Assignment Code: DA-AG-006

Statistics Advanced - 1| **Assignment**

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

Total Marks: 200

Question 1: What is a random variable in probability theory?

Answer:

A random variable is a way of turning the outcomes of a random experiment into numbers. It is really just a function or a map that takes each outcome from a sample space and assigns it a number. We normally use capital letters like X to represent a random variable.

A Random Experiment: Rolling of Dice

Let us take a basic random experiment: rolling two dice. The outcomes of this experiment include all possible pairs of numbers that can show up on the dice. So, the sample space includes combinations like (1, 1), (2, 3), and (6, 5). These pairs represent every possible result from rolling the two dice.

Now, consider which we only care about the sum of the two numbers on the dice. Here a random variable will be useful. Instead of caring about individual dice rolls, we define a random variable X such that, that takes each pair of dice rolls and returns their sum. For example −

- **X(1, 1) = 2**
- **X(2, 3) = 5**
- **X(6, 5) = 11**

In this way, we can summarize all the possible outcomes of the

dice roll with the sums. It extracts the specific piece of information we care about from the experiment.

Formal defination

A random variable is a measurable function from a sample space Ω to the real number line R. It implies the following points −

- **Sample Space − This is the set of all possible outcomes of a random experiment. For the dice example, it is the set of all pairs of dice rolls, like (1, 1), (2, 3), …,(6, 6).**
- **Real Numbers − This is just the set of numbers we use every day. They could be positive, negative, fractions, decimals, and so on. In most cases, our random variables will map outcomes to real numbers, like the sum of the dice.**
- **Measurable Function − This means that the random variable behaves nicely with the probabilities we assign to the outcomes. If we think of the sample space as all the possible outcomes, then the random variable maps those outcomes into a smaller number (like the sum in our dice example).**

Question 2: What are the types of random variables?

Answer:

Types of Random Variables

There are two main types of random variables: the discrete and continuous. Depending on what type of experiment we are conducting, the random variable can either take on a set of distinct values (discrete) or any value within a range (continuous).

Discrete Random Variables

A discrete random variable is one that can take on only specific values. Like the integer or whole numbers. For example, in our dice rolling example, the random variable can only take on the values 2 through 12. These are the possible sums of the two dice.

Discrete random variables are used in situations where we can

count the outcomes. The number of heads in a series of coin flips or the number of customers arriving at a store in an hour.

- **Takes fixed, countable values**
- **Usually whole numbers**
- **It can count the possible outcomes**

Examples:

- **Number of students in a class (30, 31, 32…)**
- **Number of goals in a match (0, 1, 2…)**
- **Number of heads in 3 coin tosses (0, 1, 2, 3)**

These are described using a Probability Mass Function (PMF).

Continuous Random Variables

A continuous random variable can take on any value within a certain range. For example, if we are measuring the time it takes for a car to complete a race. It could take any value like 1.23 seconds, 3.5 seconds, 4.98 seconds, and so on. Continuous random variables are used in situations where the outcome can vary smoothly over a range of values.

Takes any value within a range
Can include decimals
It cannot count all possible values

