CUNY SPS DATA606 HW4 - Distributions of Random Variables

Chinedu Onyeka

September 20th, 2021

Getting Started

```
#Load Required Libraries
library(tidyverse)
library(fastGraph)
```

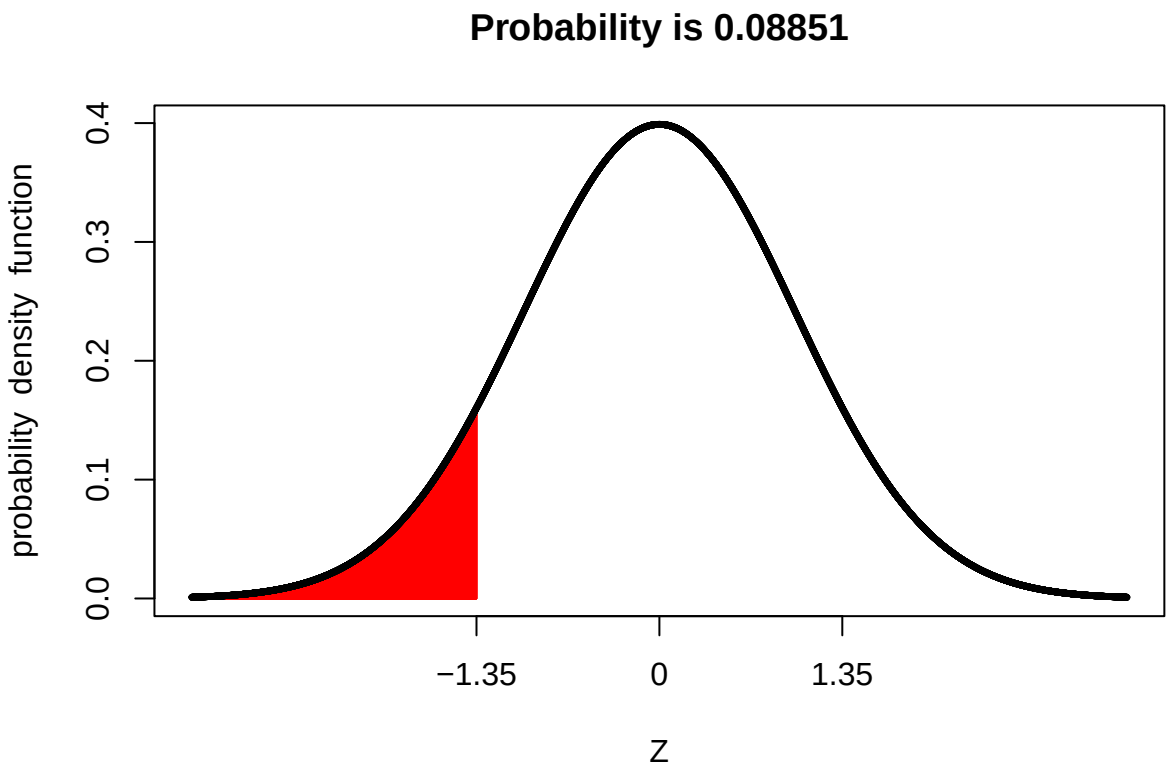
Problem 1:

