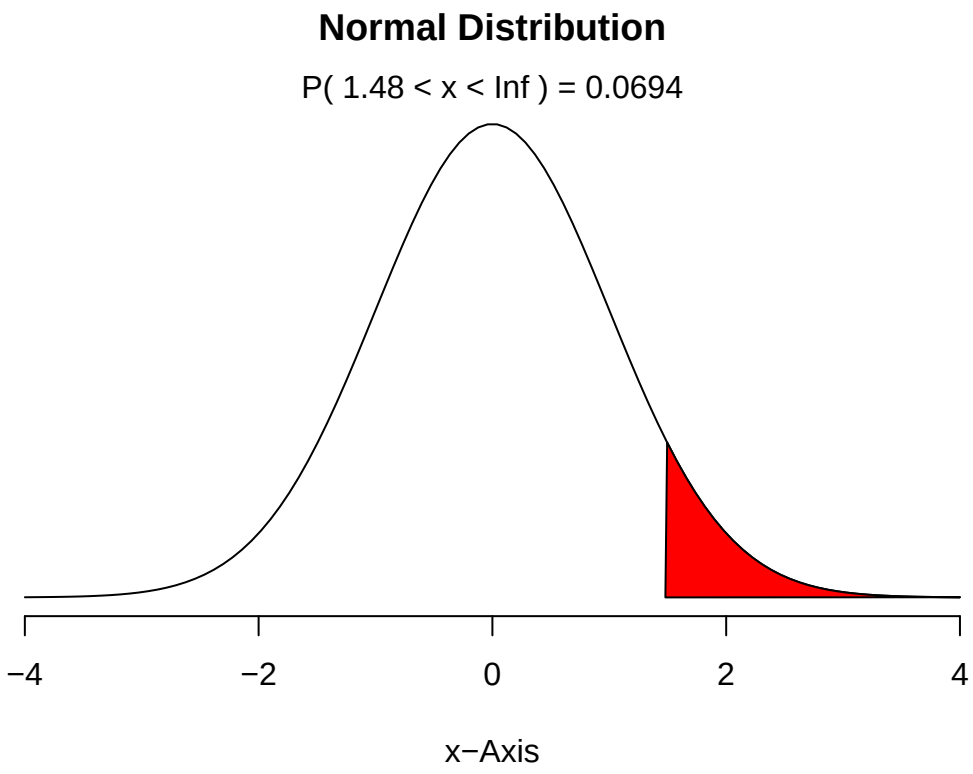
```
library(DATA606)
```

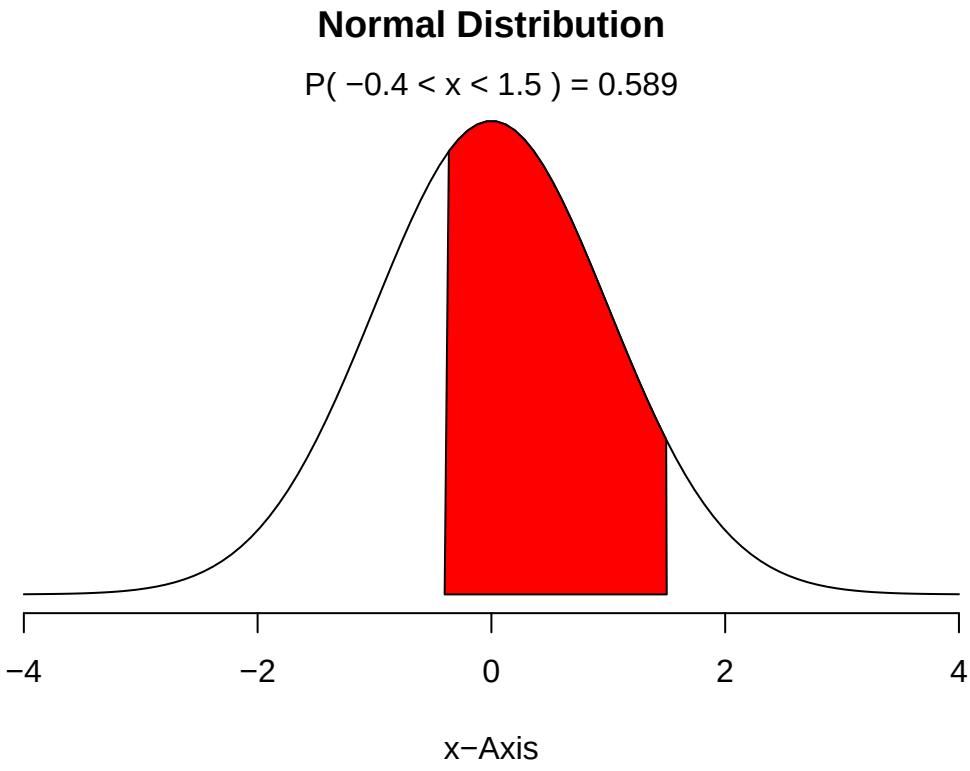
```
## Loading required package: shiny

## Loading required package: openintro

## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata

## Loading required package: OIdata

## Loading required package: RCurl

## Loading required package: maps

## Loading required package: ggplot2

## 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, 4th Edition. You can read this by typing
## vignette('os4') 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 objects are masked from 'package:openintro':
##
##     calc_streak, present, qqnormsim

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

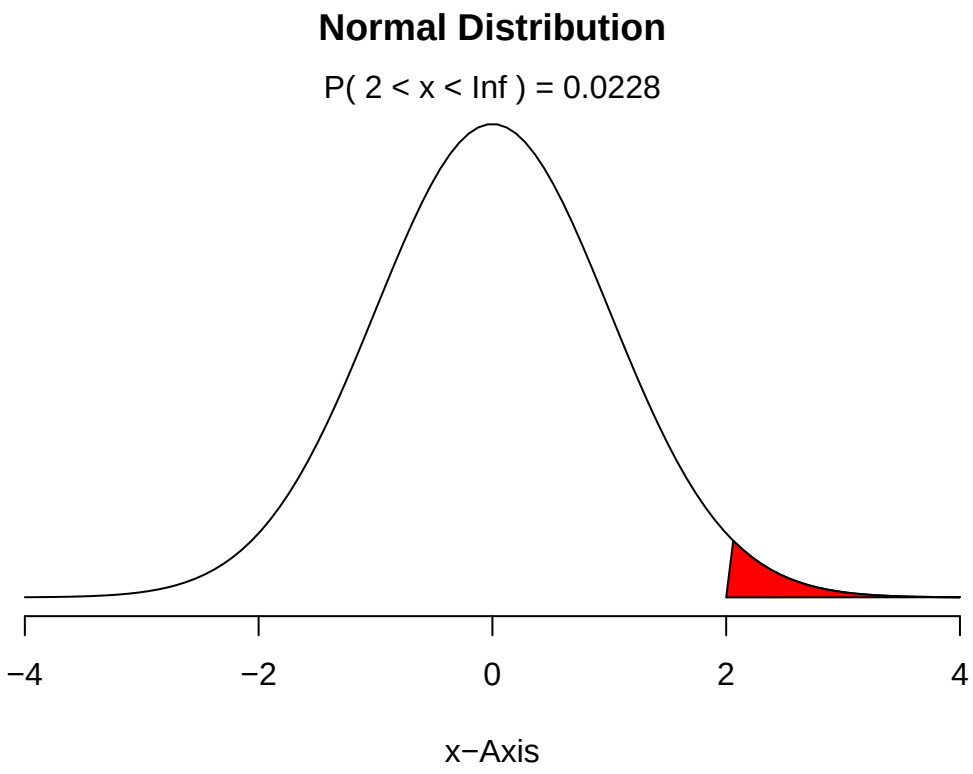
Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

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




Normal Distribution
P(1.48 < x < Inf) = 0.0694
−4
−2
0
2
4
x−Axis


Normal Distribution
P(−0.4 < x < 1.5) = 0.589
−4
−2
0
2
4
x−Axis


Normal Distribution
P(2 < x < Inf) = 0.0228
−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.

Answer ## Men = N(μ=4313,sigma =583) | Women = N(μ=5261,sigma =807)

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

Answer ## zscore= Racetime - mean / SD

#Men Ages 30-40: men_mean=4313 men_sd=583

#Women Ages 30-34: women_mean=5261 women_sd=807

Z_score_Leo<-(4948-men_mean)/men_sd Z_score_Leo

Z_score_Mary<-(5513-women_mean)/women_sd Z_score_Mary

Z-score is the number of standard deviations from the mean.

##Leo's result is 1.09 standard deviations from the mean while Mary's results is 0.31 standard deviations from the mean so Mary performed better than Leo.

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

Answer ## Mary ranks better since her result is closer to the median.

(d) What percent of the triathletes did Leo finish faster than in his group? ***Answer*** pnorm(Z_score_Leo,lower.tail=FALSE)

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

Answer pnorm(Z_score_Mary,lower.tail=FALSE)*100

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

Answer B and C would not change as the Z score does not account for distribution pattern. D,E results would get affected though as it would not give the correct proportion

---

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.

