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 labelled below it in a screenshot or text (where applicable).

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. Provided your submission adheres to the requirements outlined above, the method you use to generate the document is at your discretion.

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.

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.

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.

Solution: Each night falls into exactly one combination of outcome and prediction:

$$S = \{(A, F),\ (A, C),\ (N, F),\ (N, C)\}$$

The sample space contains 4 outcomes.

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

Solution: No, A and F are **not** mutually exclusive. Two events are mutually exclusive if they cannot occur simultaneously, i.e., $A \cap F = \varnothing$. Here, $A \cap F$ corresponds to nights where aurora was both visible *and* forecast by the model, and there are 150 such nights. As

$$P(A \cap F) = \frac{150}{400} = 0.375 \neq 0,$$

the events can (and do) occur together.

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

Solution:

$$P(A) = \frac{180}{400} = \boxed{0.45}$$

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

Solution: By the addition rule,

$$\begin{aligned} P(F \cup A) &= P(F) + P(A) - P(F \cap A) \\ &= \frac{210}{400} + \frac{180}{400} - \frac{150}{400} \\ &= \frac{240}{400} \\ &= \boxed{0.60} \end{aligned}$$

Equivalently, $F \cup A$ includes every night except those with no aurora *and* a clear prediction, so

$$P(F \cup A) = 1 - \frac{160}{400} = 0.60.$$

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

Solution:

$$P(N \cap F) = \frac{60}{400} = \boxed{0.15}$$

This is the probability of a *false alarm*: the model forecast an aurora, but no aurora was actually visible that night. 15% of all nights in the sample were false alarms, meaning eager observers may have been sent outside in the cold for nothing.

> **Comment**
> Students do not need to use the term "false alarm" to receive full marks

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.

Solution: A valid probability distribution must satisfy:

i. All probabilities must be between 0 and 1: $0 \leq P(X = x) \leq 1$ for all x.

Check: $0.35, 0.30, 0.20, 0.10, 0.05$ are all between 0 and 1.

ii. The probabilities must sum to 1: $\sum P(X = x) = 1$.

Check: $0.35 + 0.30 + 0.20 + 0.10 + 0.05 = 1.00$

Both conditions are satisfied, so this is 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 $P(X \geq 2)$.

Solution:

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4)$$
$$= 0.20 + 0.10 + 0.05 = \boxed{0.35}$$

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

Solution:

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$$
$$= 0.35 + 0.30 + 0.20 = \boxed{0.85}$$

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

Solution:

$$P(1 \leq X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3)$$
$$= 0.30 + 0.20 + 0.10 = \boxed{0.60}$$

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:

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

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

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

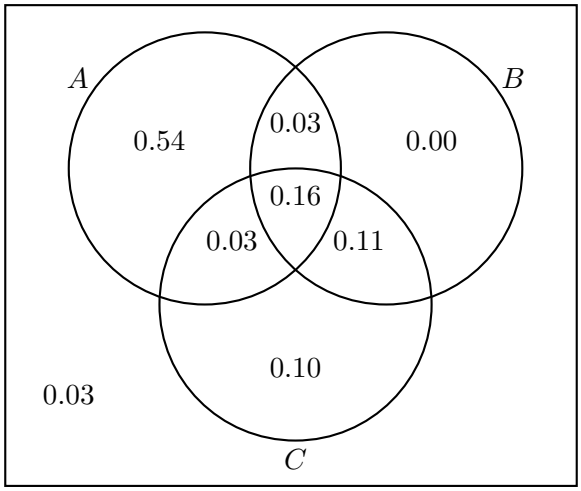
$$P(A \cap B) = 0.19, \quad P(A \cap C) = 0.19, \quad P(B \cap C) = 0.27, \quad P(A \cap B \cap C) = 0.16$$

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).

Solution:



Comment

Students may place the given probabilities (e.g., $P(A \cap B) = 0.19$) directly into the Venn diagram regions without subtracting the triple intersection. Apply a minor penalty reminding them how to label the Venn diagram regions correctly (e.g., $P(A \cap B) - P(A \cap B \cap C)$ for the A and B only region).

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

Solution: The event “exactly one of A, B, or C occurs" corresponds to the union of three disjoint events: A only, B only, and C only.

From the Venn diagram in part (a), we have:

$$P(A \text{ only}) = 0.54, \quad P(B \text{ only}) = 0.00, \quad P(C \text{ only}) = 0.10.$$

Therefore:

$$P(\text{exactly one event}) = 0.54 + 0.00 + 0.10 = \boxed{0.64}$$

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

Solution: The event "at least two of A, B, or C occur" includes:

- Exactly two events occur (from the previous part), or
- All three events occur.

