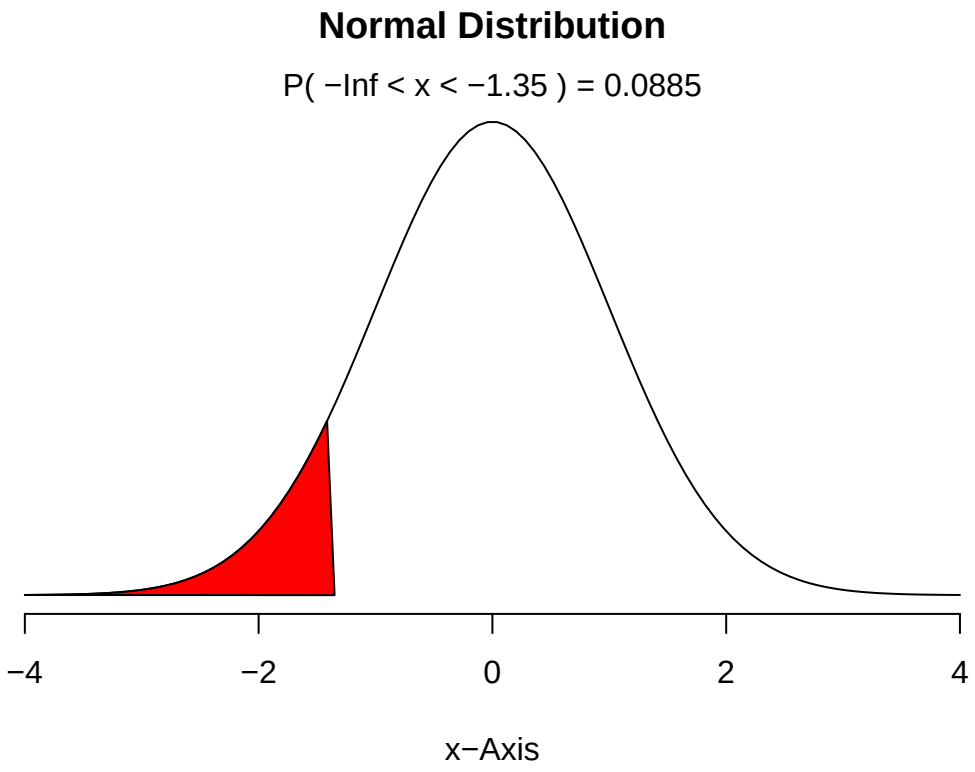
Chapter 4 - Distributions of Random Variables

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) $Z < -1.35$

The percentage of a standard normal distribution found under this lower section of the curve is 8.85%.

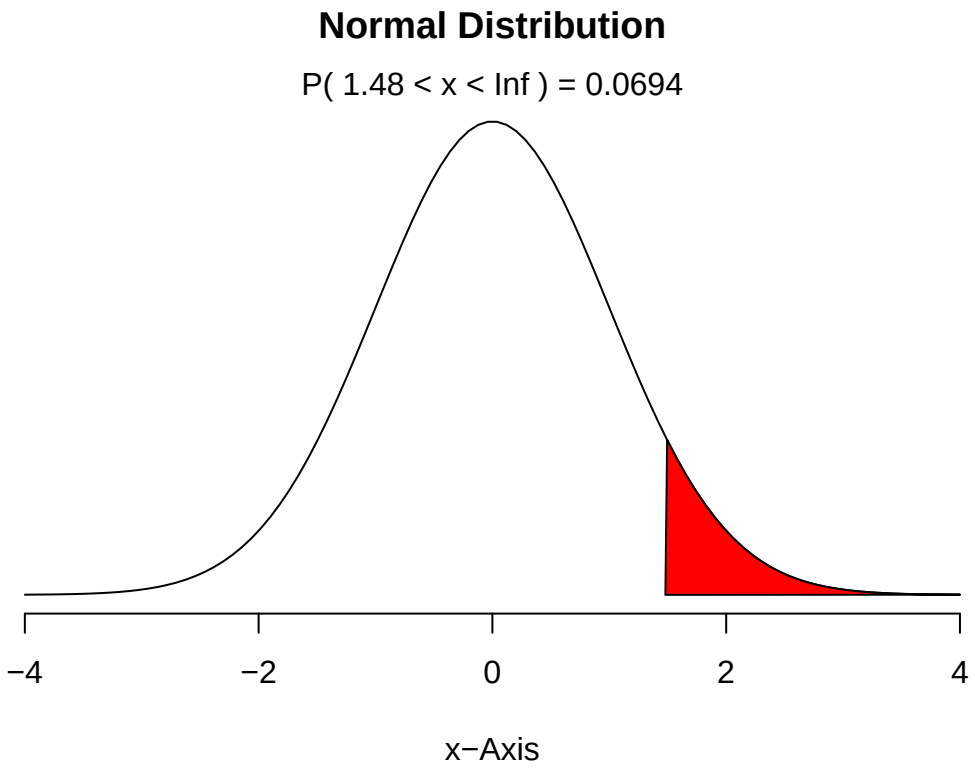
```
# trying out two methods
# library(visualize)
# visualize.norm(stat=-1.35,mu=0,sd=1,section="lower")
# normalPlot(mean =, sd = 1, bounds = c(-1,1), tails = FALSE)
normalPlot(mean = 0, sd = 1, bounds = c(-Inf, -1.35), tails = FALSE)
```



