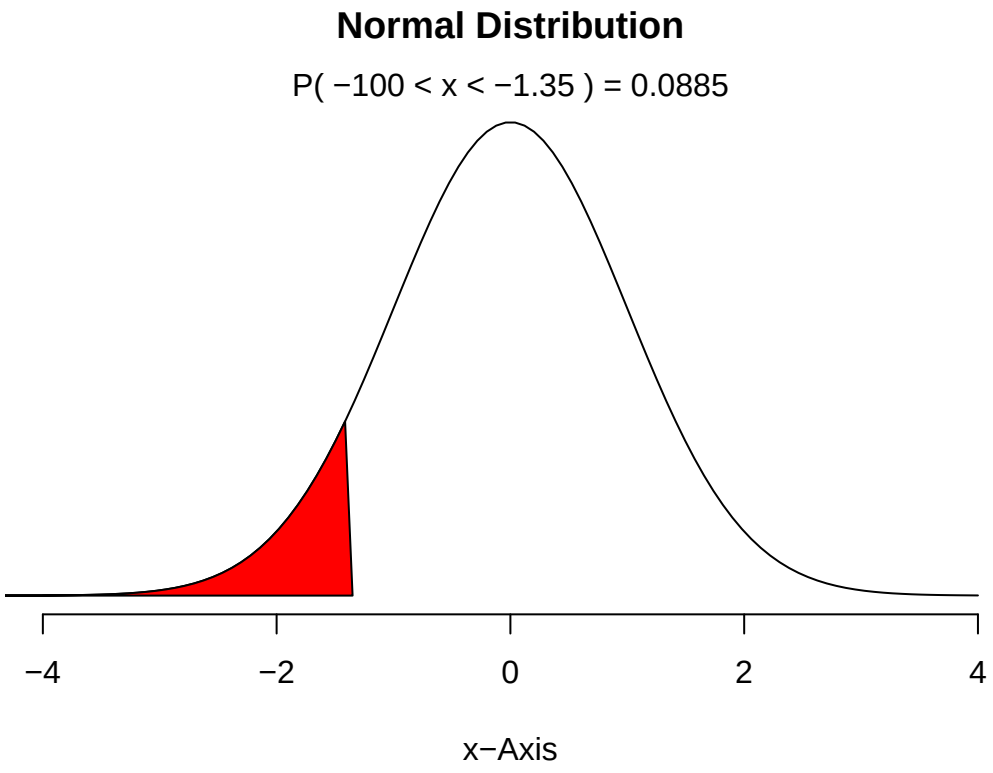
Chapter 4 - Distributions of Random Variables

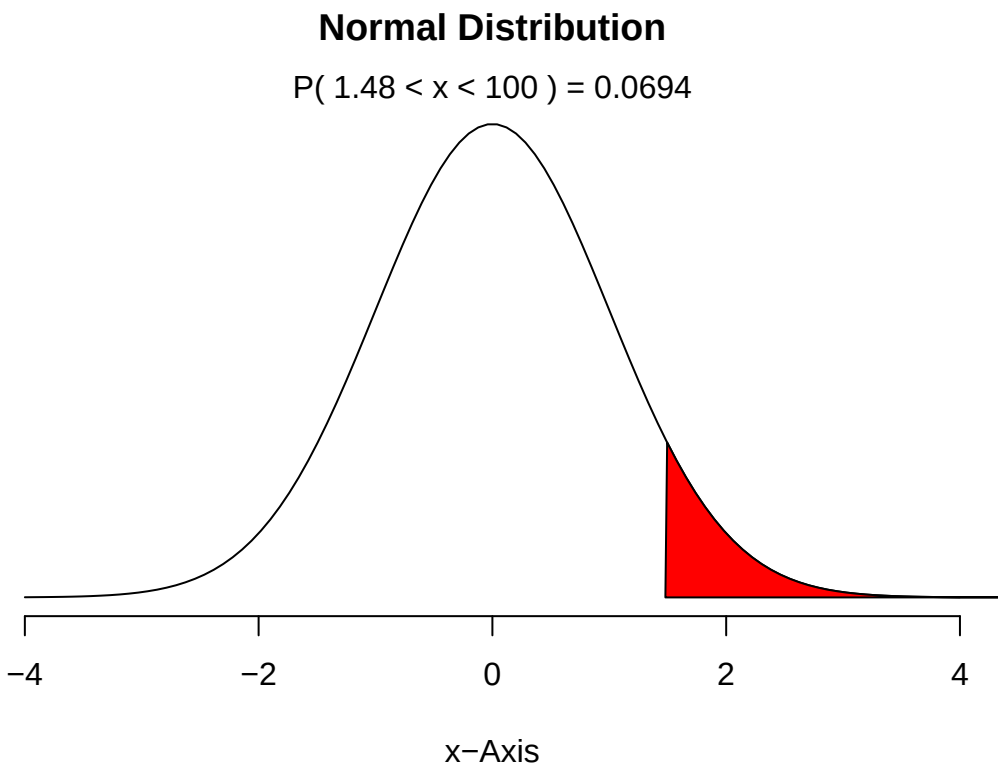
Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

