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.

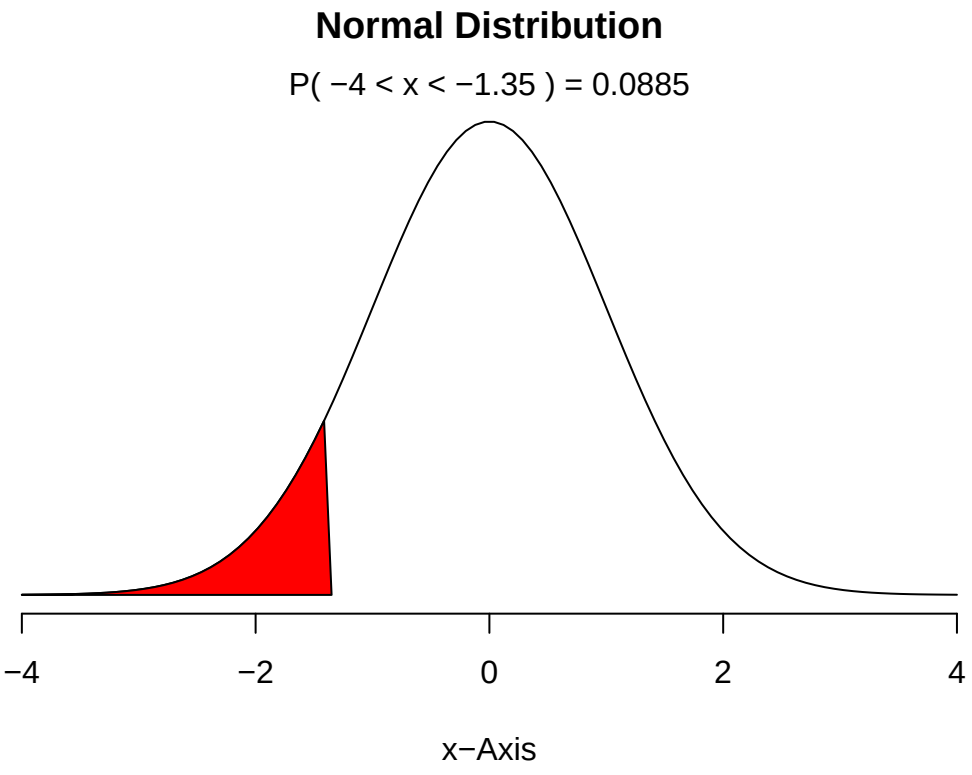
(a) $Z < -1.35$
(b) $Z > 1.48$
(c) $-0.4 < Z < 1.5$
(d) $|Z| > 2$

Answer for (a): roughly 8.85%

```
# use the DATA606::normalPlot function

# check the area under the curve
area_left <- pnorm(-1.35)
area_left
```

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

```
# plot for (a)
normalPlot(mean = 0, sd = 1, bounds = c(-4, -1.35), tails = FALSE)
```

