DS606-HW4-GC

George Cruz

9/28/2020

Area under the curve part I. (4.1 p. 142)

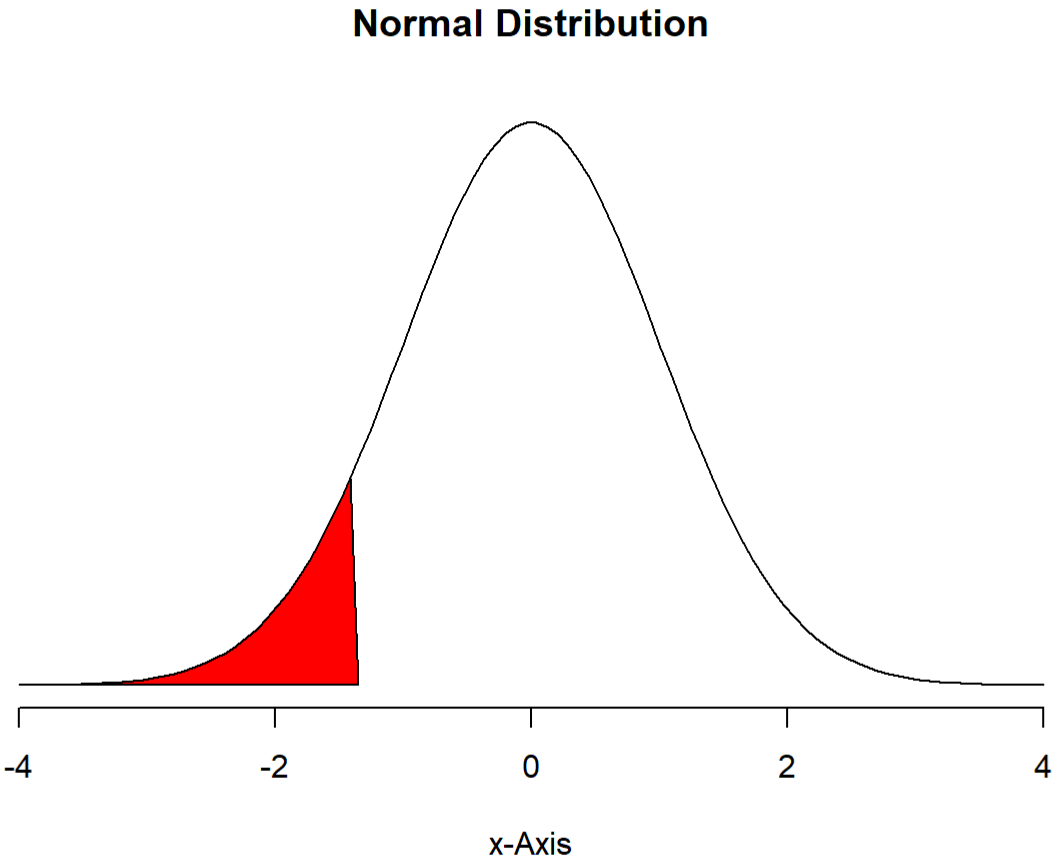
Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

a. $Z < -1.35$

```
scales::percent(pnorm(-1.35), accuracy = 0.01)
```

```
## [1] "8.85%"
```

```
normalPlot(mean=0, sd = 1, bounds = c(-1.35, Inf), tails = TRUE)
```