From the Venn diagram in part (a):

$$P(\text{exactly two}) = 0.17, \quad P(A \cap B \cap C) = 0.16.$$

Therefore:

$$P(\text{at least two}) = 0.17 + 0.16 = \boxed{0.33}$$

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

Solution: Two events are independent if $P(A \cap B) = P(A) \cdot P(B)$.

$$P(A) \cdot P(B) = 0.76 \times 0.30 = 0.228$$

$$P(A \cap B) = 0.19$$

Since $0.19 \neq 0.228$, the events are not independent. Knowing a team is on the 4-Day Work Week schedule changes the probability that they meet their deadlines.

(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?

Solution: We calculate the conditional probabilities:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.19}{0.30} \approx \boxed{0.63}$$

$$P(A \mid B^c) = \frac{P(A) - P(A \cap B)}{1 - P(B)} = \frac{0.76 - 0.19}{0.70} = \frac{0.57}{0.70} \approx \boxed{0.81}$$

Teams on the standard 5-day week (B^c) have a higher rate of meeting deadlines (81% vs. 63%) when looking at the aggregate data.

(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?

Solution: Among teams with High Complexity projects (event C):

$$P(A \mid B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{0.16}{0.27} \approx \boxed{0.59}$$

$$P(A \mid B^c \cap C) = \frac{P(A \cap C) - P(A \cap B \cap C)}{P(C) - P(B \cap C)} = \frac{0.19 - 0.16}{0.40 - 0.27} = \frac{0.03}{0.13} \approx \boxed{0.23}$$

Among teams working on High Complexity projects, those on the 4-Day Work Week have a much higher success rate (59% vs. 23%).

(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?

Solution: Among teams without High Complexity projects (C^c):

For 4-Day Week teams:

$$P(B \cap C^c) = P(B) - P(B \cap C) = 0.30 - 0.27 = 0.03$$

$$P(A \cap B \cap C^c) = P(A \cap B) - P(A \cap B \cap C) = 0.19 - 0.16 = 0.03$$

$$P(A \mid B \cap C^c) = \frac{0.03}{0.03} = \boxed{1.00}$$

For standard schedule teams:

$$P(B^c \cap C^c) = P(C^c) - P(B \cap C^c) = 0.60 - 0.03 = 0.57$$

$$P(A \cap B^c \cap C^c) = P(A \text{ only}) = 0.54$$

$$P(A \mid B^c \cap C^c) = \frac{0.54}{0.57} \approx \boxed{0.95}$$

Among teams working on Low Complexity projects, those on the 4-Day Work Week have a slightly higher success rate (100% vs. 95%).

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

Solution: Yes, the company should continue the pilot. This is an example of Simpson's Paradox:

- Overall, the 4-Day week looks harmful (63% success vs 81% standard).
- Subgroups: The 4-Day week performs better on both High Complexity projects (59% vs 23%) and Low Complexity projects (100% vs 95%).

The aggregate data is misleading because the 4-Day schedule was disproportionately assigned to the hardest projects ($P(C|B) = 0.90$), which inherently have lower success rates.

> **Comment**
> Students do not have to mention Simpson's paradox, but there does need to be justification that the aggregate data is misleading and that the 4-Day week performs better within both subgroups.

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:

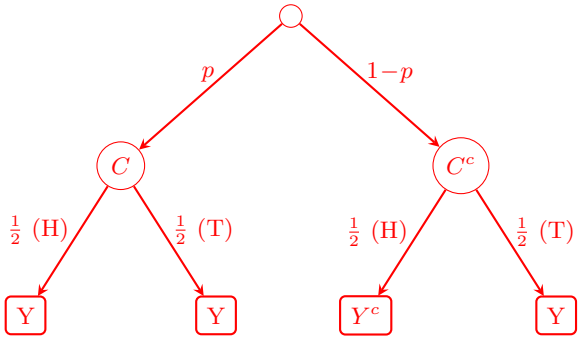
- Heads: Answer the question truthfully (*Have you ever cheated on a romantic partner?*).
- Tails: Automatically answer "YES" (regardless of the truth).

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.

