S&DS 220: Homework 4

Due Friday February 9

Braeden Cullen

Instructions

1. Complete the questions below. Upload your knitted PDF solutions to Gradescope by the due date.
2. Your solutions should be a combination of writing and R code. When writing, use complete sentences.
3. Previous homework assignments already had code chunks created for you. Now it is up to you to insert R code chunks within each problem as needed.
4. You should aim for clear and concise communication (in both words and R code).

Problem set questions

Question 1: Exercise 2.23

Suppose you do an experiment where you select ten people at random and ask their birthdays.

Here are three events:

- A: all ten people were born in February.
- B: the first person was born in February.
- C: the second person was born in January.

(a) Which pair(s) of these events are disjoint, if any?

Solution A,C

There is one disjoint pair. Events are disjoint if they cannot occur at the same time. In this case, if all ten people were born in February, then the first person was born in February, and the second person was born in January. Thus, A and C are disjoint.

(b) Which pair(s) of these events are independent, if any?

Solution B,C

Events are independent if the occurrence of one event does not affect the probability of the occurrence of the other event. This is the only possible pair of independent events.

(c) What is $P(B|A)$?

Solution The probability is 1 because if all ten people were born in February, then the first person was born in February.

Question 2: Exercise 2.26

Suppose a die is tossed three times. Let A be the event "the first toss is a 5". Let B be the event "the first toss is the largest number rolled" (the "largest" can be a tie). Determine, via simulation or otherwise, whether A and B are independent.

```
num_trials <- 1e5

# checking for A and B
trials_A_B <- replicate(num_trials, {
  die_tosses <- sample(1:6, size = 3, replace = TRUE)

  A <- die_tosses[1] == 5

  B <- die_tosses[1] == max(die_tosses)

  A & B
```

```
})

# only checking if the first number is 5
trials_A <- replicate(num_trials, {
  die_tosses <- sample(1:6, size = 3, replace = TRUE)

  A <- die_tosses[1] == 5

  A
})

# does not consider the first element being 5, only the largest number rolled
trials_B <- replicate(num_trials, {
  die_tosses <- sample(1:6, size = 3, replace = TRUE)

  B <- die_tosses[1] == max(die_tosses)

  B
})
trials_A_B_mean <- mean(trials_A_B) # 0.11619
trials_A_mean <- mean(trials_A)   # 0.16691
trials_B_mean <- mean(trials_B)   # 0.42241

trials_A_mean * trials_B_mean # P(A) * P(B) = 0.0706
```

Solution

```
## [1] 0.07008846
```

```
trials_A_B_mean * trials_A_mean # P(A intersection B) * P(A) =  0.01937676
```

```
## [1] 0.01929791
```

Ok now that we have our results, we can complete the picture. The independence of two events, A and B, are considered independent if the occurence of one does not affect the probability of the occurrence of the other. The equation representing this situation is as follows: P(A intersection B) = P(A) * P(B). In our case, we have P(A intersection B) = 0.11619, P(A) = 0.16691, P(B) = 0.42241, P(B) * P(A) = 0.0706. We can see that P(A intersection B) is not equal to P(A) * P(B), so we can conclude that A and B are **not independent.**

Question 3: Exercise 2.30

Suppose there is a new test that detects whether people have a disease. If a person has the disease, then the test correctly identifies that person as being sick 99.9% of the time (*sensitivity* of the test). If a person does not have the disease, then the test correctly identifies the person as being well 97% of the time (*specificity* of the test). Suppose that 2% of the population has the disease. Find the probability that a randomly selected person has the disease given that they test positive for the disease.

Solution to solve this problem we will make use of Bayes Theorem, which is useful in this case because it allows us to update the probabilities of an event when given new information (in this case, the result of the test). The equation for Bayes Theorem is as follows: P(A | B) = P(B | A) * P(A) / P(B). In this case, A is

the event that a person has the disease, and B is the event that a person tests positive for the disease. We are given the following information: P(B | A) = 0.999, P(A) = 0.02, and P(B) = P(B | A) * P(A) + P(B | A complement) * P(A complement). We can calculate P(B | A complement) as 1 - P(B | A complement) = 0.97, and P(A complement) = 1 - P(A) = 0.98. Plugging in these values, we get P(B) = 0.999 * 0.02 + 0.03 * 0.98 = 0.04994. Now we can calculate P(A | B) = 0.999 * 0.02 / 0.04994 = 0.3992791. So the probability that a randomly selected person has the disease given that they test positive for the disease is 0.40.