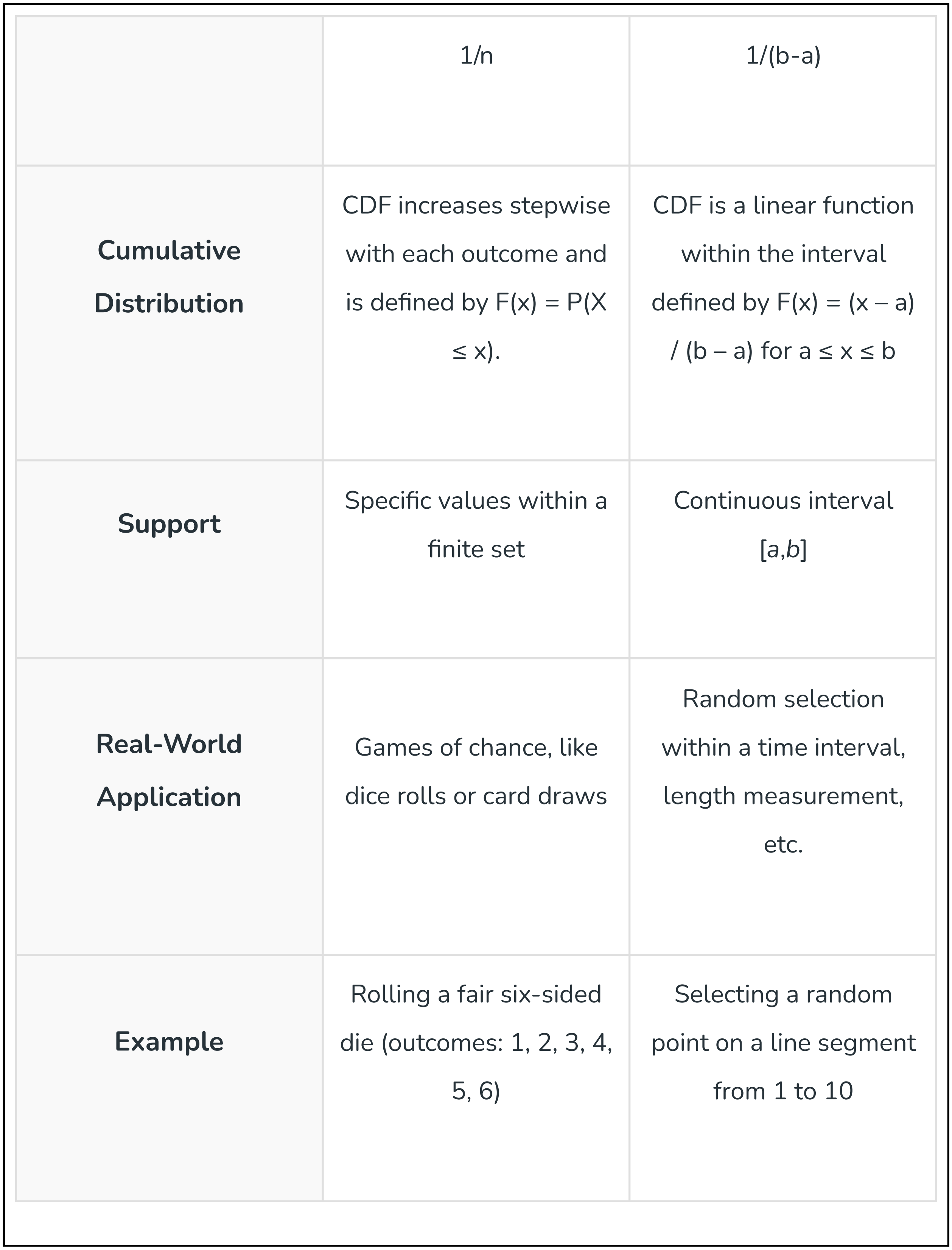
Examples:

- **Height of a student (160.2 cm, 160.25 cm, etc.)**
- **Weight of a bag**
- **Time taken to complete a test**

These are described using a Probability Density Function (PDF).

Question 3: Explain the difference between discrete and continuous distributions. **Answer:**

Basis	Discrete Uniform Distribution	Continuous Distribution
Nature of Outcomes	Finite and countable set of outcomes	Infinite and uncountable range of outcomes
Probability Function	Probability Mass Function (PMF): P(X=x)= 1/n	Probability Density Function (PDF): f(x) = 1/(b-a)
Range of Values	Specific discrete values *x*1,*x*2,...,*xn*	Continuous range of values between *a* and *b*
Probability Calculation	Equal probability for each outcome: P(X=x)=	Equal density across the interval: f(*x*)=



