DSCI 551 Lab 1

Depicting Uncertainty and Parametric Families

Contents

Lab Mechanics **2**

One-time set-up **2**

A Note on Challenging Questions **2**

A Note on Warmup Exercises **3**

(Warmup) Exercise 1: Collective Risk **4**

Exercise 2: The Slot Machine **5**
2.1. 5
2.2. 6

Excercise 3: Serendipity **8**
3.1. 8
3.2. 9
3.3. 9

Excercise 4: The Vancouver Whitecaps Football Club **11**
Challenging 13

Exercise 5: The Die **15**

Exercise 6: Choose the Distribution **16**
6.1. 16
6.2. 16

Exercise 7 **18**

(Challenging) Exercise 8 **20**

Lab Mechanics

rubric={mechanics:5}

Check off that you have read and followed each of these instructions:

- ☐ All files necessary to run your work must be pushed to your `GitHub.ubc.ca` repository for this lab.
- ☐ Paste the URL to your GitHub repo here: **INSERT YOUR GITHUB REPO URL HERE**
- ☐ You need to have **a minimum of 3 commit messages** associated with your `GitHub.ubc.ca` repository for this lab.
- ☐ **You must also submit `.Rmd` file and the rendered `pdf` in this lab to Gradescope.** You are responsible for ensuring all the figures, texts, and equations in the `.pdf` file are appropriately rendered.
- ☐ To ensure you do not break the autograder remove all code for installing packages (i.e., DO NOT have `install.packages(...)` or `devtools::install_github(...)` in your homework!
- ☐ Follow the **MDS general lab instructions**.

This lab has hidden tests. In this lab, the visible tests are just there to ensure you create an object with the correct name. The remaining tests are hidden intentionally. This is so you get practice deciding when you have provided the right numeric or text answer. This is a necessary skill for data scientists, and if we were to provide robust visible tests for all questions you would not develop this skill, or at least not to its full potential.

> **Important:** For the assignments, you are permitted to use LLMs only to gather information, review concepts, or brainstorm. You **must cite** these tools if you use them for the assignment. It is not permitted to write any given assignment by copying and pasting AI-generated responses.

One-time set-up

If you are using the `testthat` and `digest` packages for the first time and have never installed them before, do this:

1. Uncomment the two lines of code below by deleting the `#` at the start of each line.
2. Run the code cell, which will perform the installation.
3. Comment the two lines of code again by adding the `#` back to the start of each line.

```
# install.packages("testthat")
# install.packages("digest")
```

A Note on Challenging Questions

Each lab will have a few challenging questions. **These are usually low-risk questions and will contribute to a 5% out of the total lab grade of 100%**. The main purpose here is to challenge yourself or dig deeper in a particular area. When you start working on labs, attempt all other questions before moving to these questions. If you are running out of time, please skip these questions.

We will be more strict with the marking of these questions. If you want to get full points in these questions, your answers need to

- be thorough, thoughtful, and well-written (if necessary),
- provide convincing justification and appropriate evidence for the claims you make, and
- impress the reader of your lab with your understanding of the material, your analytical and critical reasoning skills, and your ability to think on your own.

A Note on Warmup Exercises

Each lab session will begin with a warmup exercise, which will be indicated when introduced in the handout. We will solve this exercise altogether during the first 10 minutes of the lab session.

Note this warmup part will not count toward the lab grade of 100%.

(Warmup) Exercise 1: Collective Risk

The BC government is assessing the risk of flu infection in a social gathering. As a data scientist, you are tasked with developing a probabilistic model to estimate the risk of the gathering. **Assuming the total population in BC is 5 million people**, of which 5,000 people have flu on a given day:

1.1. What is the probability that a randomly chosen person is infected with flu on a given day in BC?

1.2. What is the probability that a randomly chosen person is **not** infected with the flu on a given day in BC?

1.3. If we have a gathering of 20 people (e.g., a coffee shop), what is the probability that **at least** one person is infected with the flu?

1.4. Now, if we have a gathering of 200 people (e.g., a large lecture hall), what is the probability that **at least** one person is infected with the flu?

> **Note:** For **1.3** and **1.4**, we will assume that each person is **independent** from one another. Recall that if two events A and B are independent, then:
>
> $$P(A \cap B) = P(A) \cdot P(B).$$

ANSWER:

The reasoning behind the answers is the following:

1.1. *Let the event*

$$I = \text{An individual is infected with the flu on a given day in BC.}$$

Then

$$P(I) = \frac{5000}{5000000} = 0.001.$$

1.2. *We compute the complement of the previous probability:*

$$P(I^c) = 1 - \frac{5000}{5000000} = 0.999.$$

1.3. *Let the event*

$$X = \text{Number of people being infected with the flu in a gathering of 20 people.}$$

If we assume each individual as independent from one another, then:

$$P(X \geq 1) = 1 - \{[1 - P(I)]^{20}\}.$$

1.4. *Let the event*

$$Y = \text{Number of people being infected with the flu in a gathering of 200 people.}$$

If we assume each individual as independent from one another, then:

$$P(Y \geq 1) = 1 - \{[1 - P(I)]^{200}\}.$$

Exercise 2: The Slot Machine

Suppose a slot machine costs \$5 to play (call one pull of the lever a "game"). You win nothing with a probability of 0.9, \$10 with a probability of 0.09, and \$50 with a probability of 0.01. Let X be the **net gain** in one game. Then X has the following distribution:

X (net gain in one game)	-\$5	\$5	\$45
Probability	0.9	0.09	0.01

2.1.

rubric={autograde:5,reasoning:5}

Find the following quantities:

2.1.1. The expected value $\mathbb{E}(X)$.

2.1.2. $P[X = \mathbb{E}(X)]$, i.e., the probability that you achieve the average outcome in a single game.