```
## Answer

##First, let's check weather 68.27% lie within one standard deviation.

#summary(heights)
mean_heights=61.52
sd_heights=4.58

pnorm(mean_heights+sd_heights,mean=mean_heights,sd=sd_heights)
```

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

```
## Probability for falling within 1 standard deviation of the mean is 84.13% but not close to 68%.

##Second, let's check weather 68.27% lie within 2 standard deviation.

pnorm(mean_heights+2*sd_heights,mean=mean_heights,sd=sd_heights)
```

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

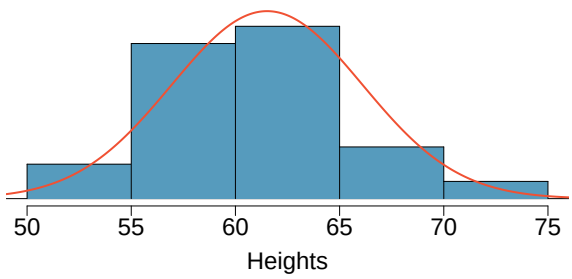
```
## Probability for falling within 2 standard deviation of the mean is 9772% but not close to 95%.

##Second, let's check weather 68.27% lie within 3 standard deviation.

pnorm(mean_heights+3*sd_heights,mean=mean_heights,sd=sd_heights)
```

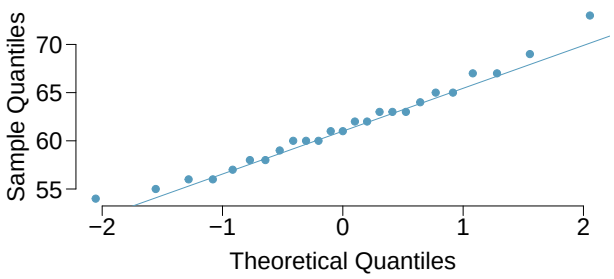
```
## [1] 0.9986501
```

```
## Probability for falling within 3 standard deviation of the mean is 99.86% which is close to 99.7%.