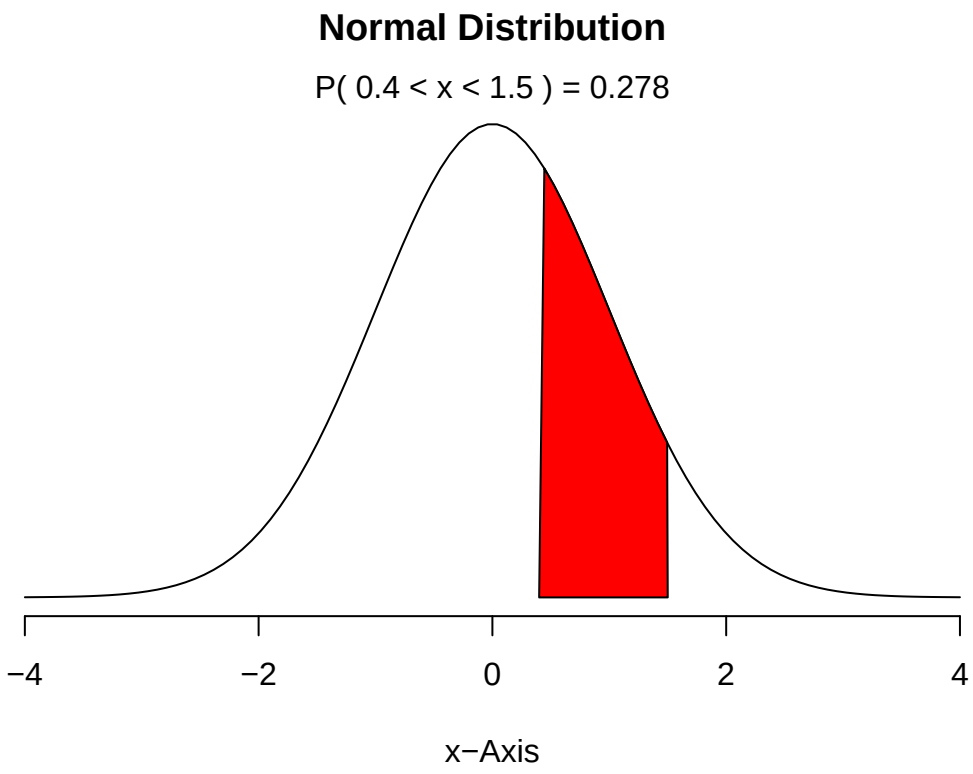
(a) $Z < -1.35$ 8.85%
(b) $Z > 1.48$ 6.94%
(c) $-0.4 < Z < 1.5$ 27.8%
(d) $|Z| > 2$ .0228*2=.0456 4.56%

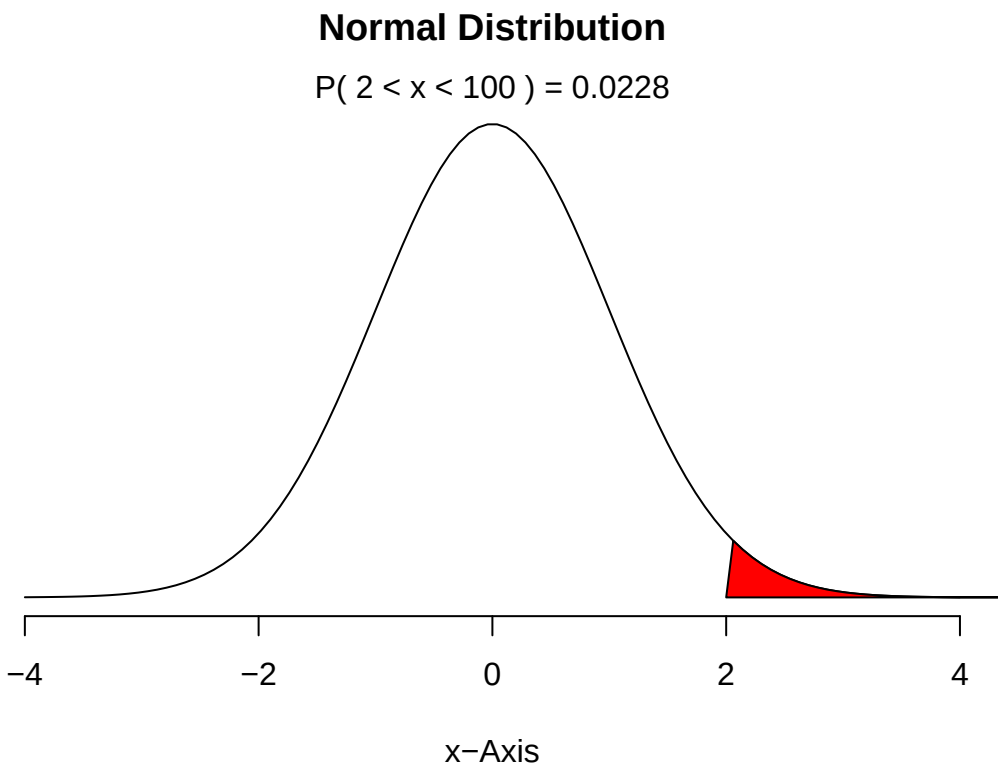
```
## Loading required package: shiny

## 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
```




