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.

Solution: The z-score is:

$$z = \frac{9.0 - 7.5}{1.2} = 1.25$$

From the standard normal table, $P(Z < 1.25) = 0.8944$, so:

$$P(X > 9.0) = 1 - 0.8944 = \boxed{0.1056}$$

So about 10.6% of non-athletes sleep more than 9 hours.

(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?

Solution: We want x where $P(X < x) = 0.20$ in the athletes' distribution ($\mu = 6.5$, $\sigma = 1.0$).

The z-score for the 20th percentile is $z = -0.84$ (from the table). Converting back:

$$x = \mu + z\sigma = 6.5 + (-0.84)(1.0) = \boxed{5.66 \text{ hours}}$$

(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.)

Solution: We need the middle 90% of the athletes' distribution ($\mu = 6.5$, $\sigma = 1.0$). The cutoffs for the outer 10% are $z = \pm 1.645$.

On the athletes' scale:

$$\text{Lower bound} = 6.5 - 1.645(1.0) = 4.855$$
$$\text{Upper bound} = 6.5 + 1.645(1.0) = 8.145$$

So athletes sleeping between $\boxed{4.855}$ and $\boxed{8.145}$ hours 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.

Solution: The z-scores tell us how far each student is from their group's mean:

$$z_{\text{athlete}} = \frac{8.0 - 6.5}{1.0} = 1.5$$
$$z_{\text{non-athlete}} = \frac{5.6 - 7.5}{1.2} = -1.58$$

André sleeps above his group average; Bob sleeps below his. Bob's sleep is more unusual since $|z| = 1.58 > 1.5$, meaning he is further from his group's mean in standard-deviation units.

(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.

Solution: For athletes:

$$z = \frac{5.5 - 6.5}{1.0} = -1.00$$

$$P(X < 5.5) = P(Z < -1.00) = \boxed{0.1587}$$

For non-athletes:

$$z = \frac{5.5 - 7.5}{1.2} = -1.67$$

$$P(X < 5.5) = P(Z < -1.67) = \boxed{0.0475}$$

(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.

Solution: By Bayes' rule,

$$P(\text{Athlete} \mid \text{Flagged}) = \frac{P(\text{Flagged} \mid \text{Athlete}) \cdot P(\text{Athlete})}{P(\text{Flagged})}$$

We need $P(\text{Flagged})$ first (law of total probability):

$$\begin{aligned} P(\text{Flagged}) &= P(\text{Flagged} \mid \text{Athlete}) \cdot P(\text{Athlete}) + P(\text{Flagged} \mid \text{Non-athlete}) \cdot P(\text{Non-athlete}) \\ &= (0.1587)(0.30) + (0.0475)(0.70) \\ &= 0.04761 + 0.03325 \\ &= 0.08086 \end{aligned}$$

So:

$$P(\text{Athlete} \mid \text{Flagged}) = \frac{0.04761}{0.08086} = \boxed{0.589}$$

Note that approximately 59% of flagged students are athletes, despite athletes being only 30% of the student body.

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.

Solution: Since $n = 36 \geq 30$, the CLT tells us $\bar{X}$ is approximately normal with:

- Mean: $\mu_{\bar{X}} = \mu = 15$ hours
- Standard error: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3.8}{\sqrt{36}} = \frac{3.8}{6} = 0.633$ hours
- Shape: approximately normal

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

Solution: The z-score for $\bar{x} = 16$:

$$z = \frac{\bar{x} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{16 - 15}{0.633} = 1.58$$

From the table, $P(Z < 1.58) \approx 0.9429$, so:

$$P(\bar{X} > 16) = 1 - 0.9429 = \boxed{0.0571}$$

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

Solution: With $n = 144$:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3.8}{\sqrt{144}} = \frac{3.8}{12} = 0.317 \text{ hours}$$

The standard error dropped from 0.633 to 0.317. It halved because the sample size quadrupled ($144/36 = 4$), and $\sqrt{4} = 2$.

(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.

Solution: No. The columnist is confusing the population with the sampling distribution. Surveying more students does not change the population's exercise habits at all: the population distribution stays exactly the same. What the CLT actually says is that the sampling distribution of the sample mean becomes approximately normal as n grows, regardless of the shape of the population.

