Statistics Advanced - 1| Assignment

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

Answer: Random Variable in Probability Theory

1. Definition:
A random variable is a function that assigns a numerical value to each outcome of a random experiment.

- It is called random because the exact outcome cannot be predicted in advance.
- It is called a variable because it can take different possible values depending on the outcome of the experiment.

Mathematically:
If S is the sample space of a random experiment, then a random variable is a function

$$X:S\rightarrow R$$

that maps each outcome in S to a real number.

2. Types of Random Variables

1. Discrete Random Variable
   - Takes countable values (finite or infinite).
   - Examples:
     - Number of heads in 3 coin tosses (0,1,2,3).
     - Outcome of rolling a die (1,2,3,4,5,6).

2.Continuous Random Variable

- Takes uncountably infinite values within a range.
- Examples:
  - Height of students in a class (e.g., 150.5 cm, 172.3 cm).
  - Time taken to complete a task.

3. Examples

- Coin Toss:
  Experiment: Toss a coin once.
  Outcomes: {Head, Tail}.
  Define X:
  - ♦ X=1 if Head occurs.
  - ♦ X=0 if Tail occurs.

Here, X is a discrete random variable.

- Dice Roll:
  Experiment: Roll a die.
  Outcomes: {1, 2, 3, 4, 5, 6}.
  Define Y as the number shown.
  Y can take values {1,2,3,4,5,6}.

- Measuring Weight:
  Experiment: Select a person at random and measure their weight.
  Possible values: Any real number within a range (e.g., 40 kg – 120 kg).
  This is a continuous random variable.

4. Importance of Random Variables

- They simplify probability problems by converting outcomes into numerical form.

- Used to define probability distributions (e.g., Binomial, Normal distribution).

- Help in statistical analysis, prediction, and decision-making.

- Provide the foundation for expectation, variance, and hypothesis testing.

5. Conclusion

A random variable is a core concept in probability and statistics. It bridges the gap between random experiments and mathematical analysis by representing uncertain outcomes as numerical values. Both discrete and continuous random variables are essential for real-world applications such as data analysis, risk modeling, and scientific research.

Question 2: What are the types of random variables?

Answer: Types of Random Variables

A random variable is a function that assigns numerical values to the outcomes of a random experiment.
Random variables are broadly classified into the following types:

1. Discrete Random Variable

- Definition: A random variable that can take only a finite or countably infinite set of values.
- Values are usually integers.
- Examples:
  - Number of heads in 3 coin tosses (0,1,2,3)
  - Outcome of rolling a die (1,2,3,4,5,6)
  - Number of students present in a class

2. Continuous Random Variable

- Definition: A random variable that can take any value within a given range or interval.
- Values are uncountably infinite and usually real numbers.
- Examples:
    - Height of a student (e.g., 165.2 cm, 170.8 cm)
    - Time taken to run 100 meters
    - Weight of a person

3. Mixed Random Variable

- Some random variables can have both discrete and continuous components.
- Example:
    - In insurance, the claim amount may be **0** (discrete part, no claim) or a continuous positive amount (continuous part, actual claim).

Conclusion

Random variables are mainly of two types: discrete and continuous. Discrete variables deal with countable outcomes, while continuous variables deal with measurements across intervals. In special cases, mixed random variables exist, which combine both.

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

Answer: Difference between Discrete and Continuous Distributions

1. Discrete Probability Distribution

- A probability distribution associated with a discrete random variable.
- The variable can take countable values (finite or infinite).
- Probability of each outcome is defined separately.
- Represented by a Probability Mass Function (PMF).
- Example:
    - Tossing 2 coins → Random variable X=number of heads.
    - Possible values: {0, 1, 2}.
    - Distribution:

$$P(X=0)=0.25,\ P(X=1)=0.5,\ P(X=2)=0.25$$

2. Continuous Probability Distribution

- A probability distribution associated with a continuous random variable.
- The variable can take uncountably infinite values within an interval.
- Probability of any exact value is 0; instead, we calculate probability over an interval.
- Represented by a Probability Density Function (PDF).
- Example:
    - Height of students in a class.
    - Random variable Y = height.
    - The probability is expressed as:

$$P(a \leq Y \leq b)=\int abf(y)dy$$

- where $f(y)$ is the probability density function.