2.1.3. $P(X < 0)$, i.e., the probability of losing money in a game.

2.1.4. The mode of X.

2.1.5. The variance $\text{Var}(X)$.

To get full marks on these questions, follow these instructions:

1. Provide all your procedure and/or reasoning (e.g., equations):
    a. You do not need to use **LaTex** to provide mathematical notation.
    b. Instead, you might work on your written answer on a separate piece of paper and take a picture of it.
    c. Then, you have to put this image in the folder `img` which is part of your lab GitHub repo.
    d. Finally, within this `R` markdown, use the following syntax: `![My caption](img/my_answer_2)`, where `my_answer_2` is your image's name. The output is the following:

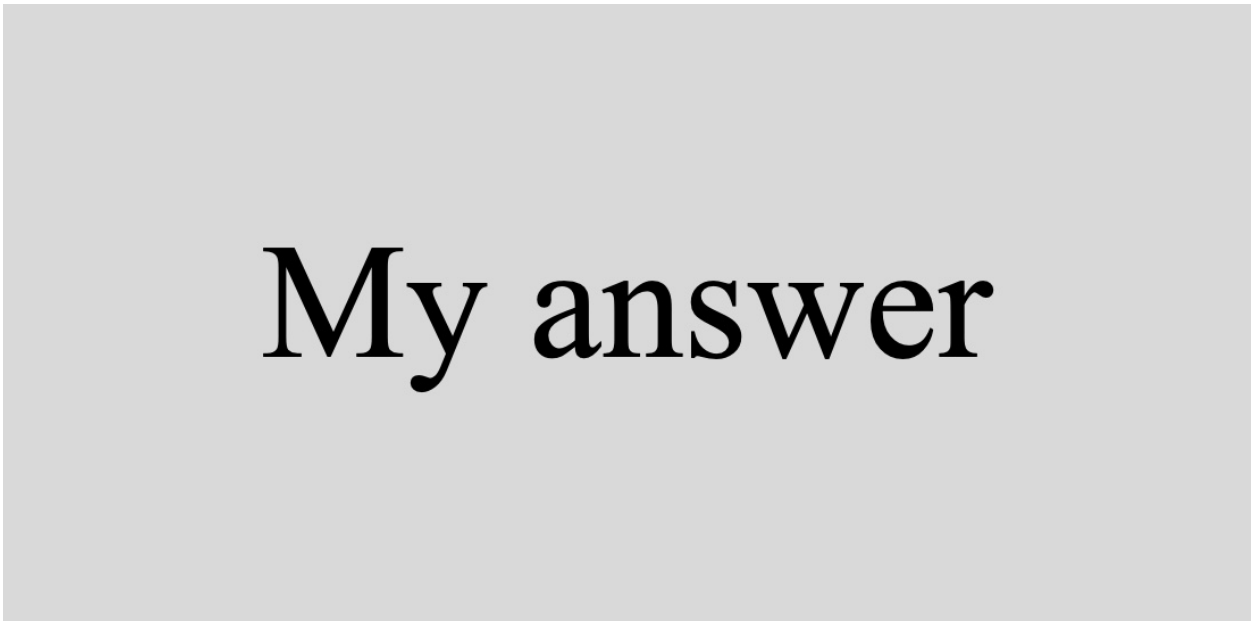

Figure 1: My caption

2. Assign your **final numeric answers** to `answer2_1_1`, `answer2_1_2`, `answer2_1_3`, `answer2_1_4`, and `answer2_1_5` in below chunk of code. **Code your computations directly in each answer if necessary.** Moreover, run the test below to validate your answers.

ANSWER:

The reasoning behind the answers is the following:

2.1.1.

$$\mathbb{E}(X) = -5(0.9) + 5(0.09) + 45(0.01) = -3.6.$$

2.1.2. *The probability mass function maps a probability of zero for all values different from -$5, $5, and $45. Then:*

$$P(X = -3.6) = 0.$$

2.1.3. *From the PMF, we know that*

$$P(X < 0) = P(X = -5) = 0.9.$$

2.1.4. *From the PMF, the largest probability is*

$$P(X = -5) = 0.9.$$

Thus, the mode is $X = -5$.

2.1.5. *The variance is computed as:*

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \\ &= (-5)^2(0.9) + 5^2(0.09) + 45^2(0.01) - (-3.6)^2 \\ &= 32.04. \end{aligned}$$

```
answer2_1_1 <- NULL
answer2_1_2 <- NULL
answer2_1_3 <- NULL
answer2_1_4 <- NULL
answer2_1_5 <- NULL

# BEGIN SOLUTION
answer2_1_1 <- -5 * (0.9) + 5 * (0.09) + 45 * (0.01)
answer2_1_2 <- 0
answer2_1_3 <- 0.9
answer2_1_4 <- -5
answer2_1_5 <- (-5)^2 * (0.9) + 5^2 * (0.09) + 45^2 * (0.01) - (-3.6)^2
# END SOLUTION
```

2.2.

rubric={autograde:3}

Now suppose you play the game 100 times. Assuming we have an infinite supply of money (so that we can always afford to play the game!), the net outcome after 100 games can be written as a new random variable:

$$Z = \sum_{i=1}^{100} X_i,$$

where all X_is are **identically distributed** random variables with the same distribution as X in **question 2.1**.