##CONCLUSION: DOES NOT FOLLOW 68-95-99.7% Rule.
```

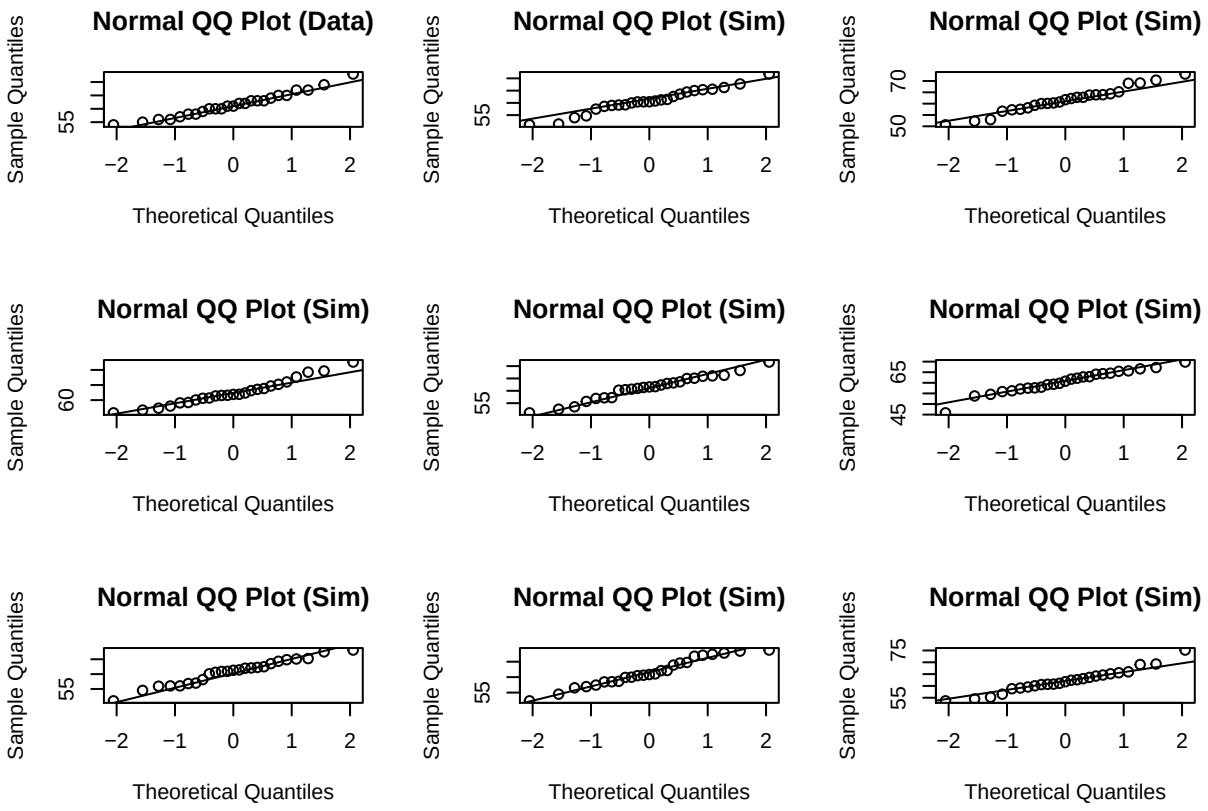
(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.

##ANSWER- based on the histogram the data is not perfectly normal distributed and looks bit right-skewed as we can see that The mean is bit greater than the median.





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




```r
#conclusion: we can see that Q-Q plot for the data is similar to that for the simulated data sets hence

##
```

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?

```
probFailure = .02
#9 is number of details before defective detail

dgeom(9,prob = probFailure)
```

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

```
#or

(1 - 0.02)^9 * 0.02
```

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

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

```
#Rate of "success"= 1-0.02. In order to find probability of getting 100 successes we have to multiply s

dbinom(0, 100, .02)
```

```
## [1] 0.1326196
```

```
#or

(1-0.02)^100
```

```
## [1] 0.1326196
```

```
#or
```

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

```
average = 1/0.02
average
```

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

```
sd=sqrt((1-0.02)/(0.02^2))
sd
```

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

(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?

```r
average = 1/0.05
average
```

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

```r
sd=sqrt((1-0.05)/(0.05^2))
sd
```

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

(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 : As probability increases mean and standard deviation decrease which means we can see the failure sooner than later .

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.

```
#Using the binomial model to calculate the probability that two of them will be boys

dbinom(2,3,0.51)
```

```
## [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.

{BGB, BBG, GBB} 3*(.49)*(.51^2)

(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). ## because we will have to write each combination separately and then add them to get the result.

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?

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

#number of cases with 2 successes and 7 failures in 9 first attempts

num_cases <- factorial(n-1)/(factorial(k-1)*(factorial(n-k)))
prob <-num_cases*(p^k)*((1-p)^(n-k))

prob
```

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

```
##OR
choose(9,2)*(.15^3)*(.85)^7
```

```
## [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?

##since the events are independent it should be 15%

(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?