(b) $Z > 1.48$

The percentage of a standard normal distribution found under this upper, highest section of the curve is 6.94%.

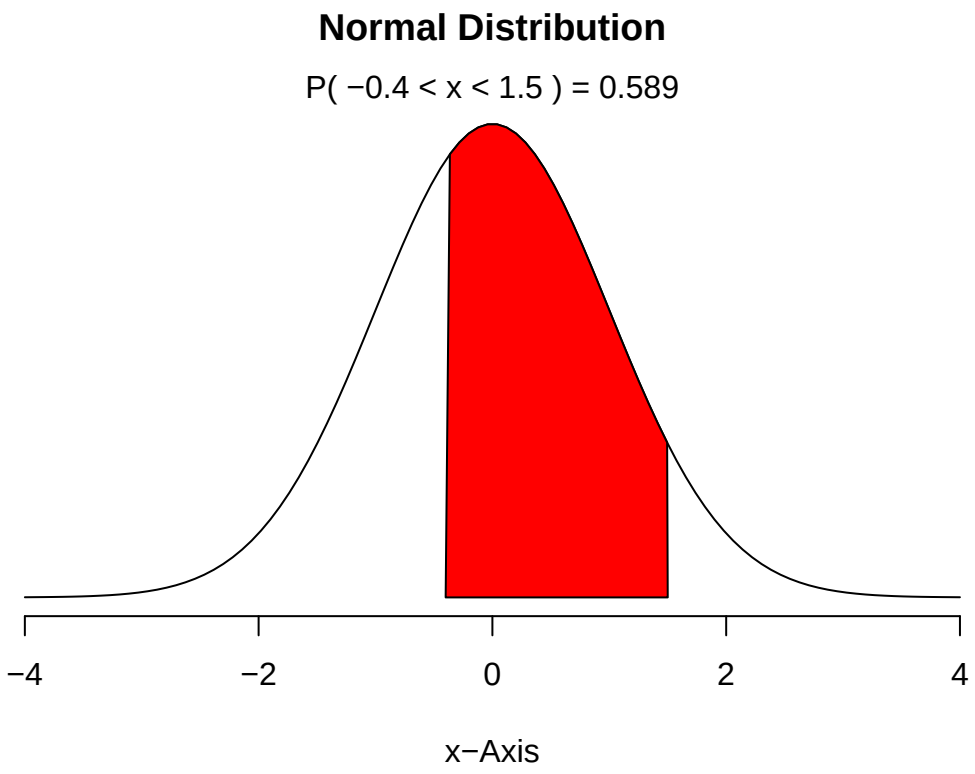
```
# visualize.norm(stat=1.48,mu=0,sd=1,section="upper")
normalPlot(mean = 0, sd = 1, bounds = c(1.48, Inf), tails = FALSE)
```



(c) $-0.4 < Z < 1.5$

The percentage of a standard normal distribution found under this middle right section of the curve is 58.9%.

