PSTAT 10: Homework 4

YOUR LAST NAME, YOUR FIRST NAME: YOUR NETID

Due on Canvas by 11:59pm on Week 5 Wednesday October 26

- Instructions for Submission
- Computing Exercises:
  - Exercise 1: Simulation
  - Exercise 2: Constructing and Plotting P.M.F.'s from Observed Data
  - Exercise 3: Doubling-Down
- Theory Exercises:
  - Exercise 4: Basic Probabilities
  - Exercise 5: Picking a Number
  - Exercise 6: Netflix
  - Exercise 7: Lost Airpods

Instructions for Submission

- This worksheet consists of two parts: Computing Exercises, and Theory Exercises.
- The submission portal on Canvas has 3 questions:
  - One for uploading your `.html` file for the Computing Exercises
  - One for uploading your `.zip` file containing both the `.html` and `.Rmd` files (along with any relevant datasets/images) for the Computing Exercises
  - One for uploading a `.PDF` of a photo or scan of your handwritten work for the Theory Exercises
- As usual, 5 Multiple Choice Questions also appear on Canvas and are a required part of this homework.
- Remember to save your `.Rmd` file as `hw05-yournetid.Rmd`

Computing Exercises:

Exercise 1: Simulation

In this problem, we will explore the following question: In 9 rolls of a fair six-sided die, what is the probability that the sum of the rolls equals 22? Specifically, we shall work toward answering this through simulation, using $10,000$ replications.

(a) Answer the question using a `for` loop. Specifically: use only one call to the `sample()` function to simulate all 9 dice rolls, and use a `for` loop to generate each of the $10,000$ replications.

Solutions:

```
d <- 9
k <- 22
nreps <- 10000

count <- 0 # Count the number of TRUE outcomes

for (rep in seq_len(nreps)) {
  rolls <- sample(1:6, d, replace = T)
  if (sum(rolls) == k) {
    count <- count + 1 # The event occurs, so we increment the counter
  }
}

count/nreps
```

```
## [1] 0.0147
```

(b) Now, Rewrite your code in two or fewer lines. Specifically:

1. use the `replicate()` function to replace the `for` loop.
2. use the `mean()` function with appropriate logical expression(s) to calculate the required probability.

Here is the key and useful idea in calculating probability using simulations: For each replication, create a logical vector which is `TRUE` if the event of interest occurs and `FALSE` if the event does not occur. Then take the mean of this vector. This estimates the probability of the event. *Make sure you understand why this works. Come to ULA/TA/instructor office hours if you aren't sure.*

To get stared with this exercise **think through this:** What is the event you are interested in focusing on? Can you write that event as a logical (T/F) expression? If so, what do you need to feed into the logical expression regarding your dice rolls? The answer to these questions helps you figure out what expression to feed into the `replicate` function. *As I mentioned, in simulation, you tend to work backwards and in handwritten problems you tend to work forwards to solve the problem.*

Solutions:

```
dice_rolls_sum <- replicate(10000, sum(sample(1:6, size =  9, replace = T)))
mean(dice_rolls_sum == 22)
```

```
## [1] 0.0153
```

(c) Are your answers to part (a) and (b) the same? Why or why not?

Solutions:

No; due to the randomness of simulations, there is no guarantee we will observe precisely the same proportion each time we run the same code.

(d) Roll three fair six-sided dice. What's the probability that the minimum of the scores is greater than 2?

Solutions:

```
reps <- replicate(1000, min(sample(1:6, 3, replace = T)))
mean(reps > 2)
```

```
## [1] 0.295
```

Exercise 2: Constructing and Plotting P.M.F.'s from Observed Data

The number of times the audio system in IV Theatre 1 failed was observed over a period of 1 year. It was found that

- 0 failures occurred in each of 9 weeks
- 1 failure occurred in each of 14 weeks
- 2 failures occurred in each of 13 weeks
- 3 failures occurred in each of 9 weeks
- 4 failures occurred in each of 4 weeks
- 5 failures occurred in each of 2 weeks
- 6 failures occurred in 1 week

(a) Use this observed data to estimate the probability distribution of the audio system failures in a week. You can use the `round` function to make the probabilities round to 2 decimals. Check that it is a valid probability distribution