(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?

Solution: From part (a), Western's standard error is $\sigma_{\bar{X}} = \frac{3.8}{\sqrt{36}} = \frac{3.8}{6} \approx 0.633$. Setting the second university's SE equal to this:

$$\frac{5.7}{\sqrt{n}} = 0.633 \implies \sqrt{n} = \frac{5.7}{0.633} = 9 \implies n = 81$$

They need $\boxed{n = 81}$, more than double Western's 36. More variable populations require larger samples to achieve the same precision. Here the SD is $1.5\times$ larger ($5.7 = 1.5 \times 3.8$), so the required n is $1.5^2 = 2.25$ times larger ($2.25 \times 36 = 81$).

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.

Solution: The 95% CI:

$$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}} = 37.0 \pm 1.96 \cdot \frac{0.6}{\sqrt{36}} = 37.0 \pm 1.96 \cdot 0.1 = 37.0 \pm 0.196$$

$$\boxed{(36.804,\ 37.196)\,^\circ\text{C}}$$

Since $37.3\,^\circ\text{C}$ lies above the upper bound (37.196), the interval does not support the company's claim.

(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?

Solution: We need the margin of error ≤ 0.1:

$$z^* \cdot \frac{\sigma}{\sqrt{n}} \leq 0.1 \implies \sqrt{n} \geq \frac{1.96 \times 0.6}{0.1} = 11.76 \implies n \geq 138.3$$

Rounding up: $\boxed{n = 139}$.

The colleague's suggestion of $n = 72$ is insufficient. Halving the margin of error requires quadrupling the sample size (not doubling), since SE depends on $\sqrt{n}$.

(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?

Solution: A 99% interval is wider because the critical value increases from $z^* = 1.96$ to $z^* = 2.576$: we need a wider net to be more confident.

The new margin of error is $2.576 \times 0.1 = 0.258$, giving roughly $(36.742, 37.258)\,^\circ\text{C}$. Even this wider interval does not reach $37.3\,^\circ\text{C}$, so the conclusion is unchanged: the data do not support 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 $36.804\,^\circ\text{C}$ and $37.196\,^\circ\text{C}$." Is this interpretation correct? Provide the correct interpretation and explain the distinction.

Solution: No. The true mean is a fixed (unknown) value which is either in the interval or is not.

The correct interpretation is: "We are 95% confident that the true mean resting temperature falls between $36.804\,^\circ\text{C}$ and $37.196\,^\circ\text{C}$." What "95% confident" really means is that if we repeated this process many times, about 95% of the resulting intervals would contain the true mean. The 95% refers to how reliable the *procedure* is, not to any one interval.

(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.

Solution: Apply the conversion to each bound:

$$\begin{aligned} \text{Lower:} \quad & \tfrac{9}{5}(36.804) + 32 = 66.247 + 32 = 98.247\,^\circ\text{F} \\ \text{Upper:} \quad & \tfrac{9}{5}(37.196) + 32 = 66.953 + 32 = 98.953\,^\circ\text{F} \end{aligned}$$

$$\boxed{(98.247,\ 98.953)\,^\circ\text{F}}$$

The margin of error transforms as $\frac{9}{5} \times 0.196 = 0.353\,^\circ\text{F}$. The additive constant $(+32)$ shifts the entire interval without changing its width; only the multiplicative factor $(\frac{9}{5})$ scales the margin of error.

Note: the centre of the interval is $\frac{9}{5}(37.0) + 32 = 98.6\,^\circ\text{F}$ — the familiar benchmark for normal body temperature.

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.

Solution: Stratified Random Sample. The organizers split attendees into strata by sport category and then randomly sample within each stratum.

(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.

Solution: Cluster Sample. Cities act as clusters: they randomly chose 8 cities, then surveyed all subscribers in those cities rather than sampling individuals directly.

(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.

Solution: Convenience Sample. They just approached whoever was around, with no randomization. Students in the student centre on a Wednesday afternoon are probably not representative of the whole campus.

(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.