3. Key Differences

Feature	Discrete Distribution	Continuous Distribution
Random Variable	Takes countable values	Takes uncountably infinite values
Function Used	Probability Mass Function (PMF)	Probability Density Function (PDF)
Probability of exact value	P(X=x)>0	P(X=x)=0
Representation	Bar graph (histogram-like)	Smooth curve
Examples	Binomial, Poisson, Geometric	Normal, Exponential, Uniform

4. Conclusion

- Discrete distributions deal with countable outcomes like coin tosses or dice rolls.
- Continuous distributions deal with measurements across a range, such as height, weight, or time.
- Together, they form the foundation of probability theory and are widely used in statistics, data analysis, and real-world modeling.

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

Answer: Binomial Distribution in Probability

1. Definition

A Binomial Distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent trials of a Bernoulli experiment (an experiment with only two possible outcomes: success or failure).

2. Conditions for Binomial Distribution

For a random variable X to follow a binomial distribution, the following must hold:

1. The experiment consists of n independent trials.
2. Each trial has only two possible outcomes – Success (with probability p) or Failure (with probability q=1−p).
3. The probability of success p remains constant for all trials.
4. The random variable X represents the number of successes in n trials.

3. Probability Mass Function (PMF)

The probability of getting exactly k successes in n trials is:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,1,2,\ldots,n$$

where:

- n = number of trials
- k = number of successes
- p = probability of success in a single trial
- q=1−p = probability of failure

- $$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

4. Mean and Variance

- Mean: E[X]=np
- Variance: Var(X)=npq

5. Applications of Binomial Distribution

- Quality control (e.g., defective vs. non-defective items).
- Medical studies (e.g., probability of recovery after treatment).
- Opinion polls (e.g., proportion of people favoring a candidate).
- Reliability testing (e.g., success/failure of a machine).

6. Conclusion

The binomial distribution is one of the most important discrete probability distributions. It models real-world situations where experiments have two possible outcomes, repeated a fixed number of times, with constant probability of success. It helps in prediction, risk analysis, and decision-making in various fields.

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

Answer: Standard Normal Distribution

1. Definition

The standard normal distribution is a special case of the normal distribution where:

- The mean (μ) = 0
- The standard deviation (σ) = 1

It is a continuous probability distribution that is symmetric about the mean.

2. Probability Density Function (PDF)

The probability density function of the standard normal distribution is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

Here, z is called the standard normal variable (z-score).

3. Properties of Standard Normal Distribution

- Symmetrical about the mean (z=0).
- Bell-shaped curve.
- Mean = 0, Median = 0, Mode = 0.
- Variance = 1, Standard deviation = 1.
- Total area under the curve = 1.
- Probabilities are calculated using z-tables.

4. Importance

- Acts as the reference distribution for all normal distributions.

- Any normal random variable $X \sim N(\mu,\sigma2)$ can be standardized into a standard normal variable using:
  $Z=X-\mu/ \sigma$
- Helps in computing probabilities and percentiles easily.
- Used in statistical inference (e.g., hypothesis testing, confidence intervals).
- Basis for many advanced models (e.g., regression, machine learning).

5. Applications

- Quality control (checking if measurements deviate from standard).
- Education (standardized test scores like SAT, IQ tests).
- Finance (stock returns, risk analysis).
- Research (finding critical regions in hypothesis testing).

6. Conclusion

The standard normal distribution is a fundamental tool in probability and statistics. By converting any normal distribution into this standard form, it simplifies the calculation of probabilities and plays a central role in hypothesis testing, estimation, and real-world decision-making.

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

Answer: Central Limit Theorem (CLT)

1. Definition

The Central Limit Theorem (CLT) states that:

When independent random samples are drawn from any population with a finite mean (μ) and finite variance ($\sigma2$), the sampling distribution of the sample mean approaches a normal distribution as the sample size becomes large, regardless of the population's original distribution.

2. Mathematical Statement

If $X1,X2,...,Xn$ are i.i.d. random variables with mean μ and variance $\sigma2$, then the standardized sample mean is:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

As $n\to\infty$, Z follows a standard normal distribution $N(0,1)$.