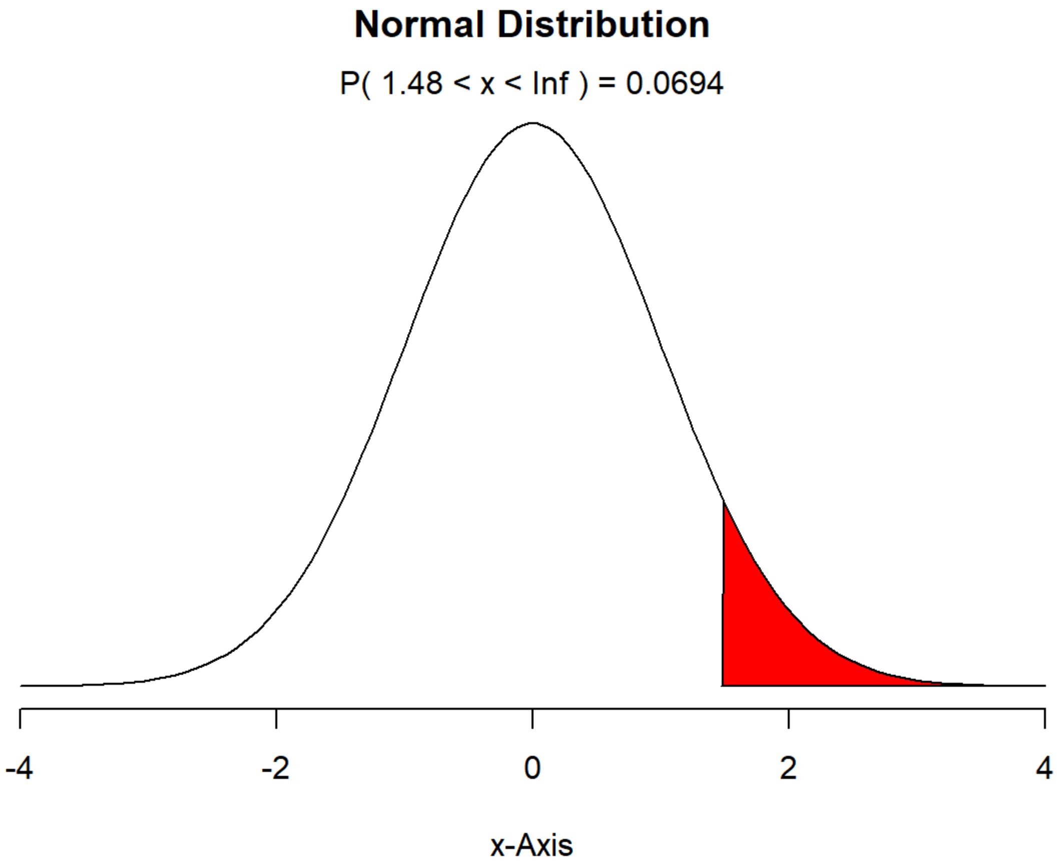


b. $Z > 1.48$

```
scales::percent(1-pnorm(1.48), accuracy = 0.01)
```

```
## [1] "6.94%"
```

```
normalPlot(mean=0, sd = 1, bounds = c(1.48, Inf), tails = FALSE)
```



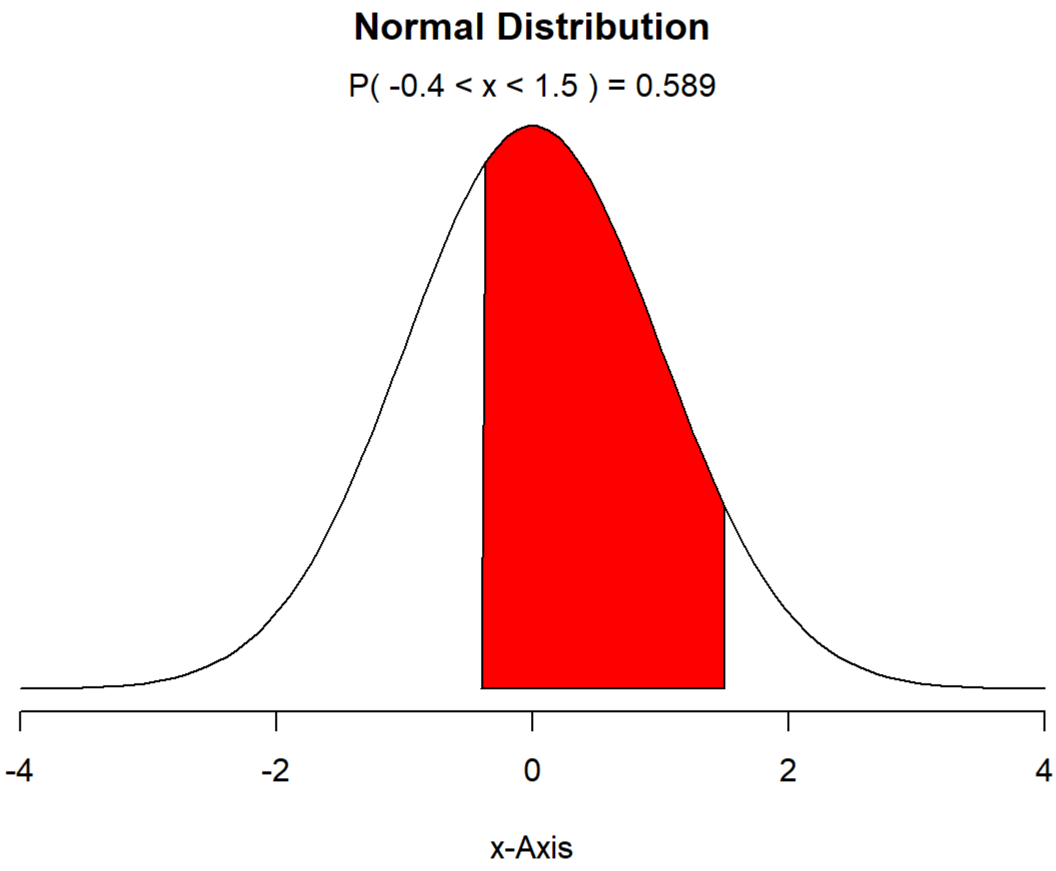
c. $-0.4 < Z < 1.5$

```
higher = pnorm(1.5)
lower = pnorm(-0.4)

scales::percent(higher-lower, accuracy = 0.01)
```

```
## [1] "58.86%"
```