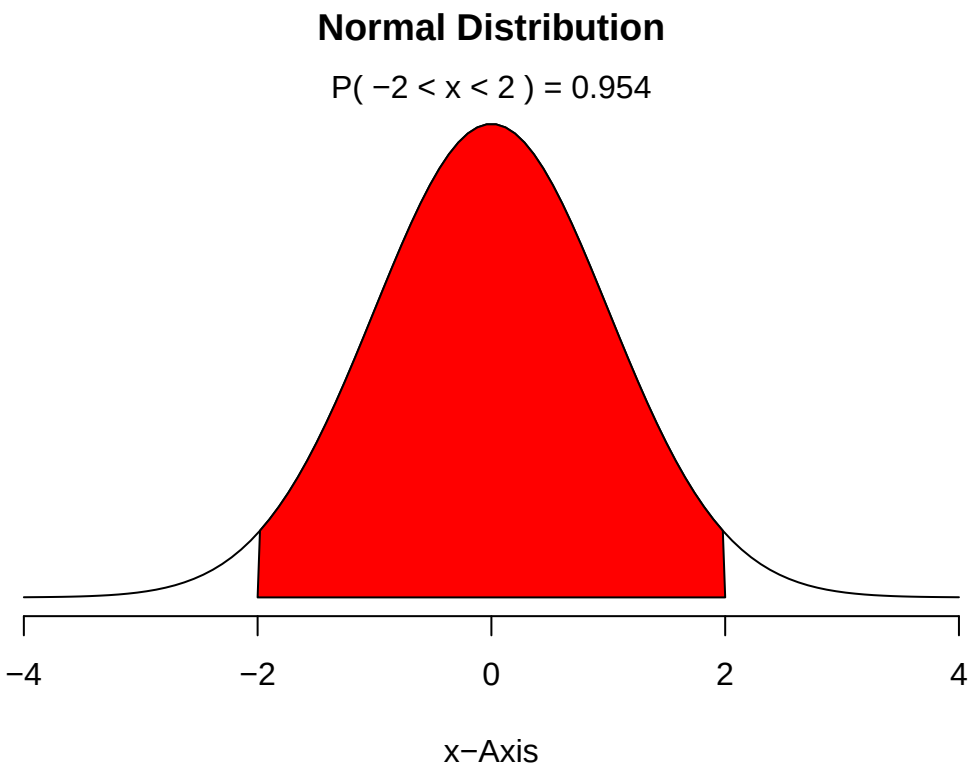
```
# visualize.norm(stat=c(-0.4,1.5),mu=0,sd=1,section="bounded")
normalPlot(mean = 0, sd = 1, bounds = c(-0.4,1.5), tails = FALSE)
```



(d) $|Z| > 2$

The percentage of a standard normal distribution found under this middle section of the curve is 95.4%.

```
# visualize.norm(stat=c(-2,2),mu=0,sd=1,section="bounded")
normalPlot(mean = 0, sd = 1, bounds = c(-2,2), tails = FALSE)
```


Normal Distribution
P(−2 < x < 2) = 0.954
−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.

```
# Shorthand for normal distribution of finishing times of women aged 25 - 29
fem_mean <- 5261
fem_sd <- 807

# Shorthand for normal distribution of finishing times of men aged 30 - 34
male_mean <- 4313
male_sd <- 583
```

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

Z-scores describe how unusual a particular variable is by calculating the number of standard deviations it is from the mean. It's formula is:

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

In our case, x is the finishing time of the individual, μ is the mean of all triathletes in the group, and σ is the standard deviation of all triathletes in the group.

Given Leo's finishing time at 4948 seconds we can calculate his score as:

```
Leo_time <- 4948
# Equivalent to (4948 - 4313) / 583
Leo_zscore <- (Leo_time - male_mean)/male_sd
Leo_zscore
```

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

The z-score for Leo is 1.089.