Solutions:

```
num_failures <- 0:6
weekly_failures <- c(9, 14, 13, 9, 4, 2, 1)
sum(weekly_failures)
```

```
## [1] 52
```

```
prob_failures = round(weekly_failures/sum(weekly_failures),2)


pdf <- prob_failures
names(pdf) <- num_failures

pdf
```

```
##    0    1    2    3    4    5    6 
## 0.17 0.27 0.25 0.17 0.08 0.04 0.02
```

```
prob_failures
```

```
## [1] 0.17 0.27 0.25 0.17 0.08 0.04 0.02
```

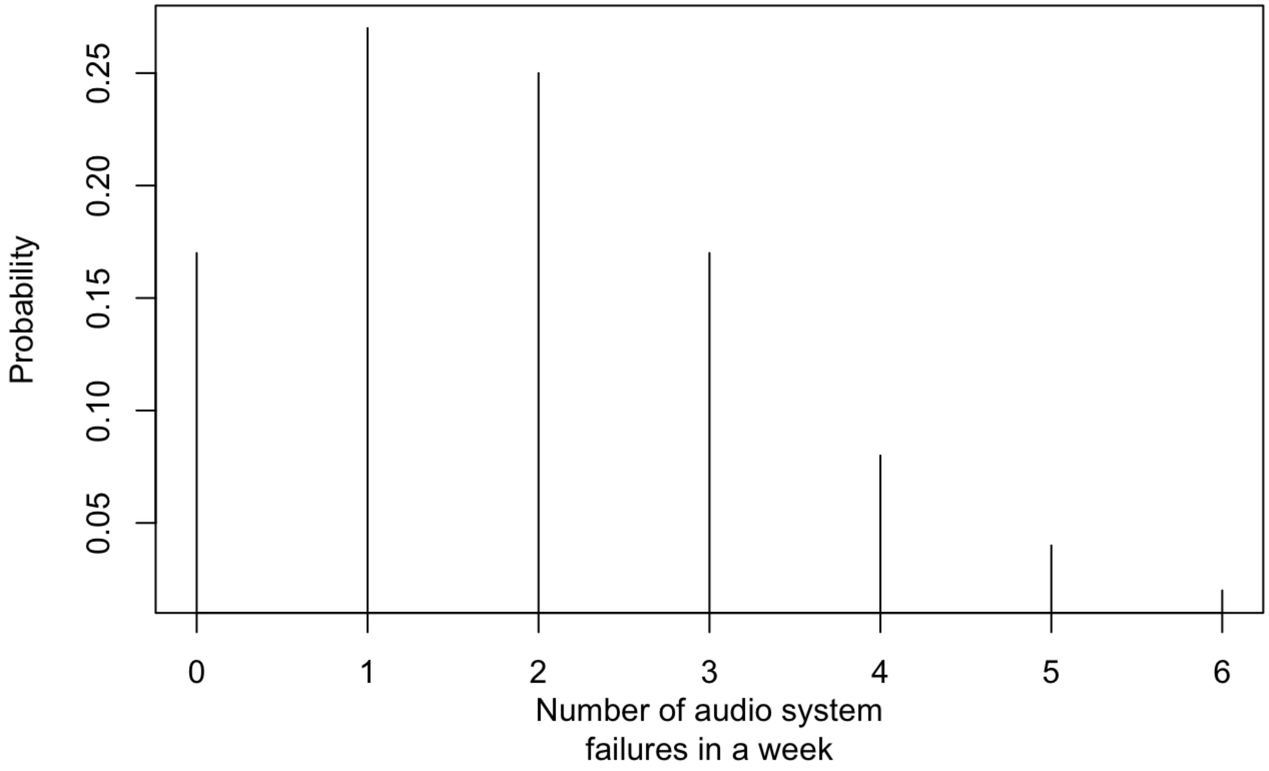
```
sum(prob_failures)
```

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

(b) Plot this distribution. Use the plot function with the `type = h` option to draw a line at each failure number. Explain what the height of each line represents in this plot.

Solutions:

```
num_failures <- 0:6
plot(num_failures, prob_failures, xlab="Number of audio system
failures in a week", ylab="Probability", type="h")
```



(c) Find the cumulative distribution function of the audio system failures in a week.