Solution: Simple Random Sample (SRS). Every household has an equal chance of selection, and they drew from the complete list using a random number generator.

(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.

Solution: Systematic Sample. After picking a random start, they selected every 25th patient, a fixed repeating interval rather than a fresh random draw each time.

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

Solution:

- Subjects: The 120 customers
- Factor: Cup type (one factor with two levels)
- Treatments: Biodegradable cup and standard cup (2 treatments)
- Response variable: Customer satisfaction rating (1–10 scale)

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

Solution: Double-blind. Neither the customers nor the baristas recording scores know which cup type was used.

(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

Solution: All 120 customers are randomly assigned to a treatment with no blocking or pairing, so this is a completely randomized design.

(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.

Solution: This confounds cup type with location. If King Street scores higher, is that because of the cup or because that location has friendlier staff, a nicer atmosphere, or different clientele? We cannot tell. Ideally, cup type should be randomly assigned *within* each location (a randomized block design, blocking on location), so that location effects are controlled and cannot be confused with the cup effect.

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.

Solution:

Python Code

```python
import numpy as np
import matplotlib.pyplot as plt

def simulate_sample_means(n, num_samples):
    means = []
    for i in range(num_samples):
        sample = np.random.randint(1, 7, size=n)
        means.append(sample.mean())
    return np.array(means)

np.random.seed(123)
test_means = simulate_sample_means(n=5, num_samples=10)
print("Sample means:", test_means)
```

Output

```
Sample means:  [3.8 3.  1.4 3.6 3.  4.  3.4 4.  3.  3.4]
```

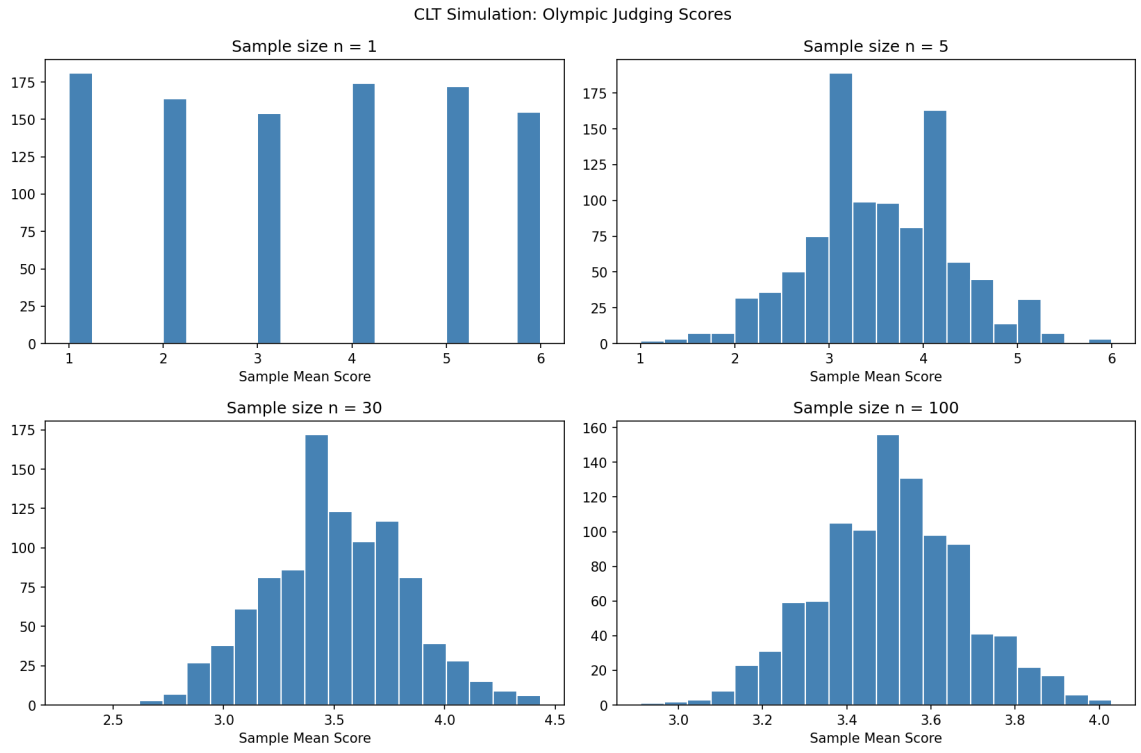
(b) (6 marks) Use your function to generate 1000 sample means for each of the following sample sizes: $n = 1$, $n = 5$, $n = 30$, and $n = 100$. Plot a histogram of the sample means for each sample size. Use appropriate titles indicating the sample size.