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



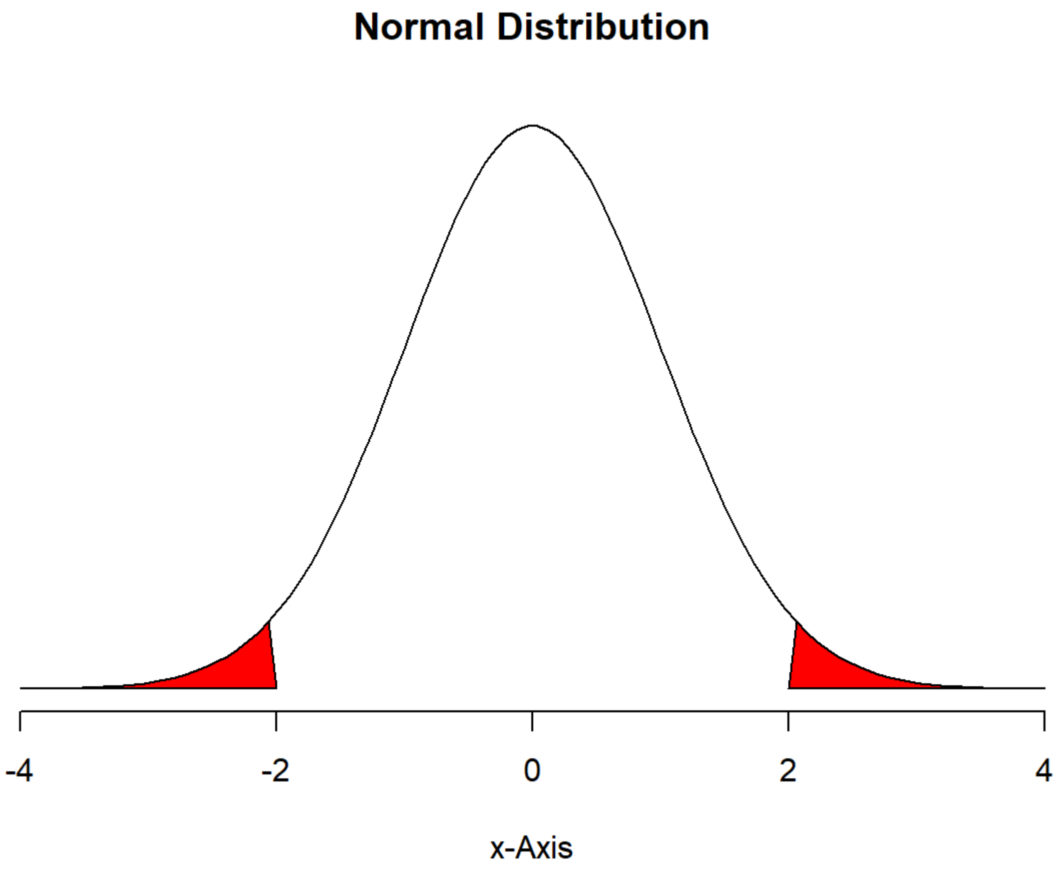
d. $|Z| > 2$

```
higher = 1-pnorm(2)
lower = pnorm(-2)

scales::percent(higher+lower, accuracy = 0.01)
```

```
## [1] "4.55%"
```

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



Triathlon times, Part I

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 group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the 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. For Men (Leo's Group): N(μ = 4313, σ = 583) For Women (Mary's Group): N(μ = 5261, σ = 807)

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

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

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

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

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

Both Leo's and Mary's Z-Scores are above the mean. Leo is 1.09 SD above the mean while Mary is only 0.31 SD above the mean.

