DS 1000B – Assignment 5

Total Marks: 100

Due date: Monday, April 6th, 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.

Part 1 – Written Responses

Student Sleep Patterns and At-Risk Screening [22 marks]

At a university, 30% of students are varsity athletes and 70% are non-athletes. Nightly sleep duration is approximately normally distributed within each group:

- Athletes: $\mu = 6.5$ hours, $\sigma = 1.0$ hours
- Non-athletes: $\mu = 7.5$ hours, $\sigma = 1.2$ hours

The campus health centre flags any student sleeping fewer than 5.5 hours per night as "at risk."

Q1. (a) (2 marks) What proportion of non-athletes sleep more than 9.0 hours per night? Show your work, including the relevant z-score.

(b) (2 marks) A varsity athlete's nightly sleep is at the 20th percentile of the athletes' sleep distribution. How many hours per night does this athlete sleep?

(c) (4 marks) The health centre is also designing a separate alert for students whose sleep falls in the most extreme 10% within their group (the lowest 5% or the highest 5%). For varsity athletes, what is the middle 90% range of nightly sleep durations? (This is the range that would *not* trigger the alert.)

(d) (4 marks) Suppose Andre is a varsity athlete who sleeps 8.0 hours per night. Bob, a non-varsity student in the same residence hall sleeps 5.6 hours per night. Determine which student's sleep duration is more unusual relative to their group.

(e) (4 marks) Using the original at-risk criterion (fewer than 5.5 hours per night), calculate the probability that a randomly selected varsity athlete is flagged. Then calculate the same probability for a randomly selected non-athlete.

(f) (6 marks) Given that a student is flagged as "at risk" (sleeps fewer than 5.5 hours), find the probability that this student is a varsity athlete.

Weekly Exercise at Western University [18 marks]

Suppose Western University seeks to estimate the average weekly time its students spend on physical exercise. Assume the true population mean is $\mu = 15$ hours per week, with a population standard deviation of $\sigma = 3.8$ hours.

Q2. (a) (5 marks) If a random sample of $n = 36$ students is selected, describe the sampling distribution of the sample mean $\bar{X}$. Include the mean, standard deviation (standard error), and shape.

(b) (4 marks) What is the probability that the sample mean weekly exercise time for a random sample of 36 students exceeds 16 hours?

(c) (2 marks) How would the standard error change if the sample size increased from 36 to 144? Calculate the new standard error.

(d) (3 marks) A health columnist claims that "taking a larger sample makes the exercise data you collect more normally distributed." Is this statement correct? Explain.

(e) (4 marks) A second university also studies weekly exercise among its students. Their population has the same mean ($\mu = 15$ hours) but a larger standard deviation of $\sigma = 5.7$ hours. Using Western's standard error from part (a) (with $n = 36$) as your target, what sample size would this university need so that their standard error equals Western's? What does this tell you about how population variability affects study design?

Campus Clinic Body Temperature [18 marks]

A campus health clinic records the resting body temperature of $n = 36$ randomly selected undergraduate students. The sample mean is $\bar{x} = 37.0\,^{\circ}\text{C}$. Assume the population standard deviation is $\sigma = 0.6\,^{\circ}\text{C}$.

Q3. (a) (4 marks) A wearable technology company's promotional materials claim that university students have an average resting body temperature of $37.3\,^{\circ}\text{C}$. Construct a 95% confidence interval for the true mean resting body temperature. State whether $37.3\,^{\circ}\text{C}$ is consistent or inconsistent with your interval, and explain your conclusion.

(b) (4 marks) A researcher is planning a follow-up study and wants the 95% confidence interval to be no wider than $0.2\,^{\circ}\text{C}$ total (i.e., a margin of error of at most $0.1\,^{\circ}\text{C}$). How many students must be sampled? A colleague suggests that doubling the current sample to $n = 72$ would be sufficient. Is the colleague correct?

(c) (3 marks) Using the same data from part (a), suppose a researcher constructs a 99% confidence interval instead. Explain how the width of the interval changes and why. Based on this wider interval, would the researcher still reach the same conclusion about the company's claim?

(d) (3 marks) A student interprets the confidence interval from part (a) as: "There is a 95% probability that the true mean resting temperature is between 36.804 °C and 37.196 °C." Is this interpretation correct? Provide the correct interpretation and explain the distinction.

(e) (4 marks) American collaborators request the results in degrees Fahrenheit. The conversion formula is $F = \frac{9}{5}C + 32$. Using z-scores, convert your 95% confidence interval from part (a) to degrees Fahrenheit. Show your work, and explain how the margin of error changes under this transformation.