Given Mary's finishing time at 5513 seconds we can calculate her score as:

```
Mary_time <- 5513
# Equivalent to (5513 - 5261) / 807
Mary_zscore <- (Mary_time - fem_mean)/fem_sd
Mary_zscore
```

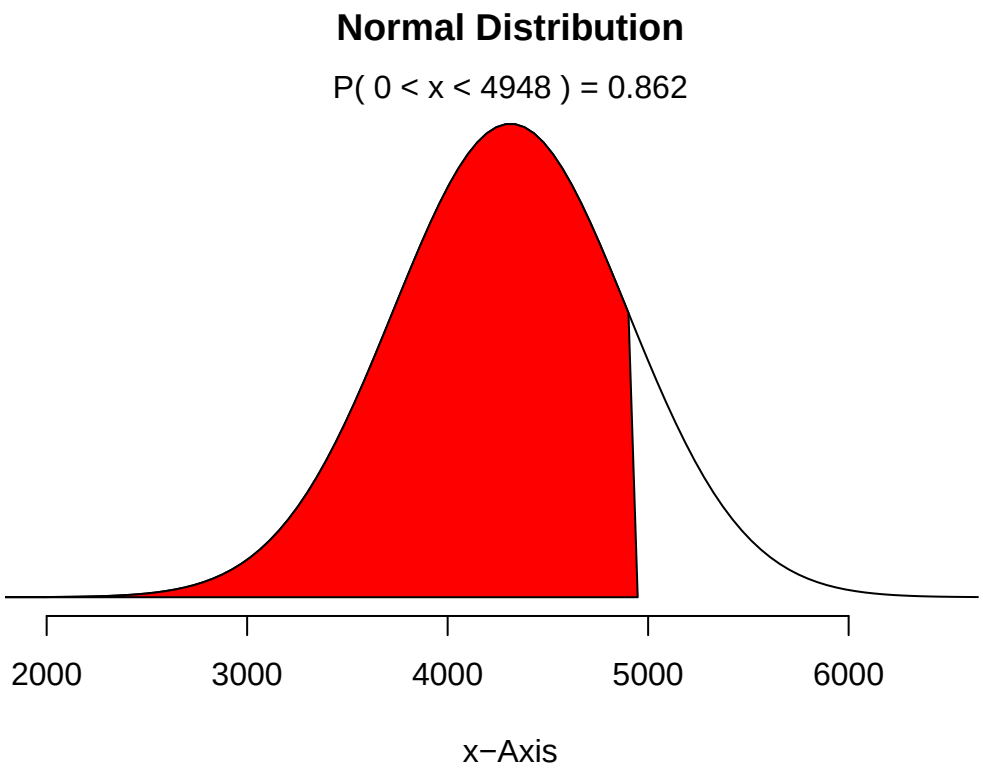
```
## [1] 0.3122677
```

The z-score for Mary is 0.3123.

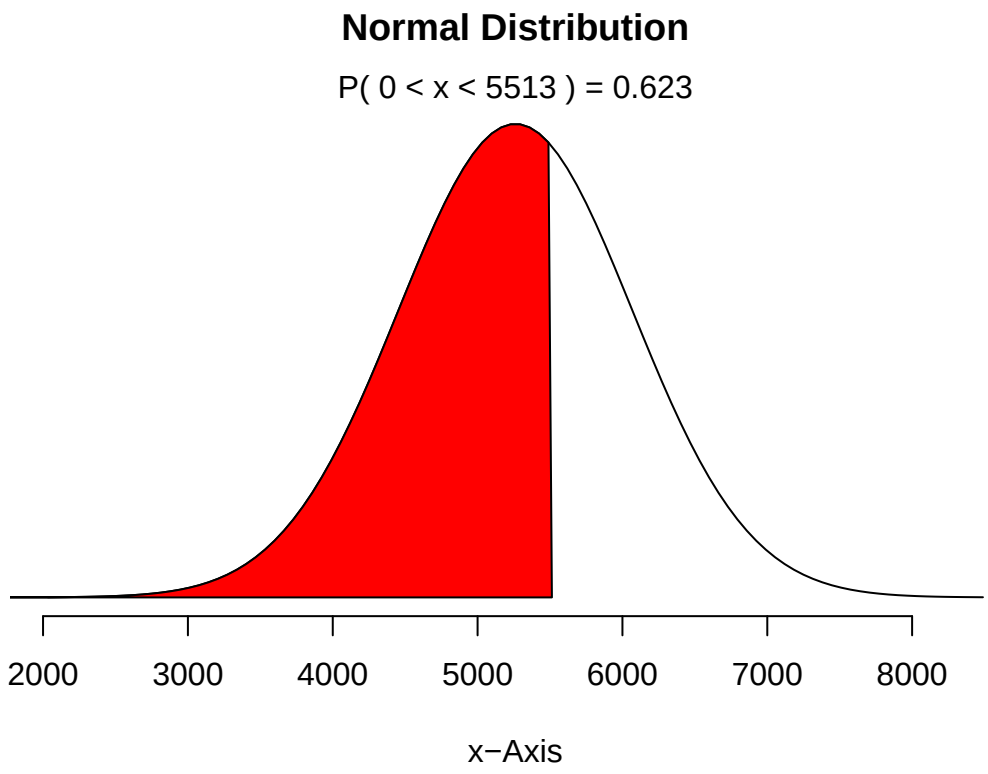
(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