Find the following quantities:

2.2.1. The maximum value of Z.

2.2.2. The minimum value of Z.

2.2.3. The expected value $\mathbb{E}(Z)$.

Assign your **numeric answers** to `answer2_2_1`, `answer2_2_2`, and `answer2_2_3`. **Code your computations directly in each answer.** You do not need to provide all your procedure (i.e., equations!) to get full marks. Moreover, run the test below to validate your answers.

ANSWER:

The reasoning behind the answers is the following:

2.2.1. *The maximum of Z occurs when we win 100 times:*

$$\max(Z) = 100(45) = 4500.$$

2.2.2. *The minimum of Z occurs when we lose 100 times:*

$$\min(Z) = 100(-5) = -500.$$

2.2.3. *Since all X_is are iid, to obtain the expected value, we apply the following property:*

$$\begin{aligned}
\mathbb{E}(Z) &= \mathbb{E}\left(\sum_{i=1}^{100} X_i\right) \\
&= \sum_{i=1}^{100} [\mathbb{E}(X_i)] \\
&= 100\mathbb{E}(X_i) \qquad \text{since they are identically distributed} \\
&= 100(-3.6) = -360.
\end{aligned}$$

```
answer2_2_1 <- NULL
answer2_2_2 <- NULL
answer2_2_3 <- NULL

# BEGIN SOLUTION
answer2_2_1 <- 100 * 45
answer2_2_2 <- 100 * -5
answer2_2_3 <- 100 * -3.6
# END SOLUTION
```

Excercise 3: Serendipity

Imagine you are living on the UBC campus with a student population of 45,000. Assume your MDS cohort has 120 students.

Now, answer the following:

3.1.

rubric={autograde:1,reasoning:2}

Assume that seeing each person from the population is equally likely and independent of seeing anyone else. If you cross paths with 100 people on campus every day, what is the probability of you running into at least one MDS student?

> **Heads-up:** *Suppose you run into 100 people* ***with replacement****, which means that you might run into the same person twice.*

To get full marks on these questions, follow these instructions:

1. Provide all your procedure and/or reasoning (e.g., equations):
   a. You do not need to use **LaTex** to provide mathematical notation.
   b. Instead, you might work on your written answer on a separate piece of paper and take a picture of it.
   c. Then, you have to put this image in the folder `img` which is part of your lab GitHub repo.
   d. Finally, within this `R` markdown, use the following syntax: `![My caption](img/my_answer_3)`, where `my_answer_3` is your image's name. The output is the following:

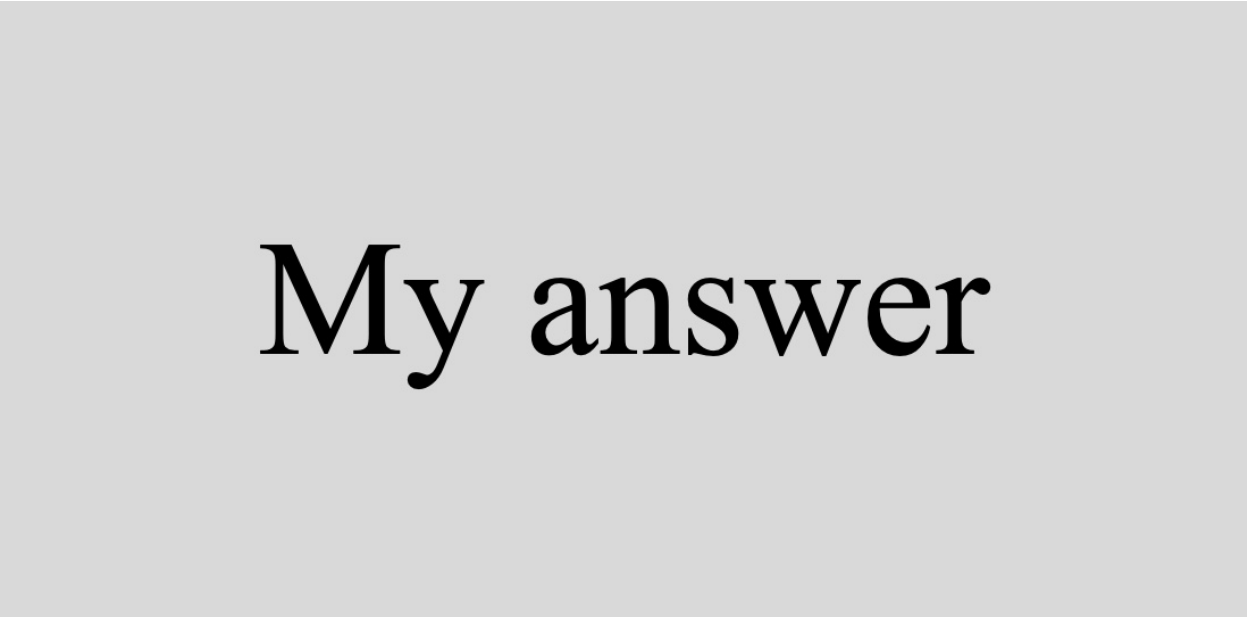

Figure 2: My caption

2. Assign your **numeric answer** to `answer3_1`. **Code your computation directly.** Moreover, run the test below to validate your answer.