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) $|Z| > 2$

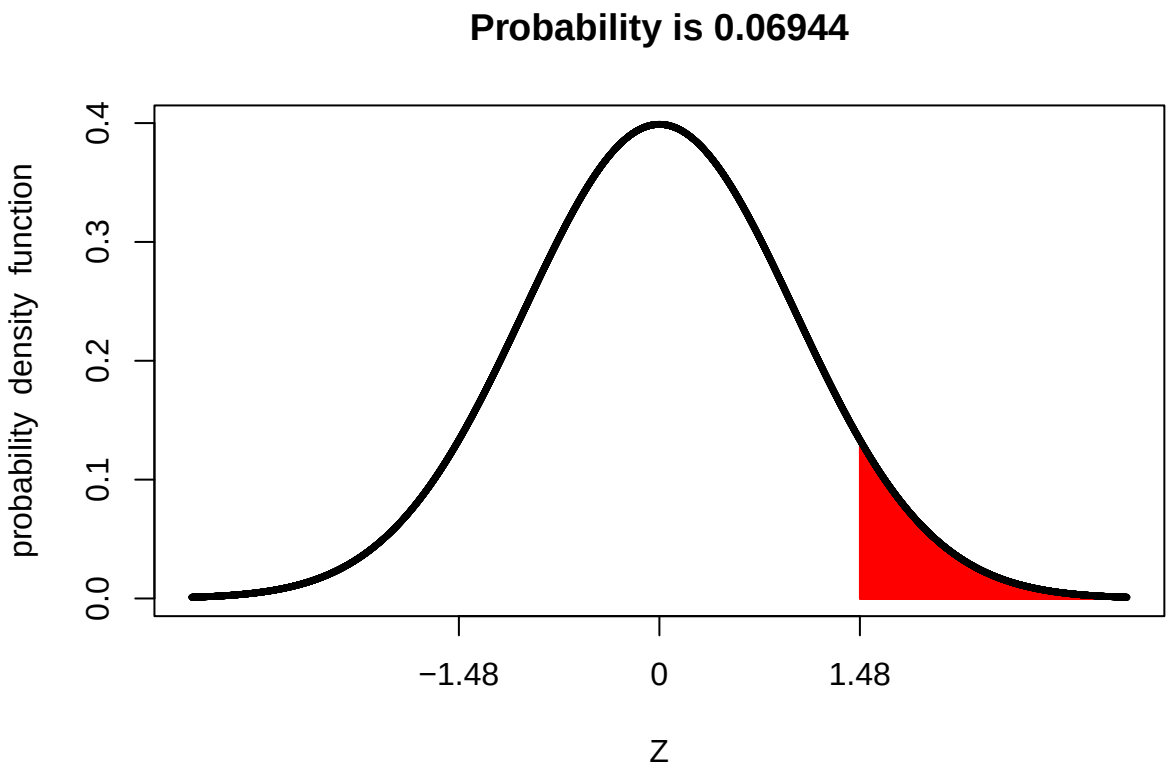
Solution 1: 1(a)
$Z < -1.35$
Ans: 8.85%

```
shadeDist(-1.35)
```



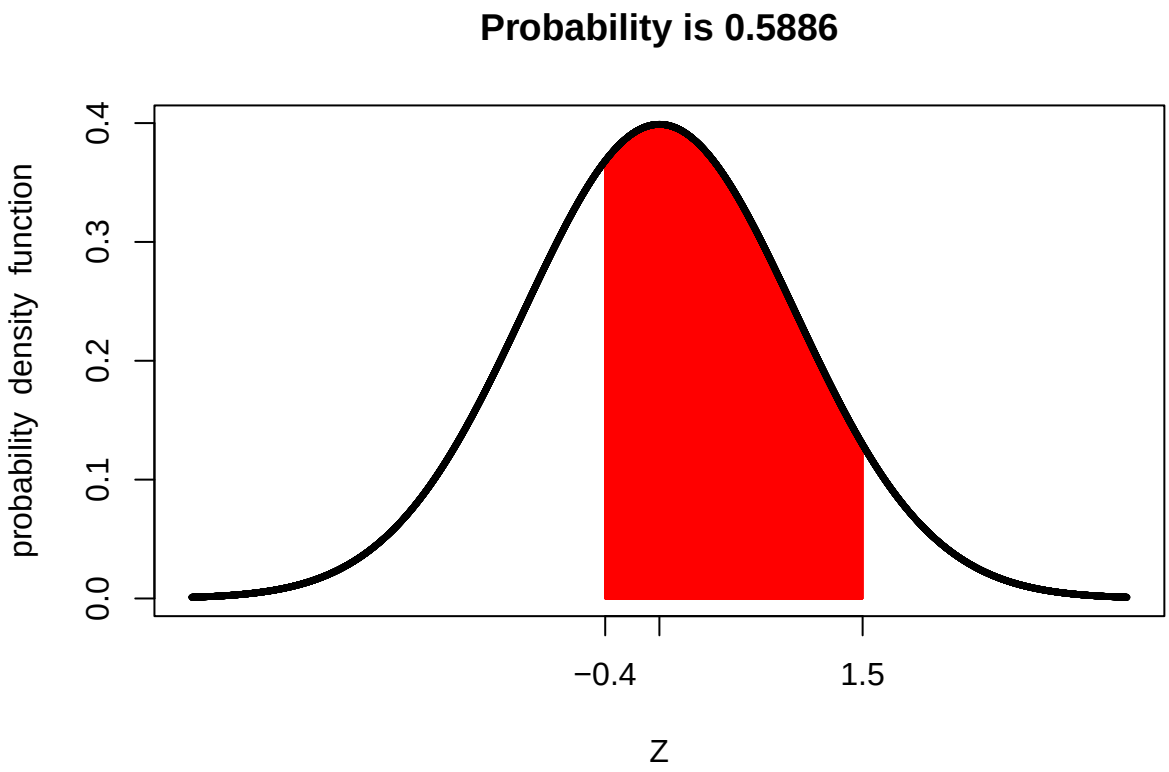
1(b)
$Z > 1.48$
Ans: 6.94%

```
shadeDist(1.48, lower.tail = FALSE)
```