Solution:

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

Python Code

```
np.random.seed(42)

means_n1   = simulate_sample_means(n=1,   num_samples=1000)
means_n5   = simulate_sample_means(n=5,   num_samples=1000)
means_n30  = simulate_sample_means(n=30,  num_samples=1000)
means_n100 = simulate_sample_means(n=100, num_samples=1000)

fig, axes = plt.subplots(2, 2, figsize=(12, 8))

axes[0, 0].hist(means_n1,   bins=20, color='steelblue', edgecolor='white')
axes[0, 1].hist(means_n5,   bins=20, color='steelblue', edgecolor='white')
axes[1, 0].hist(means_n30,  bins=20, color='steelblue', edgecolor='white')
axes[1, 1].hist(means_n100, bins=20, color='steelblue', edgecolor='white')

axes[0, 0].set_title('Sample size n = 1')
axes[0, 1].set_title('Sample size n = 5')
axes[1, 0].set_title('Sample size n = 30')
axes[1, 1].set_title('Sample size n = 100')

axes[0, 0].set_xlabel('Sample Mean Score')
axes[0, 1].set_xlabel('Sample Mean Score')
axes[1, 0].set_xlabel('Sample Mean Score')
axes[1, 1].set_xlabel('Sample Mean Score')

plt.suptitle('CLT Simulation: Olympic Judging Scores')
plt.tight_layout()
plt.savefig('figs/clt_simulation.png', dpi=150, bbox_inches='tight')
plt.close()
```



(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.

Solution: Two things happen as n grows:

- Shape: At $n = 1$ the histogram is flat (it is the population). By $n = 30$ it already looks bell-shaped, and at $n = 100$ it is clearly normal.
- Spread: The histograms get narrower; sample means cluster more tightly around 3.5 as the standard error shrinks.

This is the CLT in action: even though the underlying scores are uniform (not normal at all), the distribution of the *sample mean* becomes normal as n increases.

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.

Solution:

Python Code

```python
import pandas as pd
from scipy import stats

df = pd.read_csv("data/student_survey.csv")

mean_sleep = df['sleep_hours'].mean()
std_sleep = df['sleep_hours'].std()

print(f"Sample Mean: {mean_sleep:.2f} hours")
print(f"Sample Std Dev: {std_sleep:.2f} hours")
```

Output

```
Sample Mean:  6.79 hours
Sample Std Dev:  0.78 hours
```

(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.

Solution:

Python Code

```python
sigma = 1.25
n = len(df)
z_star = stats.norm.ppf(0.975)    # critical value for 95% CI
margin_of_error = z_star * sigma / np.sqrt(n)

lower = mean_sleep - margin_of_error
upper = mean_sleep + margin_of_error

print(f"z* = {z_star:.3f}")
print(f"Margin of error: {margin_of_error:.3f} hours")
print(f"95\\% CI: ({lower:.3f}, {upper:.3f})")
```

Output

```
z* = 1.960
Margin of error:  0.110 hours
95% CI: (6.681, 6.900)
```

(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.

Solution:

Python Code

```
np.random.seed(42)
mu = mean_sleep
z_star = stats.norm.ppf(0.975)
count = 0

for _ in range(100):
    sample = np.random.normal(mu, 1.25, 30)
    sample_mean = sample.mean()
    margin = z_star * 1.25 / np.sqrt(30)
    lower = sample_mean - margin
    upper = sample_mean + margin
    if lower <= mu <= upper:
        count += 1

print(f"Number of intervals containing the true mean: {count} out of 100")
```

Output

```
Number of intervals containing the true mean:  97 out of 100
```

Typically, about 95 out of 100 intervals will contain the true mean, reflecting the interpretation of a 95% confidence level.