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning. Because Mary's Z-Score is lower than Leo's, it indicates that she did better among her group than Leo did even though Leo ran faster than Mary.

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

```
scales::percent(1-pnorm(4948, mean=4313, sd=583), accuracy = 0.01)
```

```
## [1] "13.80%"
```

Leo ran faster than 13.8% of his competitors

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

```
scales::percent(1-pnorm(5513, mean=5261, sd=807), accuracy = 0.01)
```

```
## [1] "37.74%"
```

Mary ran faster than 37.74% of her competitors

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

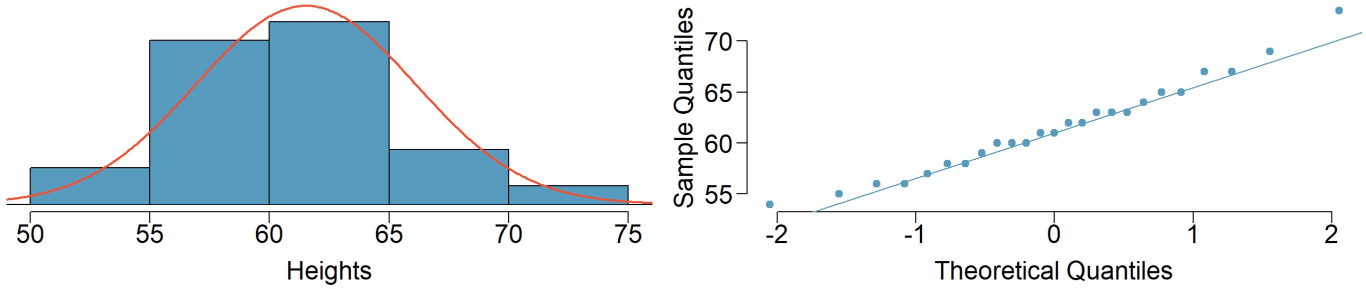
The Z-Scores (answer b) will be the same. We would not be able to calculate the other answers in a non-normal distribution.

Heights of Female College Students

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

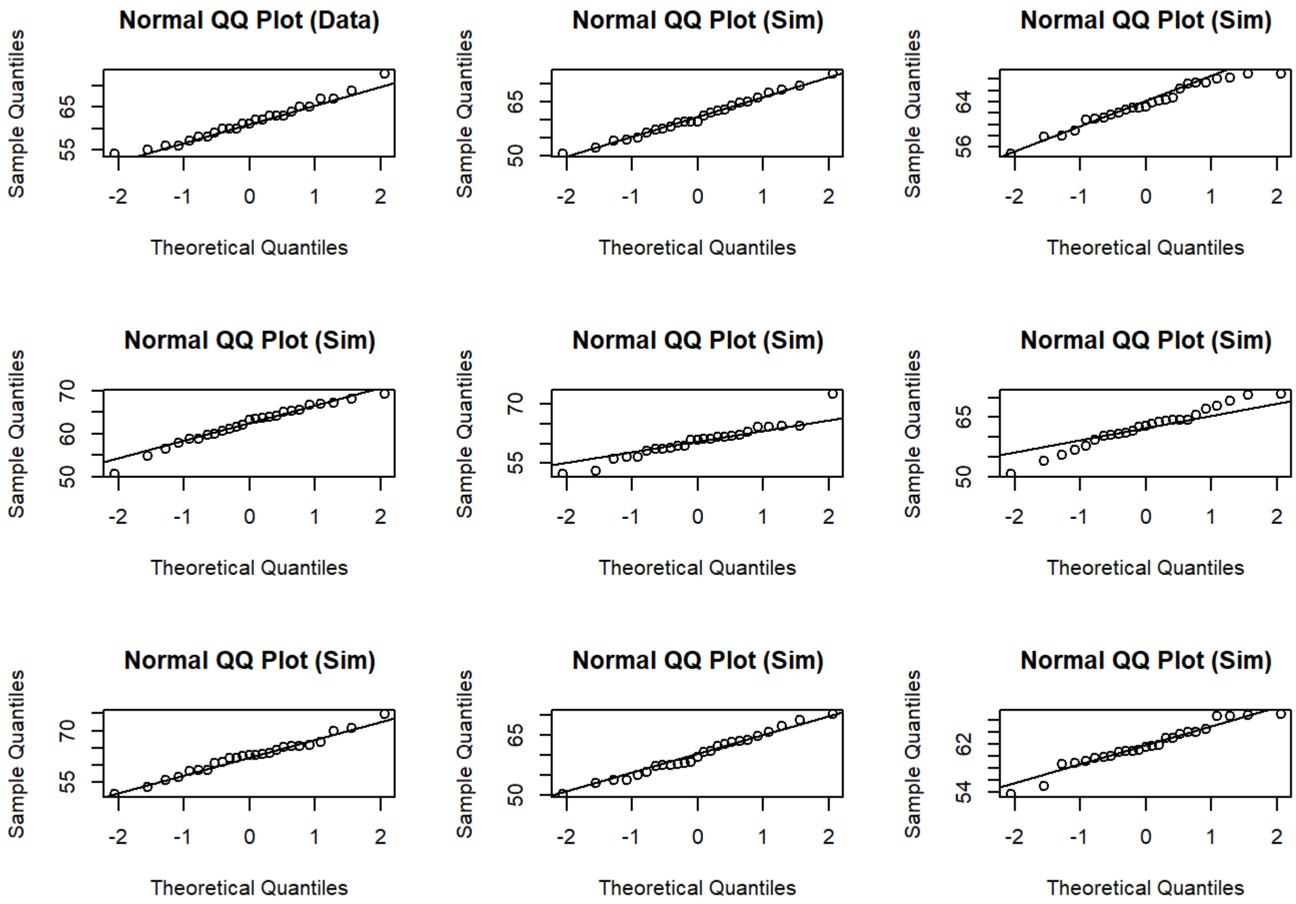
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