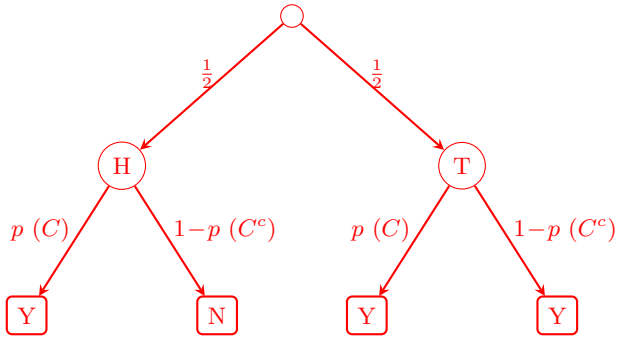
Q4. (a) (5 marks) Draw a probability tree diagram modeling this randomized response mechanism. Be sure to annotate the branches with the appropriate conditional probabilities.

Solution:

Approach 1: Status first, then coin



Approach 2: Coin first, then status



A participant in the Cheater (C) group always responds "Yes": truthfully on Heads and via the forced response on Tails. The coin toss and true status are independent, so the branching order can be reversed (Approach 1 vs. Approach 2) without affecting the probabilities.

> **Comment**
> Students may set $p = 0.20$ in the above based on what they find in part (b). Please accept this but remind them that the tree diagram should be general and not depend on subsequent components of the question.

(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.

Solution: By the Law of Total Probability,

$$P(Y) = P(Y \mid C)\, P(C) + P(Y \mid C^c)\, P(C^c)$$

Substituting the design probabilities and observed data:

$$0.60 = (1.0)(p) + (0.5)(1-p)$$

Solving the linear equation for p:

$$\begin{aligned} 0.60 &= p + 0.5 - 0.5p \\ 0.60 &= 0.5p + 0.5 \\ 0.10 &= 0.5p \\ p &= \boxed{0.20} \end{aligned}$$

Conclusion: The estimated prevalence of infidelity in the population is 20%.

(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.

Solution: We calculate the posterior probability $P(C \mid Y)$ via Bayes' Theorem:

$$P(C \mid Y) = \frac{P(Y \mid C)\,P(C)}{P(Y)}$$

Substituting the parameters ($p = 0.20$, $P(Y \mid C) = 1.0$, and $P(Y) = 0.60$):

$$P(C \mid Y) = \frac{(1.0)(0.20)}{0.60} = \frac{0.20}{0.60} = \boxed{\tfrac{1}{3} \approx 0.33}$$

Given an affirmative response (Y), there is only a 33% probability that the participant actually committed infidelity.

Conversely, there is a 67% probability that the respondent did not commit infidelity and was forced to answer "Yes" by the randomization device (the coin toss). This high level of uncertainty regarding the source of the response ensures plausible deniability, thereby protecting respondent privacy.

Comment

Students may also express $P(C \mid Y)$ in terms of the general prevalence p (without plugging in $p = 0.20$ from part (b)):

$$P(C \mid Y) = \frac{P(Y \mid C)\,P(C)}{P(Y)} = \frac{(1)(p)}{\frac{1}{2}p + \frac{1}{2}} = \frac{p}{\frac{1+p}{2}} = \frac{2p}{1+p}$$

Both forms are acceptable for full marks.

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

Solution: We apply Bayes' Theorem to compute $P(C \mid Y^c)$:

$$P(C \mid Y^c) = \frac{P(Y^c \mid C)\,P(C)}{P(Y^c)}$$

From the tree diagram, a cheater always says "Yes" (truthfully on Heads, forced on Tails), so $P(Y^c \mid C) = 0$. Therefore:

$$P(C \mid Y^c) = \frac{(0)(0.20)}{0.40} = \boxed{0}$$

A "No" response can only arise via the single path $C^c \cap$Heads, meaning it is impossible for an individual who comitted infidelity to answer "No" under this protocol. A respondent who answers "No" effectively discloses that they did not commit infidelity with certainty.

(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.

Solution: With a biased coin, the conditional probabilities become:

$$P(Y \mid C) = q + (1 - q) = 1$$
$$P(Y \mid C^c) = 0 + (1 - q) = 1 - q$$

By the law of total probability:

$$P(Y) = (1)(p) + (1 - q)(1 - p) = p + (1 - q) - (1 - q)p$$

Collecting terms in p:

$$P(Y) = p[1 - (1 - q)] + (1 - q) = qp + (1 - q)$$

> **Comment**
> Some students might keep the assumption that $P(Y) = 0.60$ from part (b) and solve for p in terms of q:
>
> $$0.60 = qp + (1 - q)$$
>
> $$0.60 - (1 - q) = qp$$
>
> $$p = \frac{0.60 - (1 - q)}{q} = \frac{q - 0.40}{q} = 1 - \frac{0.40}{q}$$
>
> This is also acceptable for full marks, but do note that there is benefit in the generality from
>
> $$p = \frac{P(Y) - (1 - q)}{q}.$$

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

Solution: When $q = 0$, every participant is forced to answer "Yes" regardless of truth (the coin always lands Tails). In this case, $P(Y) = 1$ for any value of p, and the formula yields $\frac{0}{0}$, which is undefined. The observed data carry no information about p.