By comparing the percentages of indiviudals above and below their finishing times in each group we can decide who ranks better in their respective groups.

```
# Distribution of male finishing times with Leo's finishing time selected
normalPlot(mean = male_mean, sd = male_sd, bounds = c(0, 4948))
```




```
# Distribution of female finishing times with Mary's finishing time selected
normalPlot(mean = fem_mean, sd = fem_sd, bounds = c(0, 5513))
```



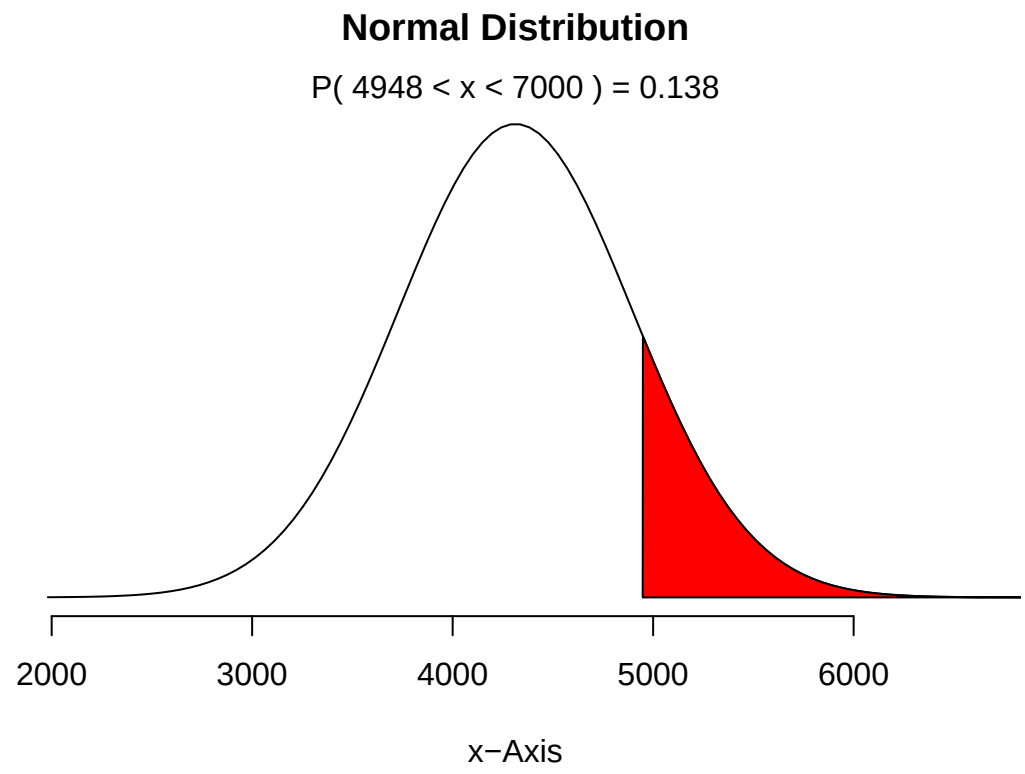
As shown in the first distribution, Leo finished with .826, or 86.2% of triathletes finishing ahead of him (because a lower time indicates earlier finishing time). While Mary finished with only 62.3% of triathletes in front of her. For this reason, Mary did better because she finished ahead of more people within her group.