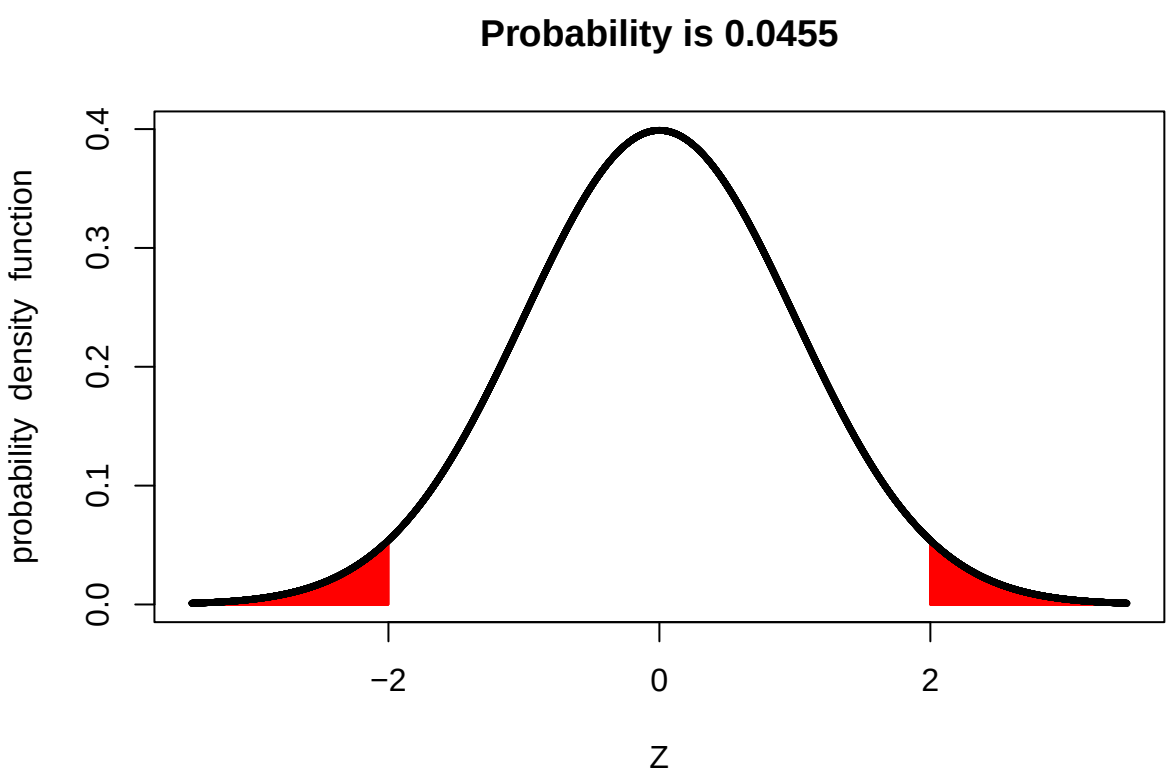
1(c)
$-0.4 < Z < 1.5$
Ans: 58.86%

```
shadeDist(c(-0.4, 1.5), lower.tail = FALSE)
```