	$1/n$	$1/(b-a)$
Cumulative Distribution	CDF increases stepwise with each outcome and is defined by $F(x) = P(X \le x)$.	CDF is a linear function within the interval defined by $F(x) = (x – a) / (b – a)$ for $a \le x \le b$
Support	Specific values within a finite set	Continuous interval $[a,b]$
Real-World Application	Games of chance, like dice rolls or card draws	Random selection within a time interval, length measurement, etc.
Example	Rolling a fair six-sided die (outcomes: 1, 2, 3, 4, 5, 6)	Selecting a random point on a line segment from 1 to 10

Question 4: What is a binomial distribution, and how is it used in probability? **Answer:**

A binomial distribution is a probability model used when you're counting how many times something happens in a fixed number of tries, where each try has only two possible outcomes.

For example if a, questions like:

> **"If I repeat the same experiment several times, what's the probability I get exactly *k* successes?"**

Conditions for a binomial distribution

A situation follows a binomial distribution if all of these are true:

1. **There is a fixed number of trials (called *n*).**
2. **Each trial has only two outcomes (often called *success* and *failure*).**
3. **The probability of success (*p*) is the same for every trial.**
4. **The trials are independent (one doesn't affect another).**

Common examples

- Flipping a coin 10 times and counting how many heads we get
- Taking a multiple-choice test and counting correct answers
- Counting how many free throws a basketball player makes out of 20 attempts

How it's used in probability

The binomial distribution lets us to calculate the probability of getting **exactly k successes** using this formula:

P(*X*=*k*)=(*nk*)*pk*(1−*p*)*n*−*k*P(X=k)=(kn)pk(1−p)n−k

Where:

- *n*n = number of trials
- *k*k = number of successes
- *p*p = probability of success on one trial
- (*nk*)(kn) = number of ways to choose *k*k successes from *n*n trials

Simple example

Suppose we flip a fair coin 5 times:

- *n*=5n=5
- *p*=0.5p=0.5

we can use the binomial distribution to find:

- The probability of getting **exactly 3 heads**
- The probability of getting **at least 1 head**
- The probability of getting **no heads**

it’s important because:

The binomial distribution is widely used in:

- Statistics and data analysis
- Quality control (defective vs. non-defective items)
- Biology and medicine (success/failure of treatments)
- Exams, surveys, and experiments

Question 5: What is the standard normal distribution, and why is it important? **Answer:**

The standard normal distribution is a special case of the normal distribution that is used as a reference model in probability and statistics.

The standard normal distribution:

- **Is bell-shaped and symmetric**
- **Has a mean of 0**
- **Has a standard deviation of 1**

It is usually denoted as $Z \sim N(0, 1)$.

Any normal distribution can be converted into the standard normal distribution by standardizing the values. This is done using a z-score:

$$Z=x-\mu /\sigma$$

Where:

- **x = original value**
- **μ = mean**
- **σ= standard deviation**

A z-score tells how many standard deviations a value is above or below the mean.

It’s important because:

The standard normal distribution is important because:

1. It simplifies probability calculations

- **Instead of having separate tables for every normal distribution, we use one standard normal table.**
- **Once data is standardized, probabilities can be looked up easily.**

2. It helps compare different data sets

- **Z-scores let you compare values from different distributions (e.g., test scores from different exams).**
- **Example: A score of +2 is high regardless of the original scale.**

3. It’s widely used in statistics

- **Hypothesis testing**
- **Confidence intervals**
- **Statistical modeling**
- **Quality control**

Simple example

If test scores are normally distributed with:

- **Mean = 70**
- **Standard deviation = 10**

A score of 85 has:

$z=85-70/10$
$=1.5z$
This means the score is 1.5 standard deviations above the mean, and you can use the standard normal distribution to find how unusual that score is.

Question 6: What is the Central Limit Theorem (CLT), and why is it critical in statistics? **Answer:**

The **Central Limit Theorem (CLT)** is one of the most important ideas in statistics because it explains *why* the normal distribution appears so often in real life.