Normal Distribution
P(1.48 < x < 100) = 0.0694
−4
−2
0
2
4
x−Axis


Normal Distribution
P(0.4 < x < 1.5) = 0.278
−4
−2
0
2
4
x−Axis


Normal Distribution
P(2 < x < 100) = 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.

Men 30-34: (M=4313,sd = 583) Women 30-34: (M=5261,sd = 807)

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

```
## [1] 1.087479
```

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

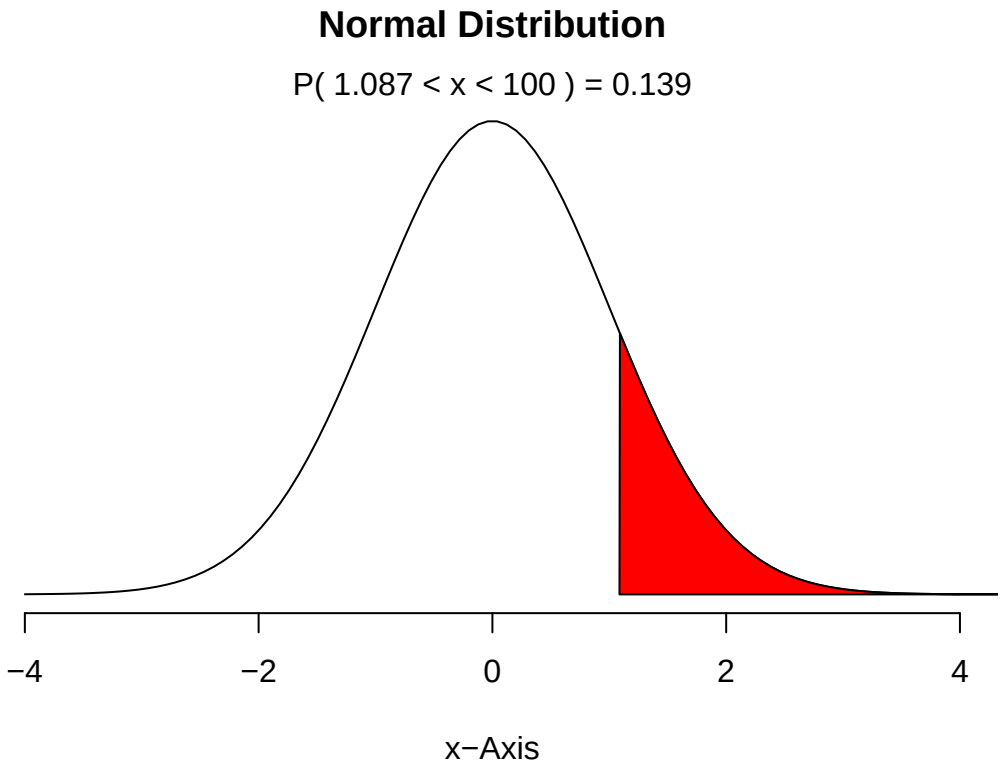
Leo Z-score = 1.087 Mary z-score = 0.312

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

Mary ranked better in her group. They were both above the average time but Mary's z-score was closer to zero which means she was closer to average than Leo was.