ANSWER:

Firstly, the probability p of running into an MDS student on campus is the following:

$$p = \frac{120}{45000} = 0.0027.$$

Now, let

X = Number of MDS students among 100 you run into on campus with replacement.

Then, using the complement and assuming independence with equal probabilities for each MDS student, we can compute the corresponding probability:

$$P(X \geq 1) = 1 - (1 - p)^{100} = 0.2343.$$

```
answer3_1 <- NULL

# BEGIN SOLUTION
answer3_1 <- 1 - (1 - 120 / 45000)^100
# END SOLUTION
```

3.2.

rubric={autograde:1}

Assume that seeing each person from the population is equally likely and independent of seeing anyone else. If you cross paths with 100 people on campus every day, what is the probability of you running into at least one MDS student?

> **Heads-up:** *Suppose you run into 100 people* ***without replacement****, or in other words, you run into 100* ***different*** *people. You can use* `R` *to perform the computation. You may find the* `choose(n, k)` *function useful. This function calculates how many ways can we choose a* k *subset (no repetition) of an* n *set, or in math,* $\binom{n}{k}$*. For example, to compute the number of groups of 2 that can be formed among 4 students, you would use* `choose(4, 2)`*.*

Assign your **numeric answer** to `answer3_2`. **Code your computation directly.** You do not need to provide all your procedure (i.e., equations!) to get full marks. Moreover, run the test below to validate your answer.

ANSWER:

Let

Y = Number of MDS students among 100 you run into on campus without replacement.

Then, using the complement and with $n = 45000$*,* $k = 100$*, and* $f = 120$*, we can compute the corresponding probability:*

$$P(Y \geq 1) = 1 - \frac{\binom{n-f}{k}}{\binom{n}{k}} = 0.2346.$$

```
answer3_2 <- NULL

# BEGIN SOLUTION
n <- 45000
k <- 100
f <- 120
answer3_2 <- 1 - (choose(n - f, k) / choose(n, k))
# END SOLUTION
```

3.3.

rubric={reasoning:2}

Compare your answers from **3.1** and **3.2**. Are they similar or different? Briefly discuss **in one or two sentences**.

ANSWER:

They are extremely similar, which makes sense since you would not be that likely to run into the same person twice anyway on such a big campus (assuming everything really is independent).

Excercise 4: The Vancouver Whitecaps Football Club

rubric={autograde:8}

The odds of a particular event reflect the likelihood that the event will take place. Odds can be calculated as:

$$\frac{\text{number of favorable events}}{\text{number of unfavorable events}}$$

or

$$\frac{p}{1-p},$$

where p is the probability of the event occurring. **In our statistical context, note that odds are not probabilities.**

Say that the Vancouver Whitecaps Football Club (a past MDS Capstone partner!) won 10 out of 34 games last season. You want to use this information to predict their performance in the first game of the upcoming season. Let us assume they continue with the same win rate this season.

Answer the following:

4.1. What is the probability of the Whitecaps winning the game?

4.2. What are the odds **in favour** of the Whitecaps winning the game? Provide the corresponding ratio.

4.3. What are the odds **against** the Whitecaps winning the game? Provide the corresponding ratio.

4.4. If a sports website states that the odds in favour of the Whitecaps winning the game is 3:4, what is the probability they win the game (according to the sports website)?

4.5. If $P(\text{winning}) \leq 0.5$ what is the maximum odds?

4.6. If $P(\text{winning}) > 0.5$ what is the maximum odds?

4.7. Questions **4.5** and **4.6** hopefully showed you that odds are not symmetrical. This makes it difficult to compare the odds in favour of an event to the odds against an event. Instead, we can take the natural logarithm of the odds to make them symmetrical. The log of the odds is called the **logit function** and allows us to map probabilities from the range $[0, 1]$ to the full range of real numbers. Therefore, what is the log of the odds 10:24?

> **Heads-up:** The **logit function** is the basis of **logistic regression**, which you will learn about later in the program.

4.8. Finally, what is the log of the odds 24:10?

Assign your **numeric answers** to `answer4_1`, `answer4_2`, ..., `answer4_7`, and `answer4_8`. **Code your computations directly in each answer if necessary.** You do not need to provide all your procedure (i.e., equations!) to get full marks. Moreover, run the test below to validate your answers.

```
answer4_1 <- NULL
answer4_2 <- NULL
answer4_3 <- NULL
answer4_4 <- NULL
answer4_5 <- NULL
answer4_6 <- NULL
answer4_7 <- NULL
answer4_8 <- NULL

# BEGIN SOLUTION
answer4_1 <- 10 / 34
```

```r
answer4_2 <- answer4_1 / (1 - answer4_1)
answer4_3 <- (1 - answer4_1) / answer4_1
answer4_4 <- 3 / 7
answer4_5 <- 1 / 1
answer4_6 <- Inf
answer4_7 <- log(10 / 24)
answer4_8 <- log(24 / 10)
# END SOLUTION
```

ANSWER:

The reasoning behind the answers is the following:

4.1. *Let $P(winning)$ be the probability of the Whitecaps winning the game. It is computed as*

$$P(\text{winning}) = \frac{\text{Number of games won}}{\text{Total number of games}} = \frac{10}{34} = 0.294.$$