Solutions:

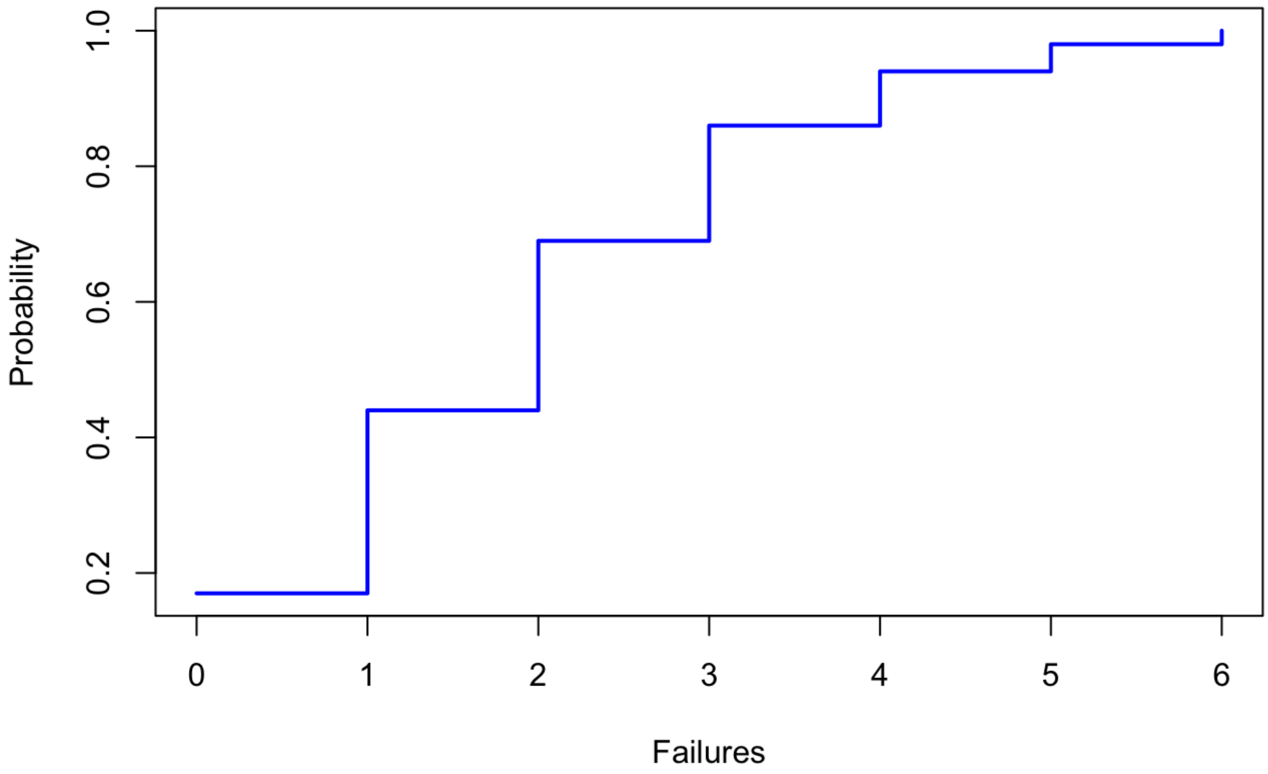
```
(cum_prob_weekly_failures <- cumsum(prob_failures))
```

```
## [1] 0.17 0.44 0.69 0.86 0.94 0.98 1.00
```

(d) Plot the c.d.f using `type="S"` to get the stair-case plot. **Correction:** This should have read `type = "s"`.

Solutions:

```
plot(num_failures, cum_prob_weekly_failures, xlab = "Failures",
ylab = "Probability", type = "s", col = "blue", lwd = 2)
```



(e) Use your p.m.f and c.d.f and the plot from (d) to approximately calculate $\mathbb{P}(X \leq 3.5)$. Which was the easiest way to calculate this probability?

Solutions:

```
# Using c.d.f
cum_prob_weekly_failures[4]
```

```
## [1] 0.86
```

```
# Using p.m.f
sum(prob_failures[1:4])
```

```
## [1] 0.86
```

(f) Use your p.m.f and c.d.f and the plot from (d) to approximately calculate $\mathbb{P}(X \leq 1)$. Understand how the calculation for (e) differs from that of (f) while using the plot.

Solutions:

```
# Using c.d.f
cum_prob_weekly_failures[2]
```

```
## [1] 0.44
```

```
# Using p.m.f
sum(prob_failures[1:2])
```

```
## [1] 0.44
```

To answer the latter part of the question, the key is to note that 3.5 is *not* a value that the random variable X can take, whereas 1 *is*. This means that the c.d.f. of X will be continuous at 3.5 whereas the c.d.f. will be discontinuous at 1. To read off the appropriate probability for (f) from the graph, we need to take the upper value at 1; i.e. $\mathbb{P}(X \leq 1) = 0.44$.