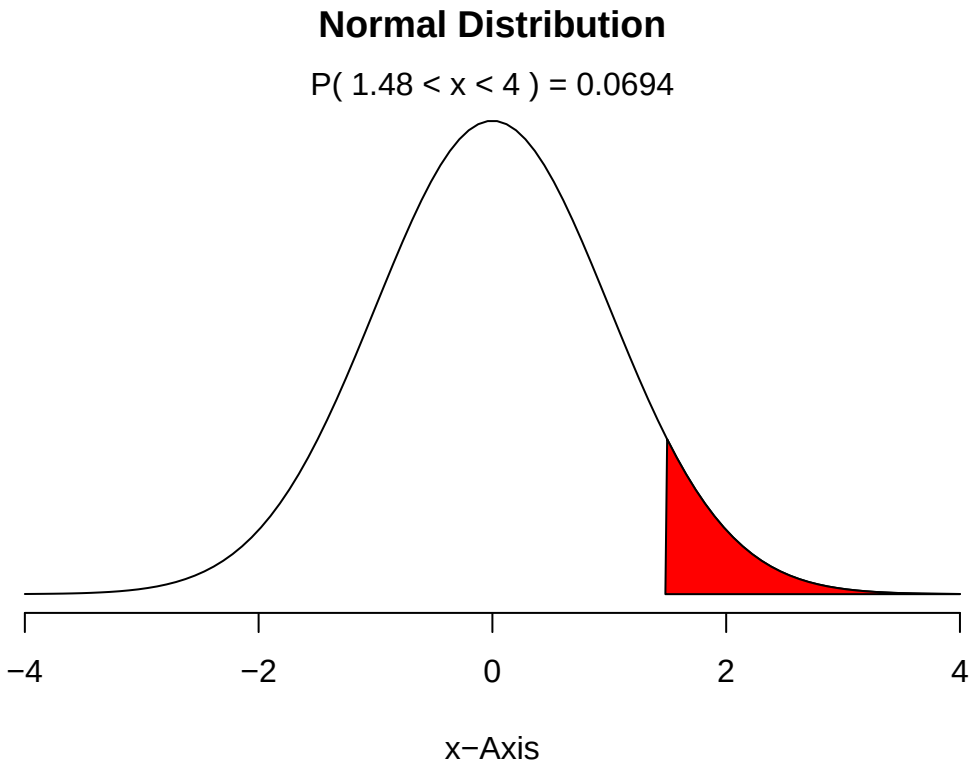


Answer for (b): roughly 6.94%

```
# use the DATA606::normalPlot function

# check the area under the curve
area_right <- 1 - pnorm(1.48)
area_right
```

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

```
# plot for (b)
normalPlot(mean = 0, sd = 1, bounds = c(1.48, 4), tails = FALSE)
```

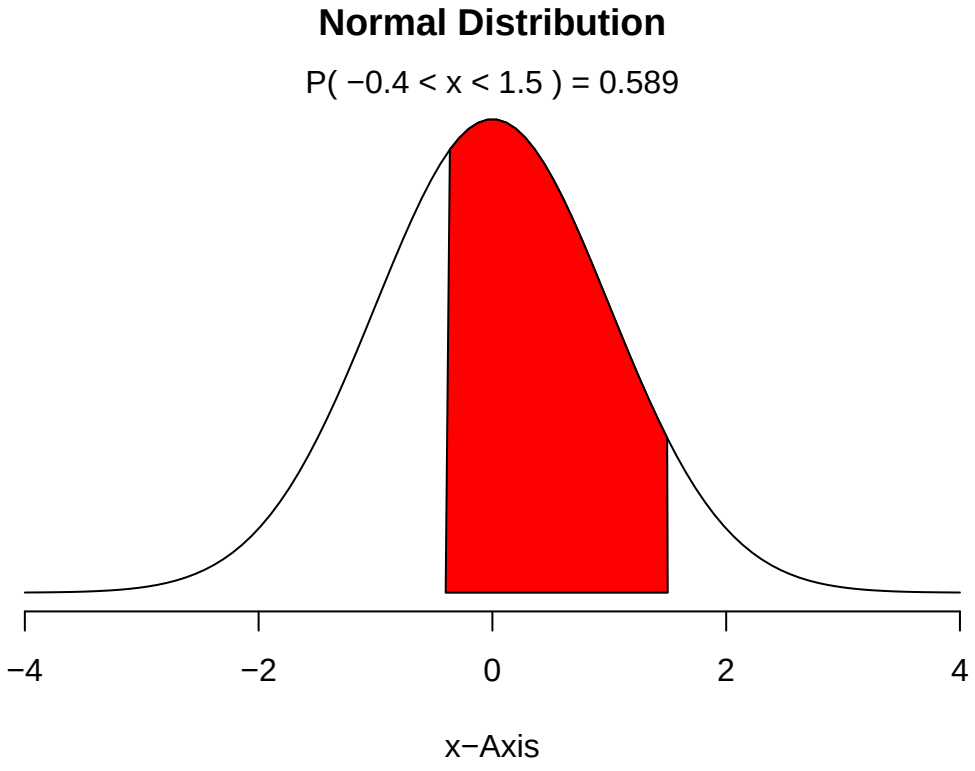


Answer for (c): roughly 58.9%

```
# use the DATA606::normalPlot function

# check the area under the curve
area_left <- pnorm(-0.4)
area_right <- 1 - pnorm(1.5)

percent <- 1 - (area_left + area_right)
percent
```