We can also compare their z-scores to see this relationship too. Leo's score of 1.089 is higher than Mary's at 0.3123. Leo's higher score means there is a lower probability that triatheletes finished after his time of 4948 seconds. Alternatively, Mary's lower score means there is a higher probability that more triathletes finished after her time of 5513. Mary's score also shows that she is much closer to the mean than Leo, which in this case, further confirms that Mary's finishing time was better in her group than Leo's in his group.

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

Leo finished faster than about 13.8% of the triathletes in his group.

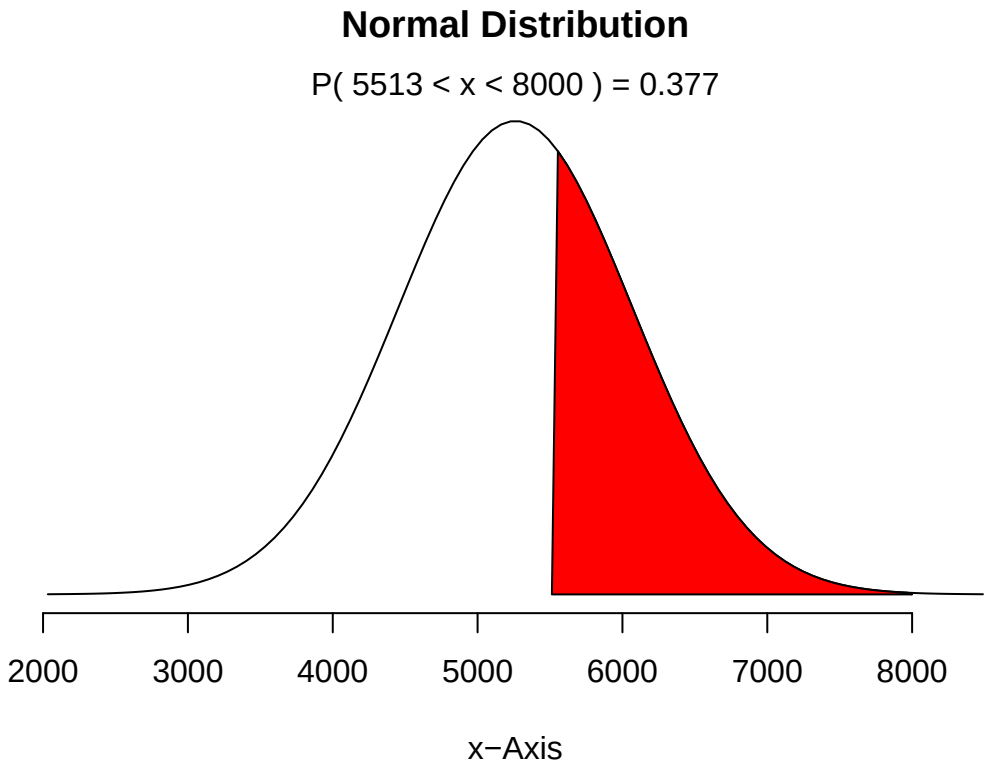
```
# Distribution of male finishing times with those after Leo
normalPlot(mean = male_mean, sd = male_sd, bounds = c(4948, 7000))
```



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

Mary finished faster than about 37.7% of the triathletes in her group.

```
# Distribution of female finishing times with those afer Mary's finish
normalPlot(mean = fem_mean, sd = fem_sd, bounds = c(5513, 8000))
```



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

Yes, if the distributions are not nearly normal, the answers would change due to the data being skewed in some way from a symmetric orientation outward from the mean. In each of these scenarios, we are able to find the probability of individuals finishing below and above the mean because the data is a nearly normal distribution.