4.2. *The odds in favour of the Whitecaps are:*

$$o(w) = \frac{P(\text{winning})}{1 - P(\text{winning})} = \frac{0.294}{1 - 0.294} = 0.42.$$

4.3. *The odds against the Whitecaps are:*

$$o(a) = \frac{1 - P(\text{winning})}{P(\text{winning})} = \frac{1 - 0.294}{0.294} = 2.4.$$

4.4. *The probability according to the sports website is*

$$P_{\text{SW}}(\text{winning}) = \frac{3}{3 + 4} = 0.43.$$

4.5. *The maximum odds occur when $P(winning) = 0.5$:*

$$\frac{0.5}{1 - 0.5} = 1.$$

4.6. *The maximum odds occur when $P(winning) = 1$:*

$$\frac{1}{1 - 1} = \infty.$$

4.7. *The log of the odds can be computed as follows:*

$$\log(o) = \log\left(\frac{\frac{\text{Wins}}{\text{Total games}}}{\frac{\text{Losses}}{\text{Total games}}}\right) = \log\left(\frac{\text{Wins}}{\text{Losses}}\right).$$

Hence, in this case, the log of the odds is

$$\log(o) = \log\left(\frac{10}{24}\right) = -0.88.$$

4.8. *Now, the log of the odds is*

$$\log(o) = \log\left(\frac{24}{10}\right) = 0.88.$$

Challenging

rubric={reasoning:3}

In **binary logistic regression**, we aim to model a set of regressors or features (namely, the Xs) versus a **binary response** (e.g., win or loss) where the event of interest (e.g., win) has a probability p. If we have a single regressor X, we call our model **simple binary logistic regression**. In this case, p will be in function of X: $p(X)$. Moreover, we also include **regression parameters** in a linear function: an intercept β_0 and a coefficient β_1.

The model is set up as

$$\log\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 X.$$

Rearrange the equation above with a standalone $p(X)$ on the left-hand side.

To get full marks, show all your work. Thus, provide all your procedure and/or reasoning (e.g., equations):

a. You do not need to use **LaTex** to provide mathematical notation.
b. Instead, you might work on your written answer on a separate piece of paper and take a picture of it.
c. Then, you have to put this image in the folder `img` which is part of your lab GitHub repo.
d. Finally, within this `R` markdown, use the following syntax: `![My caption](img/my_answer_4)`, where `my_answer_4` is your image's name. The output is the following:

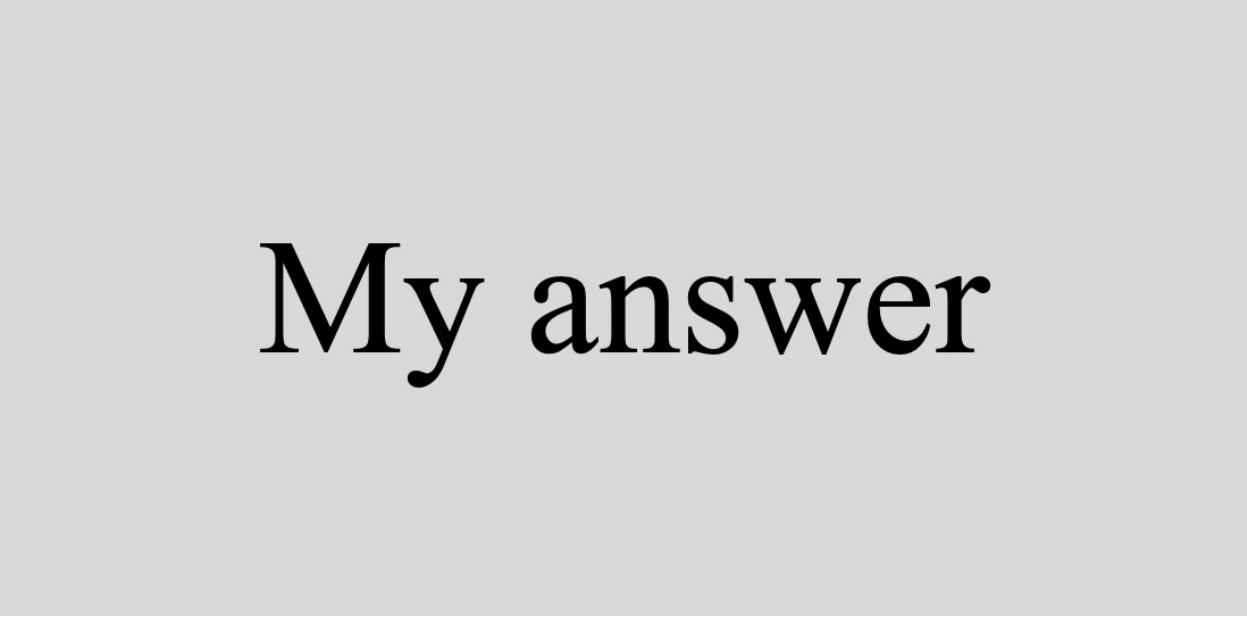

Figure 3: My caption

ANSWER:

We exponentiate both sides:

$$\frac{p(X)}{1-p(X)} = \exp(\beta_0 + \beta_1 X).$$

Then, we solve for $p(X)$:

$$p(X) = [1 - p(X)] \exp(\beta_0 + \beta_1 X).$$
$$p(X) = \exp(\beta_0 + \beta_1 X) - p(X) \exp(\beta_0 + \beta_1 X).$$
$$p(X) + p(X) \exp(\beta_0 + \beta_1 X) = \exp(\beta_0 + \beta_1 X).$$
$$p(X)[1 + \exp(\beta_0 + \beta_1 X)] = \exp(\beta_0 + \beta_1 X).$$
$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}.$$