Aside: This reveals a tradeoff in the protocol design. Decreasing q increases respondent privacy (more forced "Yes" responses provide stronger deniability), but reduces statistical precision[1] (the signal about p becomes weaker). Conversely, increasing q improves estimation quality at the cost of weaker privacy protection.

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

Solution: Two events C and Y are independent if and only if $P(C \mid Y) = P(C)$, i.e., observing Y does not change the probability of C. From part (b), $P(C) = 0.20$. From part (c), $P(C \mid Y) = \frac{1}{3} \approx 0.33$. Since $P(C \mid Y) = \frac{1}{3} \neq 0.20 = P(C)$, the events C and Y are **not independent**. □

Interpretation: Learning that a respondent answered "Yes" increases the probability that they committed infidelity (from 20% to 33%). This dependence is precisely what makes the randomized response technique useful: the observed responses carry a statistical signal about the sensitive attribute, even though individual responses remain ambiguous.

[1] outside the scope of our course

Comment

Students may alternatively use the formal definition of independence directly:

$$P(C \cap Y) = P(C) \cdot P(Y) \implies (0.20)(0.60) = 0.12 \neq P(C \cap Y) = 0.20$$

Since $P(C \cap Y) = P(Y \mid C)\, P(C) = (1)(0.20) = 0.20 \neq 0.12$, the events are not independent.

In general, one can check that C and Y are independent if and only if $p = 1$ (everyone committed infidelity, so Y carries no information) or $q = 0$ (the coin always lands Tails, forcing everyone to say "Yes", so again Y is uninformative). Both cases are degenerate and uninteresting for estimation purposes.

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.

Solution:

Python Code

```
import numpy as np
np.random.seed(20234567)
die_rolls = np.random.randint(1, 7, size=5)
print(die_rolls)
```

Output

```
[4 4 5 4 1]
```

(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.

Solution:

Python Code

```
die1 = np.random.randint(1, 7, size=5)
die2 = np.random.randint(1, 7, size=5)

sums = die1 + die2
print(f"Die 1: {die1}")
print(f"Die 2: {die2}")
print(f"Sums: {sums}")
```

Output

```
Die 1:  [2 4 6 6 3]
Die 2:  [3 2 4 5 5]
Sums:  [ 5 6 10 11 8]
```

(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.

Solution:

Python Code

```
def simulate_dice_rolls(n):
  """ Roll die 1 n times (1 to 6) """
  die1 = np.random.randint(1, 7, size=n)
  """ Roll die 2 n times """
  die2 = np.random.randint(1, 7, size=n)
  """ Calculate sum """
  sums = die1 + die2
  return sums


test_rolls = simulate_dice_rolls(10)
print(test_rolls)
```

Output

```
[ 7 6 5 8 10 4 3 6 12 12]
```

(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.

Solution:

Python Code

```
rolls_100 = simulate_dice_rolls(100)
prob_100 = np.sum(rolls_100 == 7) / 100
print(f"Empirical Probability (n=100): {prob_100:.4f}")

rolls_10k = simulate_dice_rolls(10000)
prob_10k = np.sum(rolls_10k == 7) / 10000
print(f"Empirical Probability (n=10,000): {prob_10k:.4f}")
```

Output

```
Empirical Probability (n=100):  0.1700
Empirical Probability (n=10,000):  0.1700
```

(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.

Solution:

Python Code

```
rolls_1m = simulate_dice_rolls(1000000)
prob_1m = np.sum(rolls_1m == 7) / 1000000
print(f"Empirical Probability (n=1,000,000): {prob_1m:.4f}")
```

Output

```
Empirical Probability (n=1,000,000):  0.1665
```

The empirical probability with $n = 1,000,000$ (approx. 0.1668) is closest to the theoretical probability (0.1667). The value for $n = 100$ likely deviates the most (e.g., 0.12 or 0.20), while $n = 10,000$ is much closer but still has slight variation.

This illustrates the Law of Large Numbers: as the sample size increases, the sample proportion converges to the true population probability. With more trials, random fluctuations average out and the empirical result approaches the theoretical value.