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.

```
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,
mean_heights <- mean(heights)
sd_heights <- sd(heights)
# 68% check
1-2*pnorm(mean_heights+sd_heights, mean = mean_heights, sd = sd_heights, lower = FALSE)
```

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

```
# 95% check
1-2*pnorm(mean_heights+2*sd_heights, mean = mean_heights, sd = sd_heights, lower = FALSE)
```

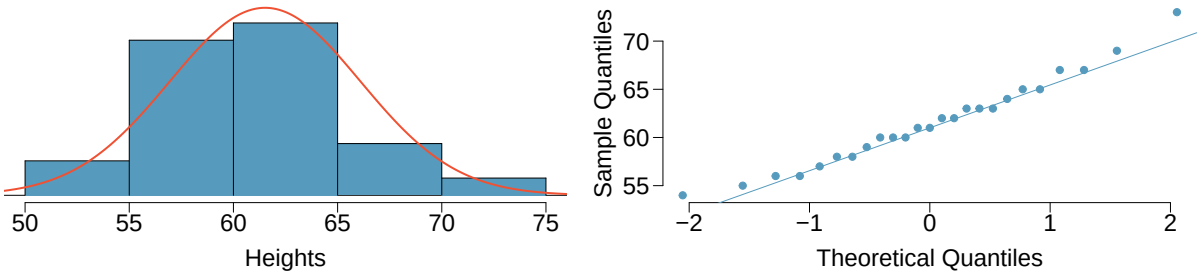
```
## [1] 0.9544997
```

```
# 99.7% check
1-2*pnorm(mean_heights+3*sd_heights, mean = mean_heights, sd = sd_heights, lower = FALSE)
```

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

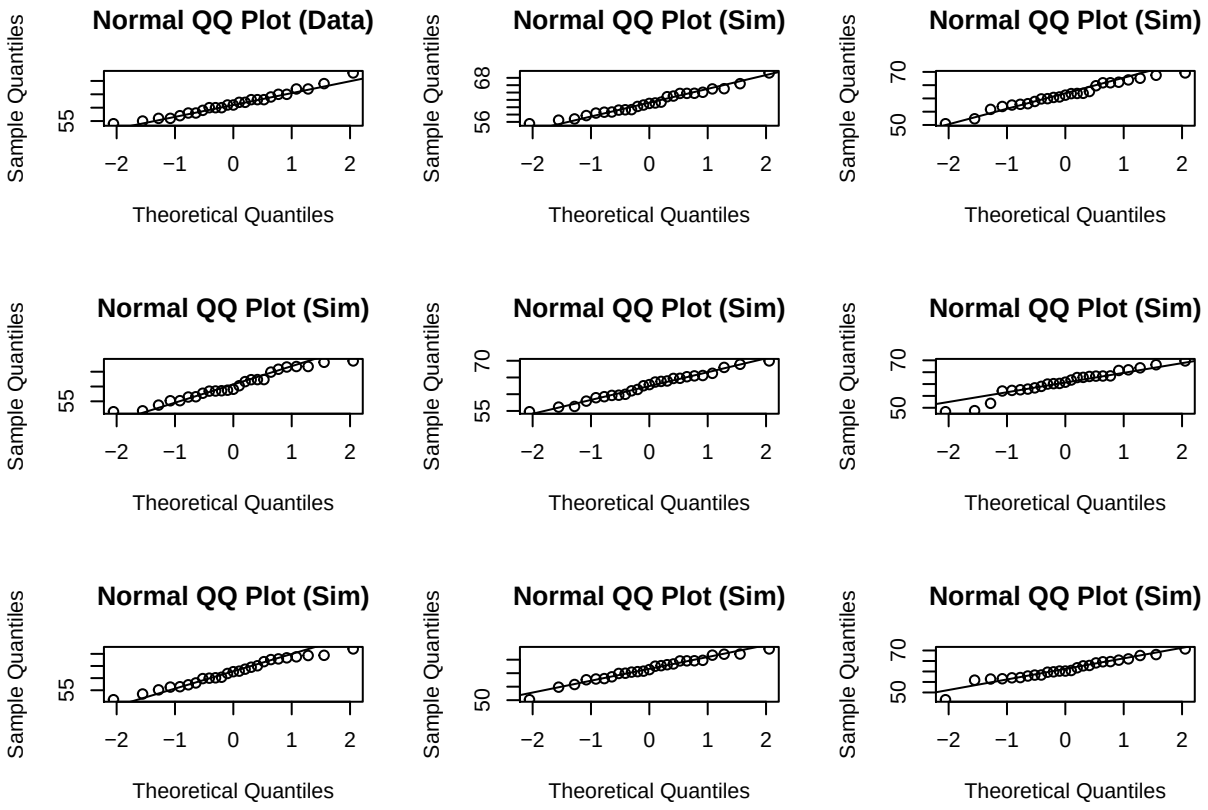
It appears the heights do follow the 68 - 95 - 99.7% rule.