The expression above would suffice. However, we usually do the following to put it in the standard form of a sigmoid function:

$$\begin{aligned} p(X) &= \frac{\frac{\exp(\beta_0+\beta_1 X)}{\exp(\beta_0+\beta_1 X}}{\frac{1+\exp(\beta_0+\beta_1 X)}{\exp(\beta_0+\beta_1 X}} \\ &= \frac{1}{\frac{1}{\exp(\beta_0+\beta_1 X)} + 1}. \end{aligned}$$

Finally:

$$p(X) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X)]}.$$

Exercise 5: The Die

rubric={reasoning:4}

Consider a weighted 6-sided die with probabilities $p_1, \ldots, p_6$. Answer the following **in two or three sentences per item**:

5.1. What values of the $\{p_i\}$ $(i = 1, \ldots, 6)$ do lead to the highest entropy distribution? Is the answer unique? Why or why not?

5.2. What values of the $\{p_i\}$ $(i = 1, \ldots, 6)$ do lead to the lowest entropy distribution? Is the answer unique? Why or why not?

ANSWER:

> **5.1.** *A uniform PMF (i.e., $p_1 = \cdots = p_6 = \frac{1}{6}$) leads to the highest entropy. It is a unique answer since this indicates we will have a uniform uncertainty over the whole range of possible outcomes.*
>
> **5.2.** *A PMF with probability 1 for a single outcome leads to the lowest entropy. It could be any of the six outcomes; thus, it is not a unique answer. We are basically reducing the uncertainty to a single outcome.*

Exercise 6: Choose the Distribution

6.1.

rubric={autograde:5}

For each of the following definitions of a random variable, select the appropriate discrete distribution family from these ones: `"Bernoulli"`, `"Binomial"`, `"Poisson"`, or `"Categorical"` (where "categorical" is any discrete distribution with a finite amount of outcomes).

> **Heads-up**: We are actually asking for the *smallest* family here. For example, the family of Bernoulli distributions is contained within the family of Binomial distributions (by taking $n = 1$ trials). Still, if a random variable here belongs to the Bernoulli family, we expect you to respond with `"Bernoulli"`.

6.1.1. The outcome of a coin flip.

6.1.2. The outcome of a 6-sided die roll.

6.1.3. The number of heads in 10 coin flips.

6.1.4. The number of times we saw a 6 out of 10 die rolls.

6.1.5. The number of times a professor says *"um"* in a lecture.

Assign your **answers as strings with the name of the distribution between quotation marks** to `answer6_1_1`, `answer6_1_2`, `answer6_1_3`, `answer6_1_4`, and `answer6_1_5`. Moreover, run the test below to validate your answers.

```
answer6_1_1 <- NULL
answer6_1_2 <- NULL
answer6_1_3 <- NULL
answer6_1_4 <- NULL
answer6_1_5 <- NULL

# BEGIN SOLUTION
answer6_1_1 <- "Bernoulli"
answer6_1_2 <- "Categorical"
answer6_1_3 <- "Binomial"
answer6_1_4 <- "Binomial"
answer6_1_5 <- "Poisson"
# END SOLUTION
```

6.2.

rubric={reasoning:2}

When discussing the binomial distribution, we often use the letters p, n and x. For example, let X be a binomial random variable

$$X \sim \text{Binomial}(n, p).$$

> **Heads-up:** *The symbol "$\sim$" means "is distributed as."*

Its PMF is defined as:

$$P(X = x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Are p, n, and x all parameters of the binomial distribution? Or is one of them different from the other two? Explain **in two or three sentences**.

ANSWER:

No, x is the argument of the PMF

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

Note n and p are the parameters that define the binomial distribution, Binomial(n, p).

Exercise 7

rubric={autograde:5}

Consider three events

$$A, B, \text{and } C.$$

Suppose that

$$P(A) = 0.12,$$
$$P(B) = 0.07,$$
$$P(C) = 0.05,$$
$$P(A \cup B) = 0.13,$$
$$P(A \cap C) = 0.04,$$
$$P(B \cap C) = 0.01,$$

and

$$P(A \cap B \cap C) = 0.01.$$

Compute the following probabilities:

7.1. $P(A^c)$.

7.2. $P(A \cap B)$.

7.3. $P(A \cap B \cap C^c)$.

7.4. $P[(A \cup B \cup C)^c]$.

7.5. $P(A \cup B \cup C^c)$.

> **Hint:** It may help to draw a Venn diagram of the events. **Check this resource**. Suppose you have an event A, note that notation A^c is equivalent to A'.