lets do this in R for more clarity

```
P_B_A <- 0.999
P_A <- 0.02
P_B_A_complement <- 1 - 0.97
P_A_complement <- 1 - 0.02
P_B <- P_B_A * P_A + P_B_A_complement * P_A_complement
P_A_B <- P_B_A * P_A / P_B
P_A_B # 0.4046173 = approx. 0.40
```

```
## [1] 0.4046173
```

Question 4: Exercise 2.35

You should answer these questions without using simulation (however, feel free to check your answer with simulation). Six standard six-sided dice are rolled.

(a) How many outcomes are there?

Solution 6^6 = 46656 lets test it using simulation

```
# determine the number of possible outcomes when six six-sided dice are rolled
outcomes <- replicate(1e6, {
  die_rolls <- sample(1:6, size = 6, replace = TRUE)
  paste(die_rolls, collapse = "")
})

length(unique(outcomes)) # 46656
```

```
## [1] 46656
```

(b) How many outcomes are there such that all of the dice are different numbers?

Solution die 1 -> can land on 6 different numbers die 2 -> can land on 5 different numbers die 3 -> can land on 4 different numbers die 4 -> can land on 3 different numbers die 5 -> can land on 2 different numbers die 6 -> can land on 1 number

$6*5*4*3*2*1 = 6! = 720$

(c) What is the probability that you obtain six different numbers when you roll six dice?

Solution total number of outcomes = 46656 number of outcomes such that all of the dice are different numbers = 720 P(all dice are different) = 720 / 46656 = 0.01543209876

Question 5: Loops in R

Read the Vignette: Loops in R at the end of chapter 3 on pages 82 and 83 in our text. Below is an example of a `while` loop that counts the number of coin tosses until the 50th time heads appears, where the probability of heads is 0.4.

```
# example of a while loop

heads_50 <- replicate(1e4, {
  # initialize the number of heads to 0
  n_heads <- 0
  # initialize the number of coin tosses to 0
  n_tosses <- 0

  while(n_heads < 50){
    # simulate a coin toss
    coin_toss <- sample(c("H", "T"), size = 1, prob = c(0.4, 0.6))

    # increment the number of tosses by 1
    n_tosses <- n_tosses + 1

    # if the coin toss is heads, increment the number of heads by 1
    if(coin_toss == "H") {
      n_heads <- n_heads + 1
    }

  }
  # return the total number of tosses
  n_tosses
})
```

(a) Write a `while` loop to count the number of times a standard six-sided die is rolled until an accumulated total of 250 or more is achieved. Include your `while` loop code inside `replicate` to simulate this experiment 10,000 times.

```
num_trials <- 1e4

n_rolls_matrix <- replicate(num_trials, {
  accumulated_total <- 0
  n_rolls <- 0

  while(accumulated_total < 250){
    # simulate a die roll
    die_roll <- sample(1:6, size = 1)

    accumulated_total <- accumulated_total + die_roll

    # increment the number of rolls by 1
    n_rolls <- n_rolls + 1
  }
  n_rolls
```

```
})

mean(n_rolls_matrix) # 71.91151
```

Solution

```
## [1] 71.9289
```

```
sd(n_rolls_matrix) # 4.13353
```

```
## [1] 4.151213
```

```
# find number of elements of n_rolls_matrix that are less than or equal to 75
sum(n_rolls_matrix <= 75) / num_trials # 0.8016
```

```
## [1] 0.8056
```

```
# find number of elements of n_rolls_matrix that are greater than 70
sum(n_rolls_matrix > 70) / num_trials # 0.6245
```

```
## [1] 0.6177
```

```
# solve for number of rolls between 71 and 75
sum(n_rolls_matrix >= 71 & n_rolls_matrix <= 75) / num_trials # 0.4261
```

```
## [1] 0.4233
```