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



For review, what would the simulations of data with the same mean and standard deviation look like compared to the origional heights data?

```
qqnormsim(heights)
```



Yes, this data appears to follow a normal distribution.Though some simulated plots fall outside of a completely symmetric normal distribution, when comparing the real data from heights with generated data we see that by far most of the points fall on the normal qqline indicating a relatively 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?

Using the equation for a geometric distribution

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

where p is the defective rate and number of transistors is represented by n,

```
(1 - 0.02)^(10-1) * 0.02
```

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

it is estimated that the probability that the 10th transistor produced will have a 0.0167 probability of being defective or 1.67% chance.

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

Given the defective rate we can find the success rate by subtracting the percentage defective from 100.

```
# probability of success converted to decimal form
success <- (100-2)/100
success
```

```
## [1] 0.98
```

With the success rate, we can find the number defective in the batch since each production is independent of the next.

```
success^100
```

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

There is about a 13.3% chance, or probability of 0.133, that the machine will produce 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?

It would take about 50 transistors before one was produced with the first defect.

```
# maintaining the rate as a decimal for calculations
1/0.02
```

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

The standard deviation is about 49.497.

```r
sqrt((1-0.02)/0.02^2)
```

```
## [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 before first defect
1/0.05
```

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

```r
# Standard deviation
sqrt((1-0.05)/0.05^2)
```

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

The average is about 20 transistors before the first one is defective. The standard deviation is 19.494.

(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?

When the probability of success is high the mean and standard deviation of the event occurring earlier is also higher. You wait less when the probability of success is higher.

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.

```
# (Total outcomes! / Two boys outcome!)*P boys^2 * P girls
(factorial(3)/factorial(2))*0.51^2*.49
```

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

The probability that two of them will be boys is 0.382.

(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.

The total possibilities of having 3 children with 2 of them as boys is:

- MMF
- FMM
- MFM

With the probability of having boys remaining at .51 and girls at .49 the new calculation using disjoint outcomes and the addition rule produces this:

```
# two boys (through some combination) with one girl
(.51*.51*.49)*3
```

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

Where the probabilities using this method and the binomial model are the same at 0.382347.

(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).

As the number of events increases the number of possibilities and values input to calculate a specific event's occurrence increases too. Computing larger values would be more tedious with the addition rule because all possibilities must be considered independently as events, then combined, when computing the likelihood that an event will occur.

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?

Using the formula:

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

Where k = 3, n = 10, and p is 0.15 we can calculate the probability

```
# n = 10
# k = 3
A <- factorial(9)/(factorial(2)*factorial(7))
B <- 0.15^3
C <- 0.85^7
A*B*C
```

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

The probability that on the 10th try she will make her 3rd successful serve is 0.03895, or about 3.89%.

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

With nine attempts, two successful serves, and the probability of success the same at 0.15 we plug into the following equation.

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

```
# n = 10 (because it is her 10th serve)
# k = 3 (because this chance is considered a success in this probability)
A <- factorial(10)/(factorial(3)*factorial(7))
B <- 0.15^3
C <- 0.85^7
A*B*C
```

```
## [1] 0.1298337
```

the probability that her 10th serve will be successful is 0.1298 or nearly 12.98%.

(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?

Looking at the formulas, there is no n-1 or k-1 term in the binomial distribution equation. Thus, there is no assumption that the last trial is a success. For this reason, the probabilities should be different. When you start with one of the possible combinations as a success (as in the negative binomials) you eliminate that as a variable outcome.

Binomial

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

Negative Binomial

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

In addition to the terms of the formulas, there is a discrepancy in each equation's intent. For a binomial, the number of trials are set and the calculation is looking for the probability of successful events. On the other hand, negative binomials are looking for the number of trials while assuming the final event is a success.