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

```
# plot for (c)

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

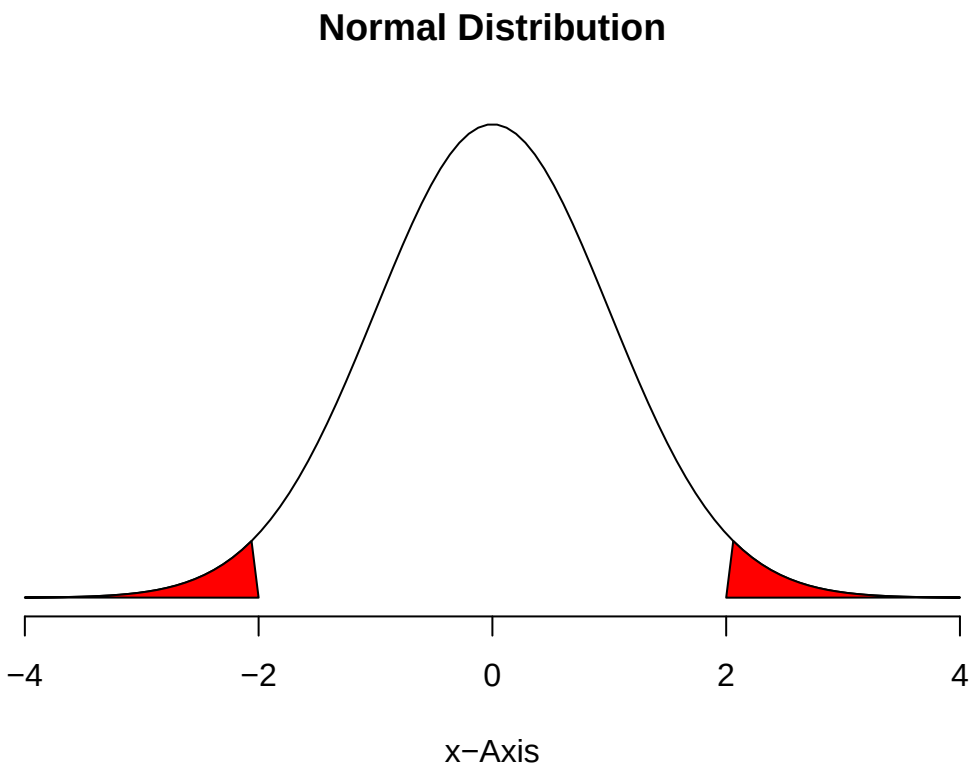


Answer for (d): roughly 4.56%

```
# use the DATA606::normalPlot function

# check the area under the curve
area_left <- pnorm(-2)
area_right <- 1 - pnorm(2)

percent <- area_left + area_right
percent
```

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

```
# plot for (c)

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



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.

The shorthand for these two normal distributions is:

- ***Men, Ages 30 - 34*:** $N(\mu = 4313, \sigma = 583)$
- ***Women, Ages 25 - 29*:** $N(\mu = 5261, \sigma = 807)$

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

The Z-score for Leo's finishing time is: 1.09

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

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

The Z-score for Mary's finishing time is: 0.312

```
(5513 - 5261) / 807
```

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

The Z-scores allow us to see how many standard deviations Leos' and Marys' times fall from the mean finish time for their respective groups. It appears that Mary's time was 0.312 standard deviations from the mean for her group and Leo's time was 1.09 standard deviations from the mean for his group.

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

Since they are both positive Z-scores, and a better performance corresponds to a faster finish, it appears that Mary ranked better in her group relative to Leo. Since Mary's Z-score was lower than Leo's, her Z-score was closer to the mean finish time for her group, and although they both finished slower than the the mean's in their respective groups, Mary's finish time was closer to her group's mean finish time than Leo's finish time was to his group's mean finish time.