Real-World Sampling Scenarios [10 marks]

Identifying Sampling Techniques

For each scenario below, identify the type of sampling method being used. Choose from: **Simple Random Sample (SRS)**, **Stratified Random Sample**, **Cluster Sample**, **Systematic Sample**, or **Convenience Sample**. Briefly justify your answer.

Q4. (a) (2 marks) The Milano Cortina 2026 Winter Olympics organizing committee wants to survey spectators about their experience. They divide attendees by sport category (ice sports, alpine skiing, cross-country/biathlon, sliding sports) and randomly select 150 spectators from each category.

(b) (2 marks) A streaming platform wants to assess user satisfaction with its new personalized recommendation system. They randomly select 8 Canadian cities and survey all subscribers in those cities.

(c) (2 marks) A campus newspaper polls students about dining hall food quality by approaching people sitting in the student centre on a Wednesday afternoon.

(d) (2 marks) A government agency wants to estimate average household income in Ontario. They assign every household a unique number, then use a random number generator to select 2000 households from the complete list.

(e) (2 marks) A hospital administrator reviews patient satisfaction by selecting every 25th patient admitted during the past year, starting from a randomly chosen record among the first 25.

The Biodegradable Cup Study [10 marks]

A Randomised Field Experiment

Q5. A coffee chain wants to test whether a new biodegradable cup design affects customer satisfaction. They recruit 120 customers from their downtown locations. Customers are randomly assigned to receive their order in either the new biodegradable cup or the standard cup. Neither the customers nor the baristas recording satisfaction scores know which cup type each customer receives. After finishing their drink, each customer rates their overall experience on a 1–10 scale.

(a) (6 marks) Identify the following elements of this experiment:

- Subjects
- Factor(s)
- Treatments
- Response variable

(b) (1 mark) What type of blinding is used in this experiment?

(c) (1 mark) This experimental design is best described as:

A. A matched pairs design

B. A randomized block design

C. A completely randomized design

D. An observational study

(d) (2 marks) Suppose the coffee chain instead tested the biodegradable cup only at their King Street location and the standard cup only at their Richmond Street location, then compared the average satisfaction scores between the two locations. Explain why this modified design is problematic.

Part 2 – Python

Olympic Figure Skating: CLT Simulation [12 marks]

Q6. In this problem, you will use simulation to demonstrate the Central Limit Theorem.

At the 2026 Milano Cortina Winter Olympics, suppose a figure skating judge awards technical element scores on a simplified scale from 1 to 6 (whole numbers only).[1] Assume each score value is equally likely, giving a uniform distribution with $\mu = 3.5$ and $\sigma \approx 1.71$.

(a) (3 marks) Write a function `simulate_sample_means(n, num_samples)` that:

- Simulates one judge scoring n performances (each scored 1–6) and calculates the mean score
- Repeats this process `num_samples` times
- Returns an array of sample means

Test your function with $n = 5$ and `num_samples`$= 10$, and print the resulting sample means.

(b) (6 marks) Use your function to generate 1000 sample means for each of the following sample sizes: $n = 1$, $n = 5$, $n = 30$, and $n = 100$. Plot a histogram of the sample means for each sample size. Use appropriate titles indicating the sample size.

(c) (3 marks) Based on your histograms from part (b), describe how the distribution of sample means changes as the sample size increases. Comment on both the shape and spread of the distributions.

[1] Scoring simplified for this exercise. The actual ISU judging system uses a different scale.

Student Wellbeing Survey [10 marks]

The file `student_survey.csv` contains data from a survey of 500 university students, including their hours of sleep per night (`sleep_hours`), study hours per week (`study_hours`), and stress level on a 1–10 scale (`stress_level`).

For this problem, assume the population standard deviation of sleep hours is known to be $\sigma = 1.25$ hours.

Q7. (a) (2 marks) Load the dataset and calculate the sample mean and sample standard deviation of the `sleep_hours` variable. Print both values.

(b) (3 marks) Using the known population standard deviation $\sigma = 1.25$, compute and print a 95% confidence interval for the true mean sleep hours. Display the critical value z^*, margin of error, and the interval bounds.

(c) (5 marks) To illustrate the meaning of "95% confidence," perform the following simulation. Treat the sample mean from part (a) as the true population mean μ, with $\sigma = 1.25$. Draw 100 random samples of size $n = 30$ from a normal distribution with these parameters. For each sample, construct a 95% confidence interval and count how many intervals contain μ. Print the result.