DS 1000B – Assignment 4

Total Marks: 100

Due date: Monday, Mar 9th, 2026 at 8:00 PM

Submission Platform: All assignments must be submitted via Gradescope. **File Format:** Submit a **single** PDF file containing all of your work. Please note that Gradescope only displays your most recent submission and this is the version that will be graded. **You will receive a grade of zero in each case where**

- Submission is not in PDF format.
- Questions have no pages assigned to them on Gradescope(i.e. did not submit anything for that question).
- Submission is illegible (e.g. blurry, too small to read comfortably without zooming in).

You must submit the following as a single PDF file:

- Part 1 – Written responses (these may be handwritten or typed). Multiple choice questions do not require work to be shown.

  Unless otherwise specified, please show your work.

- Part 2 – Python coding exercises with some written responses. *Acceptable submission formats:*
  - Screenshot showing both the code cell and the corresponding output (from Google Colab, a local Jupyter notebook, or another Python environment).
  - Copy-pasted code with the output clearly labeled below it in a screenshot or text (where applicable).

  Do not write code or its output by hand (e.g., pencil, pen, or stylus). All code and outputs must be generated and shown directly from your Python environment.

Individual Work: Each student must submit their own original work. You may discuss questions with your classmates, but you must write up your solutions independently.

Part 1 – Written Responses

Northern Lights or False Alarm? [8 marks]

A national weather service is testing a new aurora borealis prediction model. Over the course of one winter, researchers recorded data from 400 nights at high-latitude observation stations across Canada. Each night was independently classified along two criteria:

- **Actual outcome**: Aurora Visible (A) or No Aurora (N)
- **Model prediction**: Forecast Aurora (F) or Predicted Clear (C)

The results are summarized in the following table:

	Forecast Aurora (F)	Predicted Clear (C)	Total
Aurora Visible (A)	150	30	180
No Aurora (N)	60	160	220
Total	210	190	400

One night is selected at random from the 400.

Q1. (a) (1 mark) List the sample space S for this experiment, where each outcome describes a night's actual aurora activity *and* the model's prediction.

(b) (2 marks) Are the events A (Aurora Visible) and F (Forecast Aurora) mutually exclusive? Justify your answer.

(c) (1 mark) Find $P(A)$, the probability that the randomly selected night had a visible aurora.

(d) (2 marks) Find $P(F \cup A)$, the probability that the model forecast an aurora **or** an aurora was actually visible (or both).

(e) (2 marks) Find $P(N \cap F)$. Interpret this probability in the context of aurora forecasting.

Food Delivery Orders

[8 marks]

A food delivery service tracked the number of orders placed by customers on a given day. Let X be the random variable representing the number of delivery orders a randomly selected customer places in one day. The probability distribution of X is given below:

x	0	1	2	3	4
$P(X = x)$	0.35	0.30	0.20	0.10	0.05

Q2. (a) (2 marks) State and verify the conditions that must be satisfied for this to be a valid probability distribution.

(b) (2 marks) What is the probability that a randomly selected customer places at least 2 orders in a day? That is, find $P(X \geq 2)$.

(c) (2 marks) What is the probability that a customer places fewer than 3 orders? That is, find $P(X < 3)$.

(d) (2 marks) What is the probability that a customer places between 1 and 3 orders (inclusive)? That is, find $P(1 \leq X \leq 3)$.

The Four-Day Work Week [18 marks]

A technology company is piloting a 4-day work week program. Researchers tracked 300 teams to see if the new schedule affected their ability to meet project deadlines. For a randomly selected team, let A, B, and C represent:

$$\begin{aligned} A &= \text{Team meets all project deadlines on time} \\ B &= \text{Team is assigned to the 4-Day Work Week pilot} \\ C &= \text{Team is working on a ``High Complexity'' / Legacy project} \end{aligned}$$

Based on the study data, the following probabilities were observed:

$$P(A) = 0.76, \quad P(B) = 0.30, \quad P(C) = 0.40$$

$$P(A \cap B) = 0.19, \quad P(A \cap C) = 0.19, \quad P(B \cap C) = 0.27, \quad P(A \cap B \cap C) = 0.16$$

Q3. (a) (4 marks) Draw a Venn diagram showing events A, B, and C with all their intersections. In your diagram:

- Label the three circles (or closed curves) A, B, and C
- Write the probability for each of the 8 disjoint regions (the 7 regions within the three circles plus the region outside all circles)

Note: Your Venn diagram should display 8 probabilities total, one for each distinct region. You do not have to label each region (except for A, B, and C).

(b) (2 marks) What is the probability that exactly one of the three events A, B, or C occurs?

(c) (2 marks) What is the probability that at least two of the three events A, B, or C occur?

(d) (2 marks) Are the events A (Meeting Deadlines) and B (4-Day Work Week) independent? Justify your answer mathematically.

(e) (2 marks) Calculate $P(A \mid B)$ and $P(A \mid B^c)$. Which group has a higher overall success rate in meeting deadlines?