The Central Limit Theorem states that:

> **If you take sufficiently large random samples from any population with a finite mean and variance, the distribution of the sample mean will be approximately normal, regardless of the population's original distribution.**

In simple terms

- Even if the original data is **skewed**, **uniform**, or **non-normal**,
- The **averages** of large samples tend to form a **normal (bell-shaped) distribution**.

They are some Key conditions

For the CLT to apply:

1. Samples must be **random and independent**.
2. The population must have a **finite mean and variance**.
3. The sample size must be **large enough** (often $n \geq 30$n≥30 is used as a rule of thumb).

Important results of the CLT

If:

- Population mean = μ
- Population standard deviation = σ
- Sample size = n

Then:

- Mean of the sampling distribution = μ
- Standard deviation (standard error) = σ/nσ

Why the CLT is critical in statistics

1. It allows normal-based methods to be used

Because sample means are approximately normal:

- We can use **z-scores**
- Build **confidence intervals**
- Perform **hypothesis tests**

Even when the population itself is not normal.

---

2. It explains why the normal distribution is so common

Many real-world measurements are averages (test scores, survey results, measurements), and averages tend to be normally distributed due to the CLT.

---

3. It makes statistical inference possible

The CLT is the foundation of:

- Estimating population parameters
- Comparing groups
- Making predictions from sample data

Without it, many standard statistical techniques would not work.

Simple example

Suppose a population is strongly skewed (like income).

- Individual incomes are not normally distributed.
- But the **average income** from many random samples will form a normal distribution as the sample size increases.

Question 7: What is the significance of confidence intervals in statistical

analysis? **Answer:**

Confidence intervals (CIs) are important in statistical analysis because they quantify uncertainty around an estimate and help us to understand how reliable that estimate is.

1. They show the range of plausible values

A confidence interval gives a range that is likely to contain the true population parameter (such as a mean or proportion).

- **Example: A 95% confidence interval of [48, 52] suggests the true value is reasonably close to that range.**

This is more informative than a single point estimate like "the mean is 50."

2. They express uncertainty from sampling

Because data usually come from a sample (not the whole population), there is natural variability. Confidence intervals reflect how much estimates might vary if you took different samples.

- **Narrow interval → more precise estimate**
- **Wide interval → more uncertainty or less data**

3. They help assess statistical significance

Confidence intervals can be used instead of hypothesis tests:

- **If a CI for a mean difference does not include 0, the result is statistically significant.**
- **If a CI for a ratio (like a risk ratio) does not include 1, the effect is significant.**

This connects estimation and hypothesis testing in a clear way.

---

4. They indicate practical importance

CIs help judge whether an effect is not just statistically significant but meaningful in real-world terms.

- **A tiny effect with a very narrow CI may be statistically significant but practically unimportant.**
- **A large effect with a wide CI suggests potential importance but uncertainty.**

---

5. They encourage better interpretation than p-values alone

Unlike p-values, confidence intervals:

- **Show effect size**
- **Show direction of effect**
- **Show precision**

This makes conclusions more transparent and informative.

In short

Confidence intervals are significant because they:

- **Provide a range of likely values for a population parameter**
- **Reflect uncertainty due to sampling**
- **Help assess significance and practical relevance**
- **Lead to clearer, more responsible statistical conclusion.**

Question 8: What is the concept of expected value in a probability distribution? **Answer:**

The expected value (also called the mean or expectation) of a probability distribution is a way to describe the long-run average outcome of a random process.

Core idea

If we repeat a random experiment many times, the expected value is the value which we would expect to get on average.

It is not necessarily a value that actually occurs, but a weighted average of all possible outcomes, where each outcome is weighted by its probability.

Mathematical definition

For a discrete random variable
$E(X)=\sum x\cdot P(X=x)$**E(X)=∑x⋅P(X=x)**
multiply each possible value by its probability and add them all together.