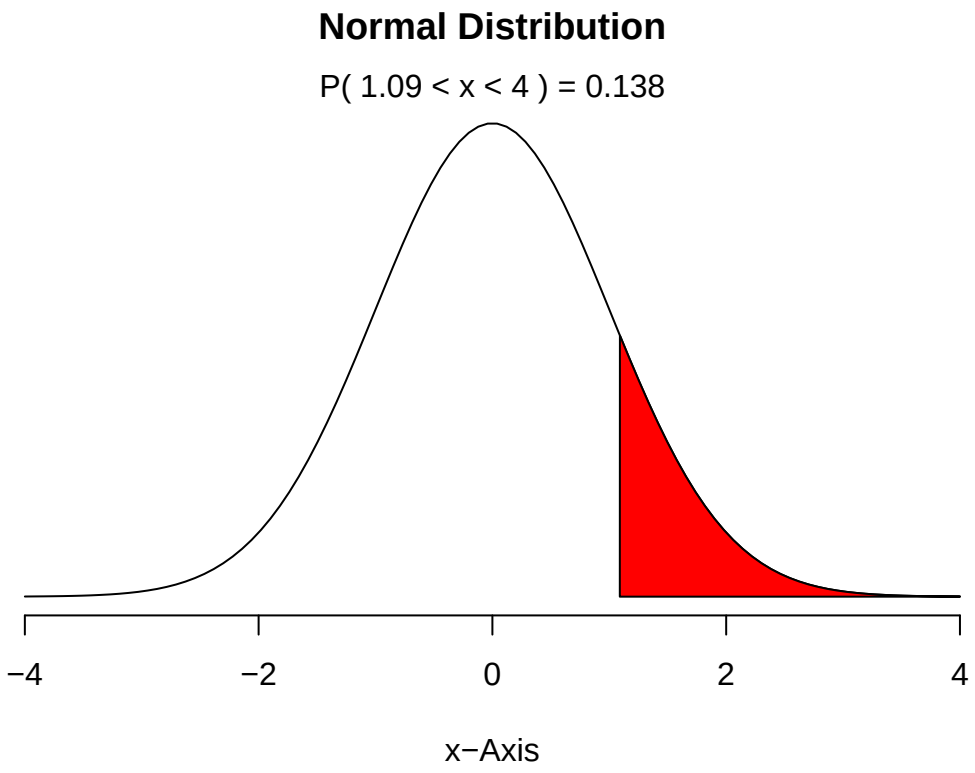
(d) What percent of the triathletes did Leo finish faster than in his group?

To find the percent of triathletes that Leo finished faster than in his group, I decided to find the area under the curve, since this is an approximately normal distribution.

```
# check the area under the curve
area_right <- 1 - pnorm(1.09)
area_right
```

```
## [1] 0.1378566
```

```
# plot for (a)
normalPlot(mean = 0, sd = 1, bounds = c(1.09, 4), tails = FALSE)
```



It looks like Leo finished faster than about 13.8% of the triatheletes in his group.

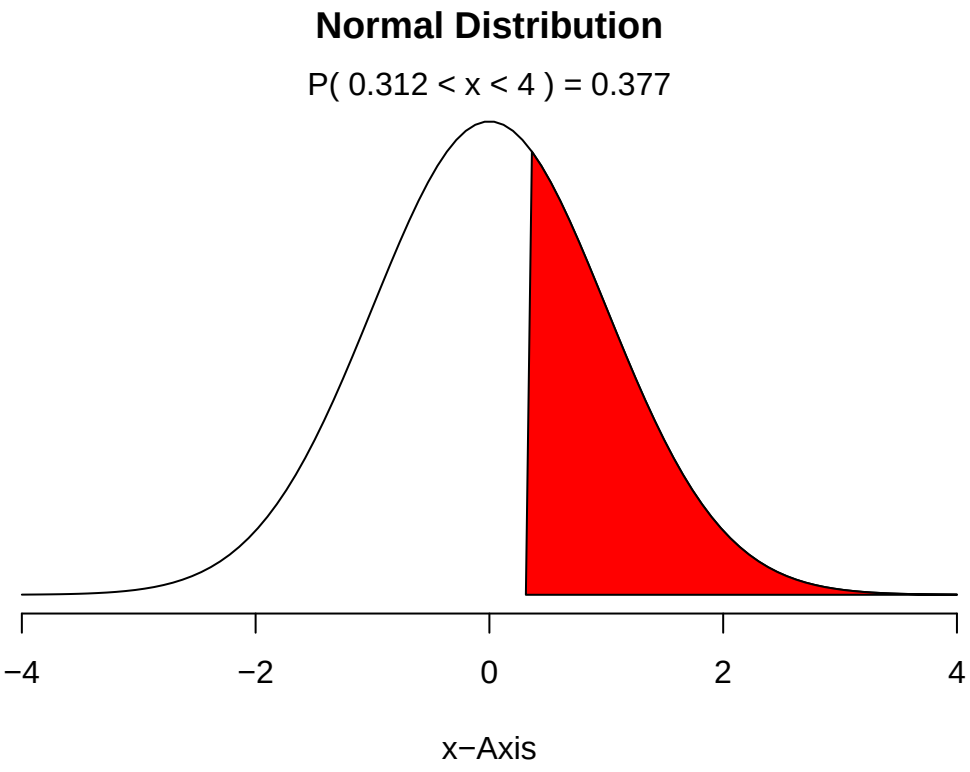
(e) What percent of the triathletes did Mary finish faster than in her group?