1(d)
$|Z| > 2$
Ans: 4.55%

```
shadeDist(c(-2,2))
```


Probability is 0.0455
probability density function
0.0
0.1
0.2
0.3
0.4
−2
0
2
Z

Problem 2:

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.

Solution 2: 2(a)
Men, Ages 30 - 34; N($\mu = 4313$, $\sigma = 583$)
Women, Ages 25 - 29; N($\mu = 5261$, $\sigma = 807$)

2(b)
z - scores: z = $\frac{x-\mu}{\sigma}$
Leo: x = 4948 ; z = $\frac{4948-4313}{583} = 1.0892$
Mary: x = 5513; z = $\frac{5513-5261}{807} = 0.3123$

2(c)
Leo: P(Leo) = 0.8620 = 86.20%
Mary: P(Mary) = 0.6226 = 62.26%
Leo did better than Mary in their respective group since he is in the 86.2 percentile of his age group while Mary is in the 62.26 percentile of her age group.

2(d)
Leo finished faster than 86.20% of triathletes in his group.

2(e)
Mary finished faster than 62.26% of triathletes in her group.

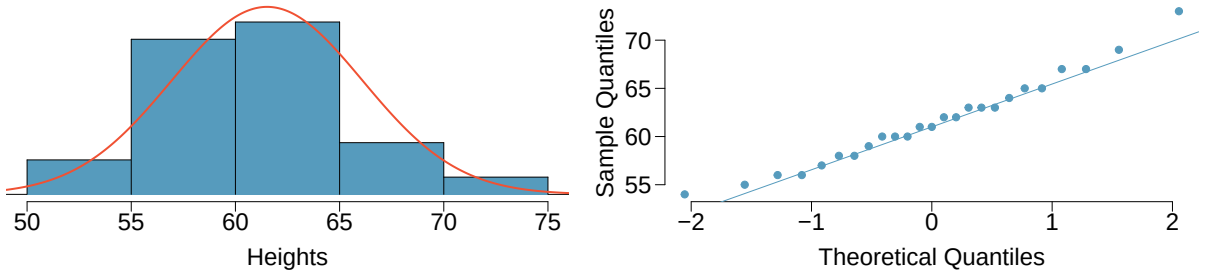
2(f)
If the distributions of finishing times are not nearly normal, my answers would be different because the assumptions I used in getting their percentile is that their finishing times are nearly normal which means that the values would be wrong if that is not the case.

Problem 3:

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




```
# Use the DATA606::qqnormsim function
```

Solution 3: 3(a)

```
female_heights <- c(54, 55, 56, 56, 57, 58, 58, 59, 60, 60, 60, 61, 61, 62, 62, 63, 63, 63, 64, 65, 65,
mean_female_heights <- mean(female_heights)
sd_female_heights <- sd(female_heights)
mu <- mean_female_heights
sd <- sd_female_heights
paste0("The mean female height is ", mu)
```

```
## [1] "The mean female height is 61.52"
```

```
paste0("The standard deviation of female heights is ", round(sd, 2))
```

```
## [1] "The standard deviation of female heights is 4.58"
```

To determine if it follows the 68-95-99.7% Rule, we check each of the categories:

68% rule: 68 percent of the data should fall within 1 standard deviation of the mean.

```
prop_68 <- length(female_heights[female_heights < (mu + sd) & female_heights > (mu - sd)]) / length(fema

prop_68*100
```

```
## [1] 68
```

We see that the total count of female heights that fall within 1 standard deviation of the mean is 68%

95% rule: 95 percent of the data should fall within 2 standard deviations of the mean.

```
prop_95 <- length(female_heights[female_heights < (mu + 2*sd) & female_heights > (mu - 2*sd)]) / length

prop_95*100
```

```
## [1] 96
```

This result shows that 96% of female heights fall within 2 standard deviations of the mean.