Exercise 3: Doubling-Down

In a past iteration of PSTAT 10, the student demographics of the class (specifically which major/s students were enrolled in) was recorded. The data has been stored in the file `roster.csv` . The dataset contains three columns:

- `id` : the student's ID number (anonymized)
- `major1` : the student's first major
- `major2` : the students second (double) major; `NA` if the student is enrolled in only one major.

(a) Load the data into your `.Rmd` file. **Hint:** `read.csv()` .

Solutions:

```
dat2 <- read.csv("./data/roster.csv")
```

(b) Write code to determine the number of students included in the dataset; do **NOT** just view the data and manually count!

Solutions:

```
n <- nrow(dat2)
n
```

```
## [1] 235
```

(c) What is the probability that a randomly selected student has Statistics as their `major1` ?.

Solutions:

Because selection is done at radom, we can use the Classical Definition of Probability to compute the desired probability as

$$\frac{\text{number of students with Statistics as their major1}}{\text{total number of students}}$$

```
length(which(dat2$major1 == 'stats')) / n
```

```
## [1] 0.3404255
```

(d) What is the probability that a randomly selected student is a double-major? **Hint:** Think carefully about how double-major-ness is encoded in the dataset.

Solutions:

The key is to note that double-majors will have non- `NA` `major2` values. Therefore, using the Classical Definition of Probability once again, we compute the desired probability as

$$\frac{\text{number of students with a non-NA major2 value}}{\text{total number of students}}$$

```
sum(!is.na(dat2$major2)) / n
```

```
## [1] 0.4893617
```

There are several different ways to get this probability: some alternate approaches include loops, or using the `na.omit()` function in `R` .

Theory Exercises:

Exercise 4: Basic Probabilities

Let A and B be two events such that $\mathbb{P}(A) = 0.2$, $\mathbb{P}(B) = 0.3$, and $\mathbb{P}(A \cap B) = 0.1$.

a. Compute $\mathbb{P}(A \cup B)$
b. Compute $\mathbb{P}(A^{\complement} \cap B^{\complement})$. **Hint:** Remember what we did in Lab Section, on Worksheet 7.
c. Are A and B independent? Why or why not?
d. Are A and B mutually exclusive? Why or why not?

Solutions:

a. By the Addition Rule, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.2 + 0.3 - 0.1 = \boxed{0.4}$

b. The complement of the event $(A \cup B)$ is $(A^{\complement} \cap B^{\complement})$ [as was seen in Lab]. Therefore, $\mathbb{P}(A^{\complement} \cap B^{\complement}) = 1 - \mathbb{P}(A \cup B) = 1 - 0.4 = \boxed{0.6}$

c. No; if they were independent then $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) = 0.3 \cdot 0.2 = 0.06$, whereas we are told $\mathbb{P}(A \cap B) = 0.1$.

d. No; if they were mutually exclusive then $(A \cap B) = \varnothing$ so $\mathbb{P}(A \cap B)$ would be 0, not 0.1.

Exercise 5: Picking a Number

A number is picked at random from the set $\{1, 2, \cdots, 100\}$.

a. What is the probability that it is a multiple of 3?
b. What is the probability that it is greater than 50?
c. What is the probability that it is a multiple of 3 that is strictly greater than 50?
d. What is the probability that it is a multiple of 3, **given** that it is strictly greater than 50?

Solutions:

For all parts, we use the Classical Definition of Probability.

a. $\mathbb{P}(\text{multiple of 3}) = \dfrac{\text{\# of multiples of 3 in } \{1, \cdots, 100\}}{\text{\# of elements in } \{1, \cdots, 100\}} = \boxed{\dfrac{33}{100}}$

b. $\mathbb{P}(\text{greater than 50}) = \dfrac{\text{\# of elements greater than 50 } \{1, \cdots, 100\}}{\text{\# of elements in } \{1, \cdots, 100\}} = \boxed{\dfrac{50}{100} = \dfrac{1}{2}}$

c. $\mathbb{P}(\text{mult. of 3 AND} > 50) = \dfrac{\text{\# of multiples of 3 that are greater than 50}}{\text{\# of elements in } \{1, \cdots, 100\}} = \boxed{\dfrac{17}{100}}$

d. $\mathbb{P}(\text{mult. of } 3 \mid > 50) = \dfrac{\mathbb{P}(\text{mult. of 3 AND} > 50)}{\mathbb{P}(> 50)} = \dfrac{\frac{17}{100}}{\frac{1}{2}} = \boxed{\dfrac{17}{50}}$

Exercise 6: Netflix

A recent *Netflix* survey sureyed several individuals about whether they prefer TV Shows, Movies, or have no strong prefernece either way. The age bracket of each individual was also recorded; the results of the survey are displayed below.