Similar to (d), I decided to find the area under the curve for Mary's finish:

```
# check the area under the curve
area_right <- 1 - pnorm(0.312)
area_right
```

```
## [1] 0.3775203
```

```
# plot for (a)
normalPlot(mean = 0, sd = 1, bounds = c(0.312, 4), tails = FALSE)
```



It looks like Mary finished faster than about 37.7% of the triatheletes 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.

If the distributions of finishing times are not nearly normal, then my answers to parts (b) - (e) would change. Since Z-scores are calculated based on the data from a respective group, I would not have been able to compare Leo's finish time to Mary's finish time if *both* groups were not nearly normal distributions. Additionally, the percentile groups and calculations would be different dependending on the type of distribution that would have been used to calculate the Z-scores if the distributions for finishing times was not nearly normal.

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.

```
library(openintro)
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)
# list out the heights vector
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
```

First, I thought it would be beneficial to find the mean and standard deviation of the 25 heights:

```
# find the mean height
avg_height <- mean(heights)

#find the standard deviation
std_height <- sd(heights)

avg_height
```

```
## [1] 61.52
```

```
std_height
```

```
## [1] 4.583667
```

The 25 heights appear to have $\mu = 61.25$ *inches* and $\sigma = 4.58$.

Next, I wanted to check to see if approximately 68% of the heights fall within one standard deviation from the mean: We can use this formula: $Area(\mu-\sigma) + Area(\mu+\sigma)$, since this denotes one standard deviation from the mean in both directions.

```
# Find area under the curve of +1 and -1 standard deviations from the mean
area_left <- pnorm(avg_height + std_height, mean = avg_height, sd = std_height,
                   lower = FALSE)
area_right <- 1 - pnorm(avg_height - std_height, mean = avg_height, sd = std_height,
                        lower = FALSE)

area_right
```

```
## [1] 0.1586553
```

```
area_left
```

```
## [1] 0.1586553
```

```
# We need it ensure that we incorporate both areas into the calculation
# (this is the area under the curve at both tails)
area_right + area_left
```

```
## [1] 0.3173105
```

```
# Now, to find the area between -1 sd and +1 sd, we need to subtract 1
1 - (area_right + area_left)
```

```
## [1] 0.6826895
```

We can confirm that roughly 68% of the heights fall within one standard deviation of the mean.

Next, we can check to see if 95% of observations fall within 2 standard deviations from the mean. To do this, we can use our same formula as above, except adjusted slightly to:

$Area(\mu - 2\sigma) + Area(\mu + 2\sigma)$

```
area_left <- pnorm(avg_height + 2 * std_height, mean = avg_height, sd = std_height,
              lower = FALSE)

area_right <- 1 - pnorm(avg_height - 2 * std_height, mean = avg_height, sd = std_height,
              lower = FALSE)

area_right
```

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

```
area_left
```

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

```
area_left + area_right
```

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

```
1 - (area_left + area_right)
```

```
## [1] 0.9544997
```

We can confirm that roughly 95% of the heights fall within two standard deviations of the mean.

Finally, we can check to see if 99.7% of observations fall within 3 standard deviations from the mean. To do this, we can use our same formula as above, except adjusted slightly to:

$Area(\mu - 3\sigma) + Area(\mu + 3\sigma)$

```
area_left <- pnorm(avg_height + 3 * std_height, mean = avg_height, sd = std_height,
              lower = FALSE)

area_right <- 1 - pnorm(avg_height - 3 * std_height, mean = avg_height, sd = std_height,
              lower = FALSE)

area_right
```

```
## [1] 0.001349898
```

```
area_left
```

```
## [1] 0.001349898
```

```
area_left + area_right
```

```
## [1] 0.002699796
```