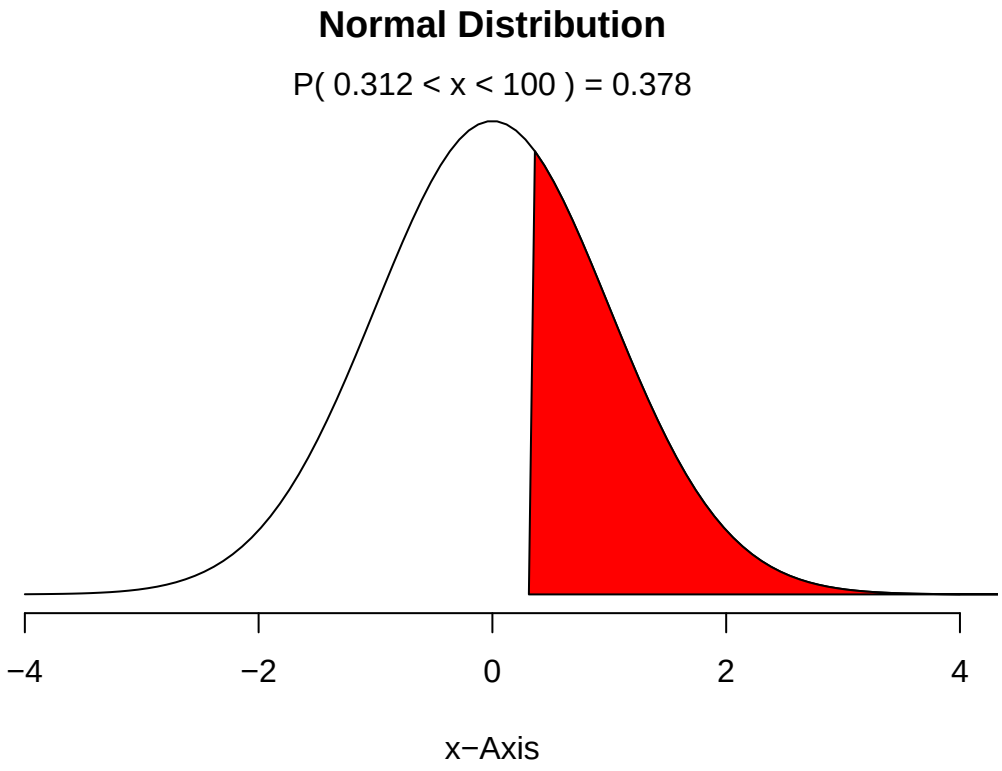
(d) What percent of the triathletes did Leo finish faster than in his group?

13.9%



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

37.8%



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

Yes the normal distribution allows for the use of z-scores and the NormalPlot function. Normal distributions make it easier to analyze and make claims about a sample. A non-normal distribution would call for a different method.

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

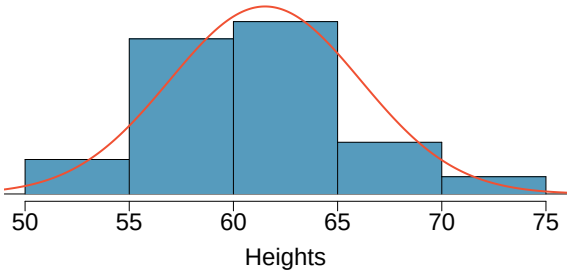
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

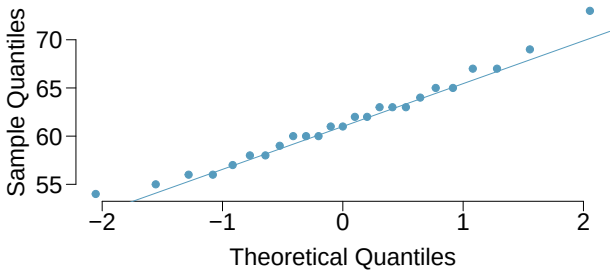
(a) The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule.

The heights approximately follow the 68-95-99.7% Rule but they are not extremely close. Under this rule the one standard deviation above the mean is 66.1 and the value for this dataset that is higher than 68% of the population is 63. 2 sd above the mean is 70.68, 95% is 68.6. 3 sd above the mean is 75.26,99.7% is 72.71.