Example:
If we roll a fair six-sided die:

$E(X)=1\cdot 16+2\cdot 16+\cdots+6\cdot 16$
=3.5E(X)
=1 ⋅ 61+2 ⋅ 61+⋯+6 ⋅ 61
=3.5

For a continuous random variable
$E(X)=\int -\infty\infty x\,f(x)\,dx$**E(X)=∫−∞∞xf(x)dx**

where $f(x)$**f(x) is the probability density function.**

Why expected value matters

1. **Summarizes the distribution**
   It gives a single number representing the "center" of the distribution.
2. **Foundation for decision-making**
   In economics, games, and risk analysis, expected value helps compare choices based on average outcomes.
3. **Key to other statistical concepts**
   Variance, standard deviation, and many statistical models are built using expected values.
4. **Predicts long-term behavior**

Over many trials, observed averages tend to move closer to the expected value (Law of Large Numbers).

Important clarifications

- **Expected value is not a guarantee of any single outcome.**
- **It can be outside the range of likely outcomes in some distributions.**
- **It reflects probability-weighted outcomes, not fairness or certainty.**

In simple terms

The expected value tells us what we should expect on average if we repeat an experiment many times, making it a central concept in probability and statistics.

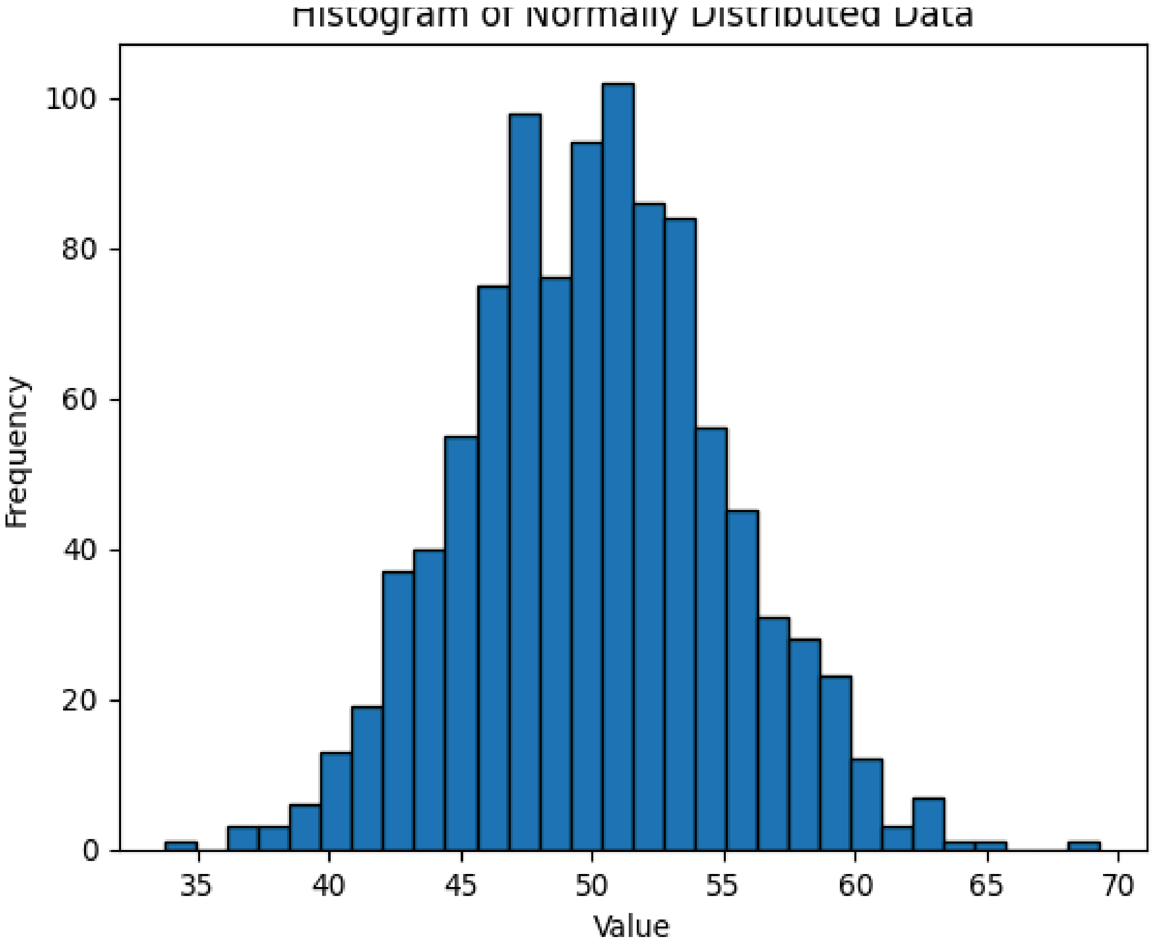
Question 9: Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution. (*Include your Python code and output in the code box below.*)