	Prefer TV Shows	Prefer Movies	No Preference
< 15	10	15	10
15 - 25	15	20	25
26 - 35	30	15	10
36 +	10	15	5

If a person is selected at random from the pool surveyed individuals, what is the probability that…

a. … they prefer TV Shows, **and** are between 15 and 25 years old?
b. … they prefer TV Shows?
c. … they are over 36 years old?
d. … they prefer movies, **given** that they are between 26 and 35 years old?

Solutions:

It may be useful to first note that there are a total of 180 people included in the survey.

a. $\text{answer} = \boxed{\dfrac{15}{180}}$

b. $\text{answer} = \dfrac{10 + 15 + 30 + 10}{180} = \boxed{\dfrac{65}{180}}$

c. $\text{answer} = \dfrac{10 + 15 + 5}{180} = \boxed{\dfrac{30}{180}}$

d. $\text{answer} = \dfrac{15}{30 + 15 + 10} = \boxed{\dfrac{15}{55}}$

Exercise 7: Lost Airpods

Sally has lost her airpods somewhere in her apartment. Based on her activity over the past few days, she has the following beliefs:

- She is 25% certain they are in her living room
- She is 40% certain they are in her bedroom
- She is 30% certain they are in the bathroom
- She is 5% certain they are not in her apartment at all.

Based on how messy each room of her apartment is, she also believes:

- A search of her living room has a 60% chance of successfully locating her airpods
- A search of her bedroom has a 80% chance of successfully locating her airpods
- A search of her bathroom has a 30% chance of successfully locating her airpods
- A search of any location outside her apartment has a 10% chance of successfully locating her airpods

Use this information to answer the following questions:

a. What is the probability that she searches her living room **and** is successful at locating her airpods?

b. What is the probability that she searches her bedroom **and** is successful at locating her airpods?

c. What is the probability that she is successful at locating her airpods?

Solutions:

Let's start (as we should with every probability problem!) by defining some events. Let:

- L = "Sally's airpods were in the living room"
- B = "Sally's airpods were in the bedroom"
- T = "Sally's airpods were in the bathroom"
- N = "Sally's airpods were in not in her apartment"

Additionally, let the event S denote "Sally is successful at locating her airpods." With this notation in mind, we the information provided in the problem can be translated as:

- $\mathbb{P}(L) = 0.25$
- $\mathbb{P}(B) = 0.40$
- $\mathbb{P}(T) = 0.30$
- $\mathbb{P}(N) = 0.05$
- $\mathbb{P}(S \mid L) = 0.60$
- $\mathbb{P}(S \mid B) = 0.80$
- $\mathbb{P}(S \mid T) = 0.30$
- $\mathbb{P}(S \mid N) = 0.10$

a. We seek $\mathbb{P}(S \cap L)$. By the Multiplication Rule, we can compute this as

$$\mathbb{P}(S \cap L) = \mathbb{P}(L) \cdot \mathbb{P}(S \mid L) = (0.25) \cdot (0.60) = \boxed{0.15}$$

b. We seek $\mathbb{P}(S \cap B)$. By the Multiplication Rule, we can compute this as

$$\mathbb{P}(S \cap B) = \mathbb{P}(L) \cdot \mathbb{P}(S \mid B) = (0.25) \cdot (0.80) = \boxed{0.20}$$

c. We seek $\mathbb{P}(S)$. By the Law of Total Probability, we compute this as:

$$\begin{aligned}
\mathbb{P}(S) &= \mathbb{P}(S \cap L) + \mathbb{P}(S \cap B) + \mathbb{P}(S \cap T) + \mathbb{P}(S \cap N) \\
&= \mathbb{P}(L) \cdot \mathbb{P}(S \mid L) + \mathbb{P}(B) \cdot \mathbb{P}(S \mid B) + \mathbb{P}(T) \cdot \mathbb{P}(S \mid T) + \mathbb{P}(N) \cdot \mathbb{P}(S \mid N) \\
&= (0.25) \cdot (0.60) + (0.25) \cdot (0.80) + (0.30) \cdot (0.30) + (0.05) \cdot (0.10) \\
&= \boxed{0.445}
\end{aligned}$$