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

Using the qqnormism function we see that the data doesn't appear to follow a relatively normal distribution. We can see this by viewing the data points and seeing that they don't deviate greatly form the the line which indicates a normal distribution. However, it isn't extremely close to a normal distribution.






```
# Use the DATA606::qqnormsim function
quantile(heights,.68)
```

```
## 68%
##  63
```

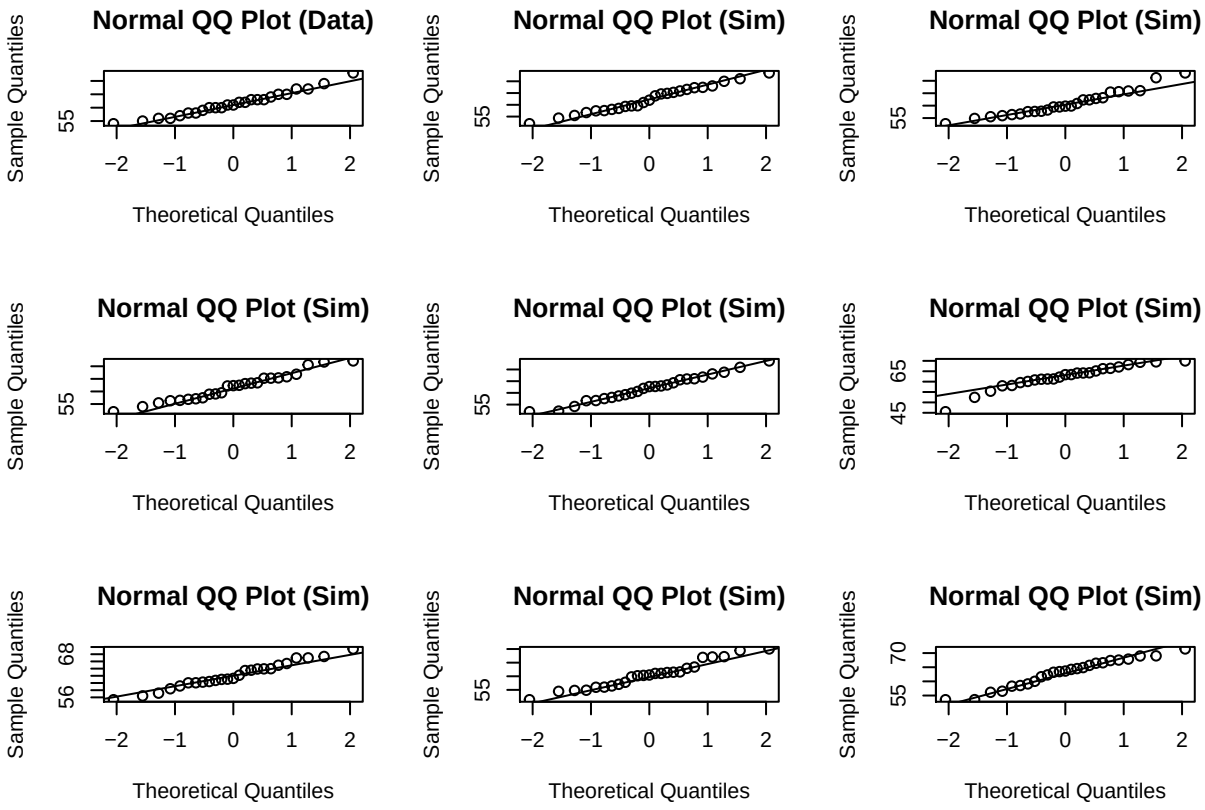
```
quantile(heights,.95)
```

```
##  95%
## 68.6
```

```
quantile(heights,.997)
```

```
##  99.7%
## 72.712
```

```
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
45
65
56
68
70

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?

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

.0167 or 1.67%

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

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

0.133 or 13.3%

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

We would expect to see 50 transistors produced before a defect (1/.02 = 50). standard deviation = 49.497

```
## [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?

We would expect to see 50 transistors produced before a defect (1/.05 = 20). standard deviation = 19.494

```
## [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?

Increasing the probability of an event causes the mean and the sd of the wait until success to decrease. This makes sense as the more likly an event the sooner it will happen and the less variability in the frequency of its occurence.

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.

0.127 or 12.7%

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

0.382 or 38.2%

(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.

P(A=Boy,B=Boy,C=Girl) P(A=Boy,B=Girl,C=Boy) P(A=Girl,B=Boy,C=Boy)

0.382 or 38.2%

yes they match

```
## [1] 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).

The approach from part a can be altered very easily when it is already written. In this example you would just need to change n to 8 and k to 3 and run the same program. The approach form part B would take much longer as you would have to rewrite the different scenarios which would take time the larger the sample gets.

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?

```
## [1] 0.007197358
```

0.00866 or 0.866%

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

The probability is 0.15 or 15% because the success of each serve is independent of the other attempts.

(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 difference is due to the scenario in (a) taking place before she has made any serves. Therefore you need to calcualte the odds she will successfully serve 2/9 and the multiply that by the odds that she makes any one serve. In the second scenario she has already successfully served 2 out of 9 times. So the probability of that is 1. Therefore the answer is 1*0.15 = 0.15.