(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.



```
# Use the DATA606::qqnormsim function

qqnormsim(heights)
```



Do 68% lie within 1 standard deviation?

```
height_avg <- mean(heights)
height_sd <- sd(heights)
scales::percent(1-2*pnorm(height_avg+ height_sd, mean = height_avg, sd = height_sd, lower.tail =
FALSE), accuracy = 0.01)
```

```
## [1] "68.27%"
```

Do 95% lie within 2 standard deviations?

```
scales::percent(1-2*pnorm(height_avg+ 2*height_sd, mean = height_avg, sd = height_sd, lower.tail
= FALSE), accuracy = 0.01)
```

```
## [1] "95.45%"
```

Do 99.7% lie within 3 standard deviations?

```
scales::percent(1-2*pnorm(height_avg+ 3*height_sd, mean = height_avg, sd = height_sd, lower.tail
= FALSE), accuracy = 0.01)
```

```
## [1] "99.73%"
```

Based on the previous results,yes, the heights follow approximately the 68-95-99.7% Rule.

(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided above. Based on the histogram and the normal probability plots, even if it’s not a perfectly normal distribution it is a normal distribution.

Defective Rate.

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?

```
def_rate <- 0.02
(1-def_rate)**(10-1)*def_rate
```

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

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

```
(1-def_rate)**100
```

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

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

```
1/def_rate
```

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

What is the standard deviation?

```
sqrt((1-def_rate)/(def_rate**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?

```
1/0.05
```

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

What is the standard deviation?

```
sqrt((1-0.05)/(0.05**2))
```

```
## [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?

The mean and the standard deviation are inversely proportional to the probability.

Male Children

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.

```
p = 0.51
k = 2
n = 3

fact_k <- factorial(k)
fact_n <- factorial(n)
fact_nk <- factorial(n-k)

(fact_n / (fact_k * fact_nk)) * p ** k * ( 1-p)**(n-k)
```

```
## [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.

```
ord_1 <- c("boy", "boy", "girl")
ord_2 <- c("boy", "girl", "boy")
ord_3 <- c("girl", "boy", "boy")

( p * p * (1-p)) + ( p * (1-p) * p ) + ((1-p) * p * p )
```

```
## [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).

Because it would increase the sequence of numbers we would need to multiply and add to get the value. We would also need to account for all the possible permutations of 8 kids with 3 being boys.

Serving in Volleyball

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?

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

fact_n <- factorial(n-1)
fact_k <- factorial(k-1)
fact_nk <- factorial(n-k)

fact_n/(fact_k * fact_nk) * p ** k * (1-p)**(n-k)
```

```
## [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? 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? On b we’re calculating for a single independent event, in a, we’re calculating for a sequence of events to happen within 10 occurences.