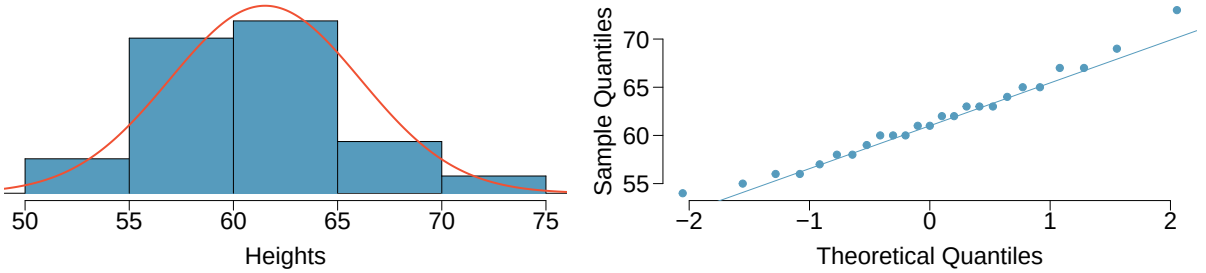
```
1 - (area_left + area_right)
```

```
## [1] 0.9973002
```

We can confirm that roughly 99.7% of the heights fall within three standard deviation of the mean.

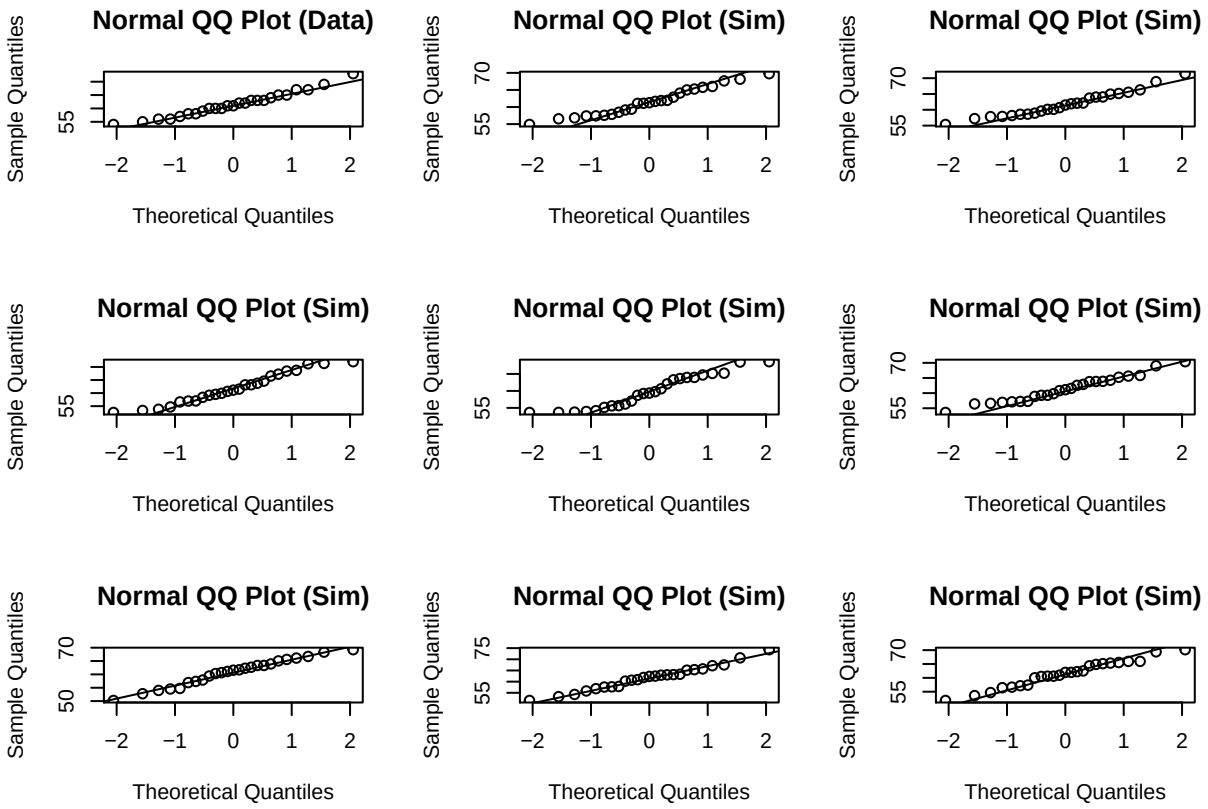
Therefore, in the end, we can confirm that this dataset of heights does approximately follow the 68-95-99.7% Rule.