(f) (2 marks) Among teams working on High Complexity projects (C), calculate:

- The probability of meeting deadlines for teams on the 4-Day Work Week (B).
- The probability of meeting deadlines for teams on the standard schedule (B^c).

Which group has a higher success rate among High Complexity projects?

(g) (2 marks) Among teams working on Low Complexity projects (C^c), calculate:

- The probability of meeting deadlines for teams on the 4-Day Work Week (B).
- The probability of meeting deadlines for teams on the standard schedule (B^c).

Which group has a higher success rate among Low Complexity projects?

(h) (2 marks) Based on your analysis, should the company continue the 4-Day Work Week pilot? Explain your reasoning.

Randomized Response Methodology [36 marks]

As we will see in Chapter 8, observational studies frequently use sample data to estimate population parameters. A significant challenge in survey methodology is the reduction of non-sampling error caused by social desirability bias. When queried regarding sensitive topics, such as infidelity, respondents are often inclined to withhold truthful answers due to social stigma. To estimate the true prevalence of a sensitive attribute while strictly preserving respondent privacy, researchers employ Randomized Response Techniques. This method introduces a stochastic component to the reporting process, affording participants plausible deniability. **The Protocol:**

A participant is told to toss a fair coin in private and observe its outcome. They are instructed **not** to reveal the coin toss result and to adhere to the following decision rule regarding the outcome of the coin flip:

- Heads: Answer the question truthfully (*Have you ever cheated on a romantic partner?*).
- Tails: Automatically answer "YES" (regardless of the truth).

Because a "YES" response may originate from either a truthful admission (on Heads) or a forced response (on Tails), the researcher cannot distinguish the source of any individual answer (as long as the participant does not reveal the coin toss outcome). **Assumption:** Assume that the coin tosses are independent of the participants' true status regarding infidelity and that all participants follow the instructions correctly. **Notation:** Let C denote the event that a participant possesses the sensitive attribute (has committed infidelity). Let Y denote the event that a participant responds "Yes". Moreover, let $p = P(C)$ represent the true, but unknown, prevalence of infidelity in the population.

Q4. (a) (5 marks) Draw a probability tree diagram modeling this randomized response mechanism. Be sure to annotate the branches with the appropriate conditional probabilities.

(b) (5 marks) Suppose that are out 1000 person study resulted in 600 respondents which responded in the affirmative "YES". Estimate the prevalence of individuals who committed infidelity in the population.

(c) (7 marks) Calculate the probability that a respondent who answers "YES" has actually committed infidelity, i.e., find $P(C \mid Y)$. If we are interested in increasing a respondent's privacy, would we want $P(C \mid Y)$ to be higher or lower? Explain.

(d) (4 marks) Find $P(C \mid Y^c)$. What does a "No" response reveal about the respondent?

(e) (7 marks) Suppose the protocol is modified so that the coin has probability q of landing Heads (where $0 < q < 1$), rather than $q = \frac{1}{2}$. Express the estimated prevalence p as a function of the observed proportion $P(Y)$ and the design parameter q.

(f) (4 marks) Using your answer from part (e), for what values of q is estimation of $p = P(C)$ impossible, and why?

(g) (4 marks) Determine whether the events C and Y are independent. Justify your answer using the formal definition of independence (i.e. $P(A \cap B) = P(A)P(B)$).

Part 2 – Python

Simulation and Probability [30 marks]

While we can calculate probabilities theoretically (like in Part 1), in Data Science we often use simulation to approximate probabilities for complex events. Although the following questions involve rolling dice, the same principles apply to more complicated scenarios where theoretical calculations are intractable.

Q5. Consider an experiment where you roll two fair six-sided dice and calculate their sum.

(a) (3 marks) 🐍 Import the `numpy` library and set the random seed to your student number (e.g., `np.random.seed(20234567)` if your student number is 20234567). Then, use `np.random.randint` to simulate rolling one die 5 times and print the results.

(b) (3 marks) 🐍 Now simulate rolling *two* dice 5 times each. Store the results in variables `die1` and `die2`, then calculate and print their sums.

(c) (6 marks) 🐍 Write a function called `simulate_dice_rolls(n)` that:

- Takes an integer n as input
- Simulates rolling two dice n times
- Returns the sums as a numpy array

Test your function with $n = 10$ and print the results.

(d) (8 marks) 🐍 Use your function to simulate $n = 100$ rolls and $n = 10,000$ rolls. For each simulation, calculate the empirical probability of getting a sum of 7 (i.e., number of 7s divided by n) and print both probabilities.

(e) (10 marks) 🐍 Finally, simulate $n = 1,000,000$ rolls and calculate the empirical probability. Compare the results for $n = 100$, $n = 10,000$, and $n = 1,000,000$ to the theoretical probability of approximately 0.1667.

Aside: What we see at play here is the so-called Law of Large Numbers.