Answer:

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

# Set random seed for reproducibility (optional)
np.random.seed(42)

# Generate 1000 random numbers from a normal distribution
data = np.random.normal(loc=50, scale=5, size=1000)

# Compute mean and standard deviation
```

```
mean = np.mean(data)
std_dev = np.std(data)

# Print results
print("Computed Mean:", mean)
print("Computed Standard Deviation:", std_dev)

# Plot histogram
plt.hist(data, bins=30, edgecolor='black')
plt.title("Histogram of Normally Distributed Data")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.show()
```

output:



```
Computed Mean: 50.096660279111624
Computed Standard Deviation: 4.893631038736771
```

[redacted]

Question 10: You are working as a data analyst for a retail company. The company has collected daily sales data for 2 years and wants you to identify the overall sales trend.

daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,

235, 260, 245, 250, 225, 270, 265, 255, 250, 260]

- Explain how you would apply the Central Limit Theorem to estimate the average sales with a 95% confidence interval.
- Write the Python code to compute the mean sales and its confidence interval.

(*Include your Python code and output in the code box below.*)

Answer:

As the data analyst for a retail company.

1. Applying the Central Limit Theorem (CLT)

The Central Limit Theorem states that when you take a sufficiently large sample (typically n≥30n \ge 30n≥30, but often acceptable for smaller samples in practice), the sampling distribution of the sample mean will be approximately normally distributed, regardless of the shape of the original data.

In this case:

- **We have daily sales observations (n=20n = 20n=20)**
- **We want to estimate the true average daily sales**
- **Using CLT, we can approximate the distribution of the sample mean as normal**
- **This allows us to construct a 95% confidence interval (CI) for the mean**

Steps:

1. **Compute the sample mean (x¯\bar{x}x¯)**
2. **Compute the sample standard deviation (sss)**
3. **Compute the standard error:**
   SE=snSE = \frac{s}{\sqrt{n}}SE=ns
4. **For a 95% confidence level, use the z-score:**
   z=1.96z = 1.96z=1.96
5. **Construct the confidence interval:**

x¯±z×SE\bar{x} \pm z \times SEx¯±z×SE

---

2. Python Code to Compute Mean and 95% Confidence Interval

```
import numpy as np
from math import sqrt


# Daily sales data
daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,
               235, 260, 245, 250, 225, 270, 265, 255, 250, 260]


# Sample size
n = len(daily_sales)


# Mean
mean_sales = np.mean(daily_sales)


# Sample standard deviation
std_sales = np.std(daily_sales, ddof=1)


# Standard error
standard_error = std_sales / sqrt(n)


# Z-score for 95% confidence
z = 1.96
```

```
# Confidence interval

lower_bound = mean_sales - z * standard_error

upper_bound = mean_sales + z * standard_error


mean_sales, lower_bound, upper_bound
```

Output:

```
(np.float64(248.25),

 np.float64(240.68312934041109),

np.float64(255.81687065958891))
```

3. Interpretation

- **Average daily sales: 248.25 units**
- **95% Confidence Interval: (240.69, 255.81)**

This means we are 95% confident that the true average daily sales over the two-year period lie between approximately 241 and 256 units, indicating a fairly stable overall sales level.