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



Before deterimining whether or not this distribution appears to follow a normal distribution, it would be helpful to see if simulated data, using this same mean and standard deviation, would display a similar distribution as to the real heights data. To test this, we can use the qqnormsim function in R to run a few simulations:

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



We can see from this plot, that many of the simulations appear to have similar q-q plot distributions to the real heights data. However, the real heights data seems to follow the normal q-q plot line even more than most of the simulations, especially as values deviate farther from the mean. We can see from both the q-q plot simulations as well as the histogram and normal probability plot that the real height data seems to follow a normal distribution.

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?

By using the geometric distribution, we can calculate this problem (since they are independent events).
$(1-p)^{n-1}p$

```
p <- 0.02
p_non_defect <- (1 - 0.02)

(p_non_defect) ^ (10-1) * (p)
```

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

The probability that the 10th transistor produced is the first with a defect is roughly 1.67%.

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

```
p_non_defect ^ 100
```

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

The probability that the machine produces no defective transistors in a batch of 100 is roughly 13.2%

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

To find the average, we need to do $\mu = \frac{1}{p}$

```
1 / (p)
```

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

We can see, that on average the first transistor that will be produced with a defect will be the 50th. The standard deviation can be found by doing $\sigma = \sqrt{\frac{1-p}{p^2}}$

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

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

The standard deviation is 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?

Similar to the previous problem, we can do the following calculations:

```r
p_other_machine <- 0.05
1 / (p_other_machine)
```

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

On average we would expect that 20 transistors would be produced with this other machine before the first defect.

```r
sqrt((1 - p_other_machine) / (p_other_machine ^ 2))
```

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

The standard deviation for this other machine is 19.49.

(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?

Based on my answers to parts (c) and (d), by increasing the probability of an event occurring, we will see the mean and standard deviation of the wait time until success 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.

We can use the following binomial equation to calculate the probability:

$\binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$

```
# full calculation
((3*2*1) / (2*1) * (1)) * ((0.51) ^ 2) * ((1 - 0.51) ^ (3 - 2))
```

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

```
# check
p <- 0.51
dbinom(2, 3, p)
```

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

The probability of having two boys is 0.38.

(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.

```
# scenarios
s1 <- 0.51 * 0.51 * 0.49
s2 <- 0.51 * 0.49 * 0.51
s3 <- 0.49 * 0.51 * 0.51

# additional rule for disjoint outcomes
s1 + s2 + s3
```

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

This confirms that answers (a) and (b) match. The probability is roughly 0.38.

(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).

If we wanted to calculate the probability that a couple who plans to have 8 kids will have 3 boys, part (b) would be much more tedious than the approach from part (a) because there would be many more permutations ($\binom{8}{3} = 56$ permutations) that we'd have to write out and add together to calculate the probability. Therefore, we would save a lot of time being able to utilize the formula from part (a) to make this calculation.

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?

We would use the negative binomial distribution to answer this question, since the last trial must be a success. We can use the following formula:

$\binom{n-1}{k-1}p^k(1-p)^{n-k}$

```
sequences <- (factorial(10-1) / (factorial(3-1) * (factorial(10-3))))
sequences * (0.15 ^ 3) * ((1 - 0.15) ^ (10 - 3))
```

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

The probability that on the 10th try she will make her 3rd successful serve is about 3.90%.

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

Since each attempt is independent from one another, the probability of each attempt is also independent from one another. Therefore, each attempt has a probability of success of 0.15 - this means that the probability that her 10th serve will be successful is also 0.15.

(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?

The reason why the probabilities are different between (a) and (b) is due to the type of information we would like to find. For (b), we are not concerned about the probabilities of the previous serves since we only would like to know the probability that the 10th serve is successful - since all attempts are independent from one another, the probability of previous attempts does not have any affect on the probability of the tenth serve. However, for (a), we would like to know the number of successes (k) out of the number of attempts (n), this follows the negative binomial distribution, where we would like to account for successes and attempts.