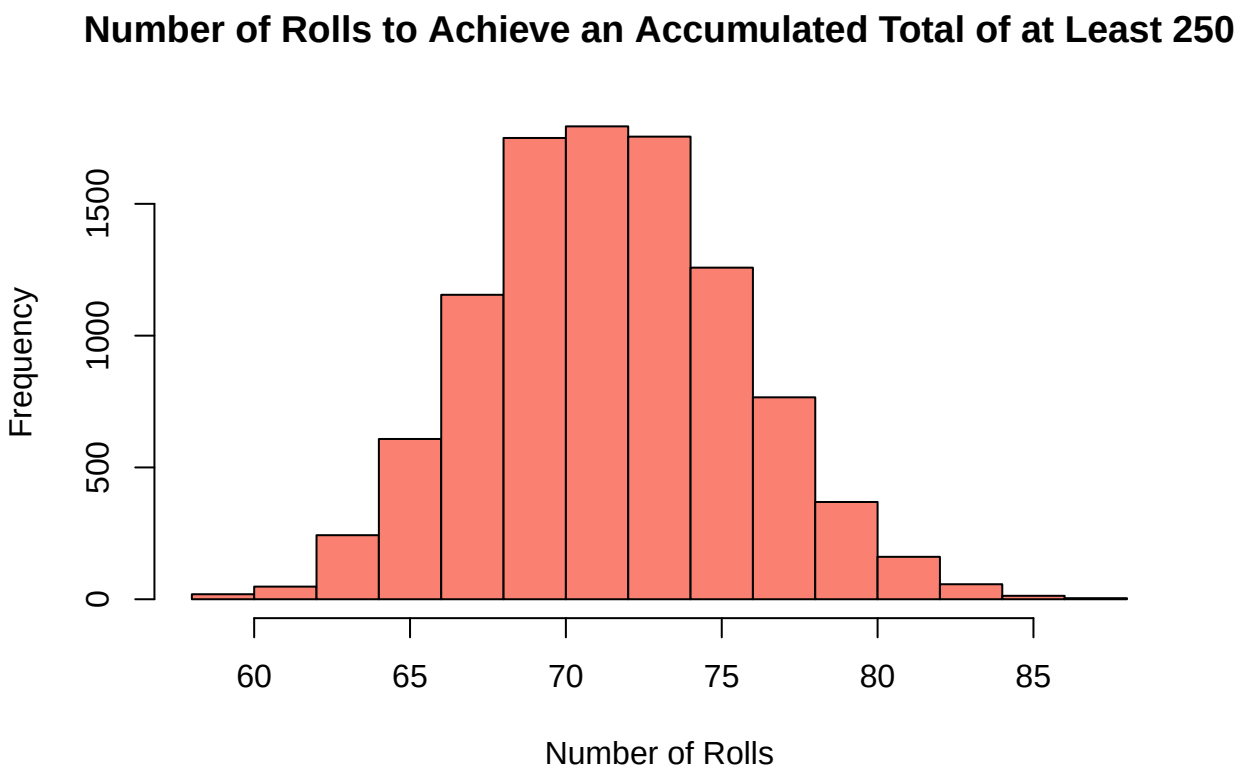
Use the results of the simulation you wrote in part (a) to answer parts (b) through (e). In other words, don't re-simulate the experiment.

(b) Estimate the mean and standard deviation of the number of rolls.

Solution mean = 71.91151 standard deviation = 4.13353

(c) Make a histogram of the number of rolls. Give the histogram an appropriate title and axes labels.

```
# construct historgram with number of rolls
hist(n_rolls_matrix, main = "Number of Rolls to Achieve an Accumulated Total of at Least 250", xlab = "
```



(d) Estimate the probability that 75 or fewer rolls are required.

Solution 0.8016

(e) Suppose after 70 rolls, the accumulated dice total less than 250. What is the probability an accumulated total of 250 or more will be achieved in the next 5 rolls?

Solution the formula P(A | B) = P(A intersection B) / P(B) describes the probability of event A occurring given that event B has occurred.

In the context of this problem: A is the event that the accumulated total is 250 or more after the next 5 rolls. B is the condition that, after 70 rolls, the accumulated total is less than 250.

P(B) = probability that the number of rolls will be greater than 70 = 0.62 P(A intersection B) = probability that the number of rolls will be between 71 and 75 = 0.42

P(A intersection B) / P(B) = 0.42 / 0.62 = 0.68

0.68