Assign your **numeric answers** to `answer7_1`, `answer7_2`, `answer7_3`, `answer7_4`, and `answer7_5`. You do not need to provide all your procedure (i.e., equations!) to get full marks. Moreover, run the test below to validate your answers.

```
answer7_1 <- NULL
answer7_2 <- NULL
answer7_3 <- NULL
answer7_4 <- NULL
answer7_5 <- NULL

# BEGIN SOLUTION
answer7_1 <- 0.88
answer7_2 <- 0.06
answer7_3 <- 0.05
answer7_4 <- 0.86
answer7_5 <- 0.99
# END SOLUTION
```

ANSWER:

> *The reasoning behind the answers is the following:*

7.1. $P(A^c) = 1 - P(A) = 1 - 0.12 = 0.88.$

7.2. *We know that*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Hence

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.12 + 0.07 - 0.13 = 0.06.$$

7.3.

$$\begin{aligned} P(A \cap B \cap C^c) &= P(A \cap B) - P(A \cap B \cap C) \\ &= 0.06 - 0.01 \\ &= 0.05 \end{aligned}$$

7.4.

$$\begin{aligned} P[(A \cup B \cup C)^c] &= 1 - P(A \cup B \cup C) \\ &= 1 - [P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + \\ &\qquad P(A \cap B \cap C)] \\ &= 1 - (0.12 + 0.07 + 0.05 - 0.06 - 0.01 - 0.04 + 0.01) \\ &= 0.86. \end{aligned}$$

7.5.

$$\begin{aligned} P(A \cup B \cup C^c) =& P(A) + P(B) + P(C^c) - P(A \cap B) - P(B \cap C^c) - P(A \cap C^c) + \\ & P(A \cap B \cap C^c) \\ =& P(A) + P(B) + \underbrace{[1 - P(C)]}_{P(C^c)} - P(A \cap B) - \underbrace{[P(B) - P(B \cap C)]}_{P(B \cap C^c)} - \\ & \underbrace{[P(A) - P(A \cap C)]}_{P(A \cap C^c)} + \underbrace{[P(A \cap B) - P(A \cap B \cap C)]}_{P(A \cap B \cap C^c)} \\ =& 0.12 + 0.07 + (1 - 0.05) - 0.06 - (0.07 - 0.01) - (0.12 - 0.04) + (0.06 - 0.01) \\ =& 0.99. \end{aligned}$$

(Challenging) Exercise 8

rubric={autograde:2,reasoning:2}

Let X be a discrete random variable defined on the positive integers ($i = 1, 2, 3, \ldots$), with the following PMF:

$$P(X = i) = \frac{C}{i^2} \qquad \text{for some normalizing constant } C.$$

> **Heads-up:** It turns out that $C = \frac{6}{\pi^2} \approx 0.61$. We could have at least reasoned that $C < 1$, since otherwise we would have $P(X = 1) \geq 1$, which does not make sense.

Find the following quantities:

8.1. $P(X < 5)$.

8.2. $P(X > 2 \cup X < 4)$.

8.3. $P(X < 2 \cup X < 4)$.

8.4. $P(X > 2 \cup X > 4)$.

8.5. $P(X < 2 \cup X > 4)$.

8.6. $P(X > 2 \cap X < 4)$.

8.7. $P(X < 2 \cap X < 4)$.

8.8. $P(X > 2 \cap X > 4)$.

8.9. $P(X < 2 \cap X > 4)$.

8.10. $\mathbb{E}(X)$.

To get full marks on these questions, follow these instructions:

1. Provide all your procedure and/or reasoning (e.g., equations):
    a. You do not need to use **LaTex** to provide mathematical notation.
    b. Instead, you might work on your written answer on a separate piece of paper and take a picture of it.
    c. Then, you have to put this image in the folder `img` which is part of your lab GitHub repo.
    d. Finally, within this `R` markdown, use the following syntax: `![My caption](img/my_answer_8)`, where `my_answer_8` is your image's name. The output is the following:

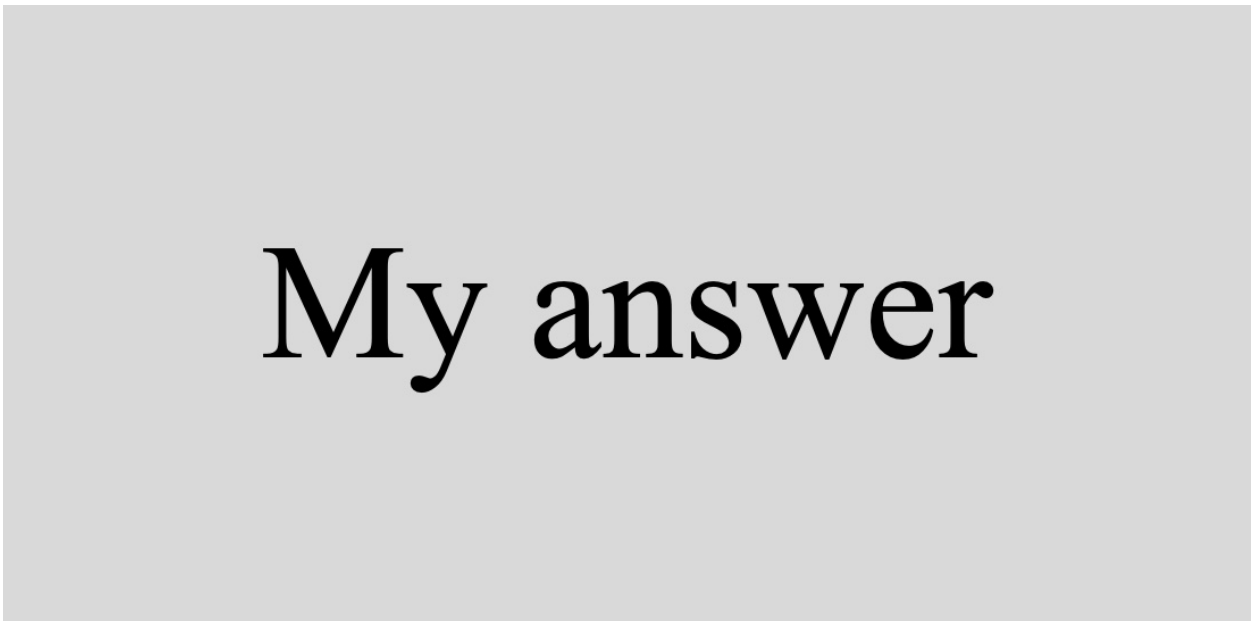

Figure 4: My caption

2. Assign your **numeric answers** to `answer8_1`, `answer8_2`, ..., `answer8_9`, and `answer8_10`. **Code your computations directly in each answer if necessary.** Moreover, run the test below to validate your answers.

ANSWER:

The reasoning behind the answers is the following:

8.1.

$$P(X < 5) = \frac{C}{1^2} + \frac{C}{2^2} + \frac{C}{3^2} + \frac{C}{4^2} = 0.87.$$

8.2.

$$\begin{aligned} P(X > 2 \cup X < 4) &= P(X > 2) + P(X < 4) - P(X > 2 \cap X < 4) \\ &= \underbrace{[1 - P(X \leq 2)]}_{P(X>2)} + P(X < 4) - \underbrace{P(X = 3)}_{P(X>2 \cap X<4)} \\ &= \left(1 - \frac{C}{1^2} - \frac{C}{2^2}\right) + \left(\frac{C}{1^2} + \frac{C}{2^2} + \frac{C}{3^2}\right) - \frac{C}{3^2} \\ &= 1. \end{aligned}$$

8.3.

$$\begin{aligned} P(X < 2 \cup X < 4) &= P(X < 2) + P(X < 4) - P(X < 2 \cap X < 4) \\ &= P(X < 2) + P(X < 4) - \underbrace{P(X = 1)}_{P(X<2 \cap X<4)} \\ &= \frac{C}{1^2} + \left(\frac{C}{1^2} + \frac{C}{2^2} + \frac{C}{3^2}\right) - \frac{C}{1^2} \\ &= \frac{C}{1^2} + \frac{C}{2^2} + \frac{C}{3^2} \\ &= 0.83. \end{aligned}$$

8.4.

$$\begin{aligned} P(X > 2 \cup X > 4) &= P(X > 2) + P(X > 4) - P(X > 2 \cap X > 4) \\ &= \underbrace{[1 - P(X \leq 2)]}_{P(X>2)} + \underbrace{[1 - P(X \leq 4)]}_{P(X>4)} - \underbrace{[1 - P(X \leq 4)]}_{P(X>2 \cap X>4)} \\ &= 1 - \frac{C}{1^2} - \frac{C}{2^2} \\ &= 0.24. \end{aligned}$$

8.5.

$$\begin{aligned} P(X < 2 \cup X > 4) &= P(X < 2) + P(X > 4) - P(X < 2 \cap X > 4) \\ &= P(X < 2) + [1 - P(X \leq 4)] - P(X < 2 \cap X > 4) \\ &= \frac{C}{1^2} + \left(1 - \frac{C}{1^2} - \frac{C}{2^2} - \frac{C}{3^2} - \frac{C}{4^2}\right) - 0 \\ &= 0.74. \end{aligned}$$

8.6.

$$P(X > 2 \cap X < 4) = P(X = 3) = \frac{C}{3^2} = 0.07.$$

8.7.

$$P(X < 2 \cap X < 4) = P(X \leq 1) = \frac{C}{1^2} = 0.61.$$

8.8.

$$\begin{aligned} P(X > 2 \cap X > 4) &= P(X > 4) \\ &= 1 - P(X \leq 4) \\ &= 1 - \left(\frac{C}{1^2} + \frac{C}{2^2} + \frac{C}{3^2} + \frac{C}{4^2}\right) \\ &= 0.13. \end{aligned}$$

8.9. *The intersection is an empty subset:*

$$P(X < 2 \cap X > 4) = 0$$

8.10.

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} iP(X = i) = \sum_{i=1}^{\infty} \frac{C}{i} = \infty$$

```
C  <- 6 / pi^2 # Where appropriate, write your answers in terms of C.

answer8_1 <- NULL
answer8_2 <- NULL
answer8_3 <- NULL
answer8_4 <- NULL
answer8_5 <- NULL
answer8_6 <- NULL
answer8_7 <- NULL
answer8_8 <- NULL
answer8_9 <- NULL
answer8_10 <- NULL
```

```
# BEGIN SOLUTION
answer8_1 <- C * ((1 / 1^2) + (1 / 2^2) + (1 / 3^2) + (1 / 4^2))
answer8_2 <- 1
answer8_3 <- C * ((1 / 1^2) + (1 / 2^2) + (1 / 3^2))
answer8_4 <- 1 - (C * ((1 / 1^2) + (1 / 2^2)))
answer8_5 <- 1 - (C * ((1 / 2^2) + (1 / 3^2) + (1 / 4^2)))
answer8_6 <- C / 3^2
answer8_7 <- C
answer8_8 <- 1 - C * ((1 / 1^2) + (1 / 2^2) + (1 / 3^2) + (1 / 4^2))
answer8_9 <- 0
answer8_10 <- Inf
# END SOLUTION
```