99.7% rule: 99.7 percent of the data should fall within 3 standard deviations of the mean.

```
prop_99.7 <- length(female_heights[female_heights < (mu + 3*sd) & female_heights > (mu - 3*sd)]) / leng

round(prop_99.7*100, 2)
```

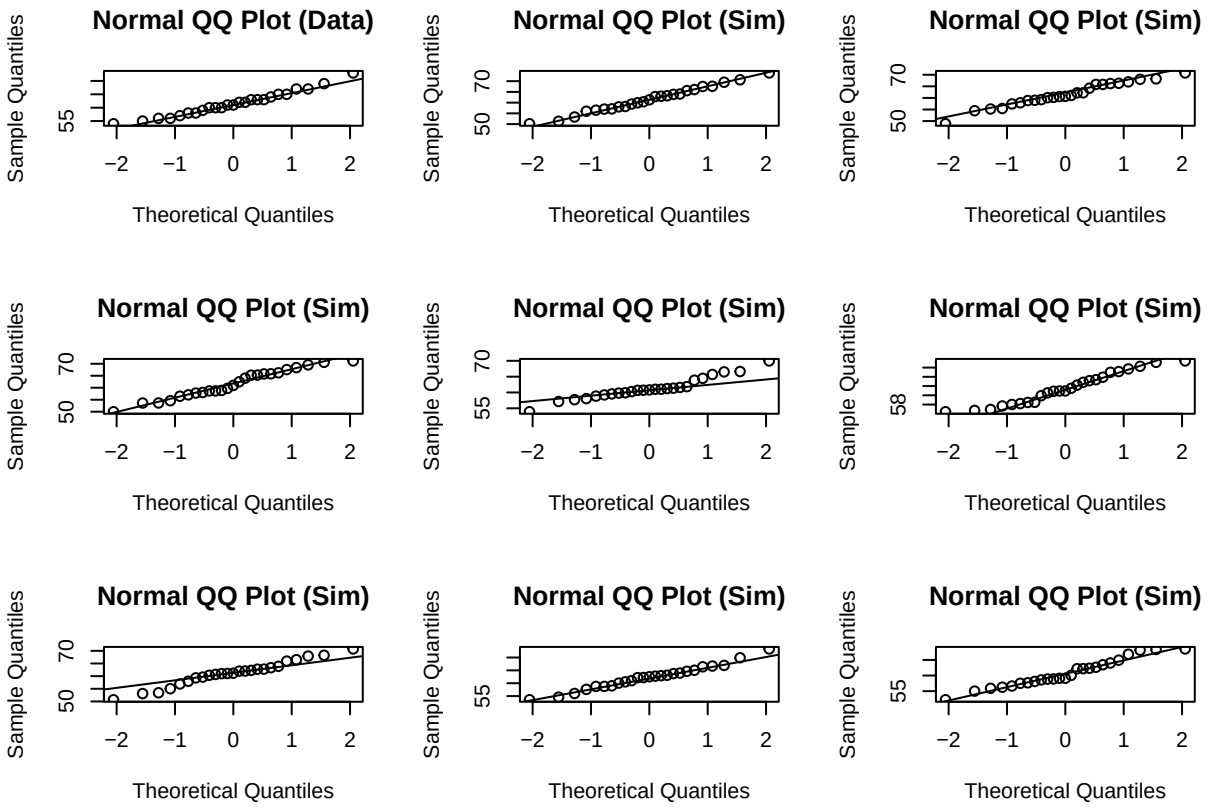
```
## [1] 100
```

In this case, 100% of the data is within 3 standard deviations of the mean which satisfies the requirement that 99.7% of the data fall within 3 standard deviations.

3(b)
Since the data satisfies the 68-95-99.7% rule, and from the histogram provided and theoretical quantiles curve provided, I can conclude that the given distribution is nearly normal and follows a normal distribution.

Using the qqnormsim function, we get the graphs below which further supports that the given female heights follows a normal distribution.

```
qqnormsim(female_heights)
```



Problem 4:

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 4: Transistor has 2% defective rate. p = 2% = 0.02

4(a)
Probability that the 10^{th} transistor produced is the first with a defect.
P(10^{th} transistor is the first defective) = $(1-0.02)^{10-1} * 0.02 = 0.016675 = 0.0167$
The probability that the 10^{th} transistor produced is the first with a defect is 1.67%

4(b)
Probability of no defects in a batch of 100 = $(1-0.02)^{100} = 0.1326$
The probability of no defects in a batch of 100 transistors is 13.26%

4(c)
On average we would expect $\frac{1}{0.02} = 50$ transistors before the first with a defect.
$\mu = \frac{1}{p} = \frac{1}{0.02} = 50$
$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.02}{0.02^2}} = 49.4975 = 49.5$ The mean is 50 and the standard deviation is 49.5

4(d)
$\mu = \frac{1}{p} = \frac{1}{0.05} = 20$
$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.05}{0.05^2}} = 19.4936 = 19.5$

4(e)
As the probability increases, the mean or expected value decreases as they have a clear inverse relationship. Also, the standard deviation decreases as the probability increases.

Problem 5:

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 5: p = probability of having a boy = 0.51

5(a)
P(k = 2, n = 3)

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

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

The probability of having 2 boys of 3 children is 38.23%

5(b)
All possible outcomes of 3 children, 2 of whom are boys are: 1st = boy, 2nd = boy, 3rd = girl; 1st = boy, 2nd = girl, 3rd = boy; and 1st = girl, 2nd = boy, 3rd = boy.
P(2 boys of 3 children) = P(1st = boy, 2nd = boy, 3rd = girl) + P(1st = boy, 2nd = girl, 3rd = boy) + P(1st = girl, 2nd = boy, 3rd = boy)
P(2 boys of 3 children) = (0.51)(0.51)(0.49) + (0.51)(0.49)(0.51) + (0.49)(0.51)(0.51)
= 0.127449 + 0.127449 + 0.127449 = 0.382367 = 0.3824
Therefore, the probability of having 2 boys of 3 children is 38.24%

It can be seen that the answers for part(a) and part(b) both match.

5(c)
This approach will be tedious as we would need to find the possible arrangements of 3 boys in 8 children (${}_{3}^{8}\mathrm{C} = 56$). This means that there are 56 ways this could happen and we would need to find all 56 ways of these events which is not an easy task. Using the probability formula, the probability of having 3 boys out of 8 children is 21%

```
p_3boys <- dbinom(3, 8, 0.51)
p_3boys
```

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

Problem 6:

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?

Solution 6: probability of making a successful serve, p = 0.15
This is a negative binomial distribution.

6(a)
probability of 3^{rd} successful serve on 10^{th} trial, k = 3, n = 10
P(k = 3, n = 10, p = 0.15) = $\binom{n-1}{k-1}(1-p)^{n-k}p^k = \binom{10-1}{3-1}(1-0.15)^{10-3}p^3 = 0.0281$
The probability of 3^{rd} successful serve on 10^{th} trial is about 2.81%

6(b)
The probability that her 10^{th} serve will be successful is 15%. Since each trial is independent of the other and the probability of success is the same for each of the trials, the probability of the serve being successful on the 10^{th} trial is still the same 15% irrespective of the previous outcomes.

6(c)
There is discrepancy in the answers for parts (a) and (b) as can be seen from the answers. Part (a) considers the 10^{th} trial to be the 3ˆrd successful trials out of 10 trials which means that there have been 2 successful trials in the previous 9 trials before the 10^{th} trial while part (b) considers just the 10^{th} trial which is independent of the previous 9 trials.