3. Key Points

- Works for any population distribution (normal or non-normal).
- The larger the sample size n, the closer the sampling distribution of X^{-} is to normal.
- In practice, $n\geq30$ is usually considered sufficient.

4. Why is CLT Critical?

1. Foundation of inferential statistics – allows us to make conclusions about populations using sample data.
2. Enables hypothesis testing – most statistical tests rely on normality assumptions.
3. Justifies confidence intervals – since X¯ becomes approximately normal, we can calculate probability ranges.
4. Applies widely – works even if the population is skewed or irregular.
5. Simplifies complex problems – many real-world distributions are unknown, but CLT lets us use the normal distribution as an approximation.

5. Applications

- Opinion polls: Estimating population preferences from small samples.
- Quality control: Checking if average product weight/length meets standards.
- Finance: Modeling average returns of investments.
- Medical research: Analyzing average effects of treatments.

6. Conclusion

The Central Limit Theorem is critical because it explains why the normal distribution is so widely used in statistics. It allows statisticians to apply normal probability models to sample means, making it possible to perform reliable inference, testing, and prediction in almost every field.

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

Answer: Significance of Confidence Intervals in Statistical Analysis

1. Definition

A confidence interval (CI) is a range of values, derived from sample data, that is likely to contain the true population parameter (such as mean or proportion) with a certain level of confidence.
For example: A 95% confidence interval for the population mean is the range of values that has a 95% probability of containing the true mean.

2. Formula (for population mean when σ is known)

$$CI = \bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Where:

- X¯ = sample mean
- σ = population standard deviation
- n = sample size
- Zα/2 = critical value from the standard normal table

If σ is unknown, we use the t-distribution instead of Z.

3. Interpretation

- A 95% confidence interval does not mean there is a 95% chance the true parameter lies within the interval.
- It means: If we take many random samples and construct confidence intervals for each, then 95% of those intervals would contain the true population parameter.

4. Significance / Importance

1. Estimation Accuracy – Provides a range, not just a single point estimate, making results more reliable.
2. Accounts for Uncertainty – Reflects the variability due to sampling.
3. Helps Decision-Making – Narrow intervals mean precise estimates; wide intervals indicate more uncertainty.
4. Used in Hypothesis Testing – If a hypothesized value (e.g., μ=0) does not lie within the CI, we reject the null hypothesis.
5. Communicates Reliability – Confidence intervals give more information than just reporting the mean or proportion.

5. Applications

- Medical research: Estimating the effect of a new drug.
- Elections: Predicting candidate's vote share with margin of error.
- Business & economics: Estimating average income, customer satisfaction, or product demand.
- Quality control: Ensuring product measurements fall within acceptable limits.

6. Conclusion

Confidence intervals are significant in statistical analysis because they provide a range of plausible values for population parameters, while quantifying the uncertainty of sample estimates. They make statistical conclusions more informative, transparent, and reliable than single point estimates.

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

Answer: Expected Value in a Probability Distribution

1. Definition

The expected value (EV) of a random variable is the long-run average value it takes when an experiment is repeated many times.

- It represents the theoretical mean of a probability distribution.
- In simple terms, it is the weighted average of all possible values, where the weights are the probabilities.

2. Formulas

- For a discrete random variable X with outcomes x1,x2,...,xn and probabilities p(xi):

$$E[X] = \sum_{i=1}^{n} x_i \cdot p(x_i)$$

- For a continuous random variable X with probability density function f(x):

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

3. Explanation

- The expected value is not necessarily an outcome that can occur in a single trial, but rather the average outcome if the experiment is repeated many times.
- Example: In rolling a fair die, the expected value is 3.53.53.5, even though you can never roll a 3.5.

4. Examples

(a) Discrete Example – Dice Roll

- Outcomes: {1, 2, 3, 4, 5, 6}
- Probabilities: 1/6 each

$$E[X] = \sum_{i=1}^{6} x_i \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = 3.5$$

(b) Continuous Example – Uniform Distribution (0,1)

$$E[X] = \int_{0}^{1} x \cdot 1\, dx = \frac{1}{2}$$

5. Significance of Expected Value

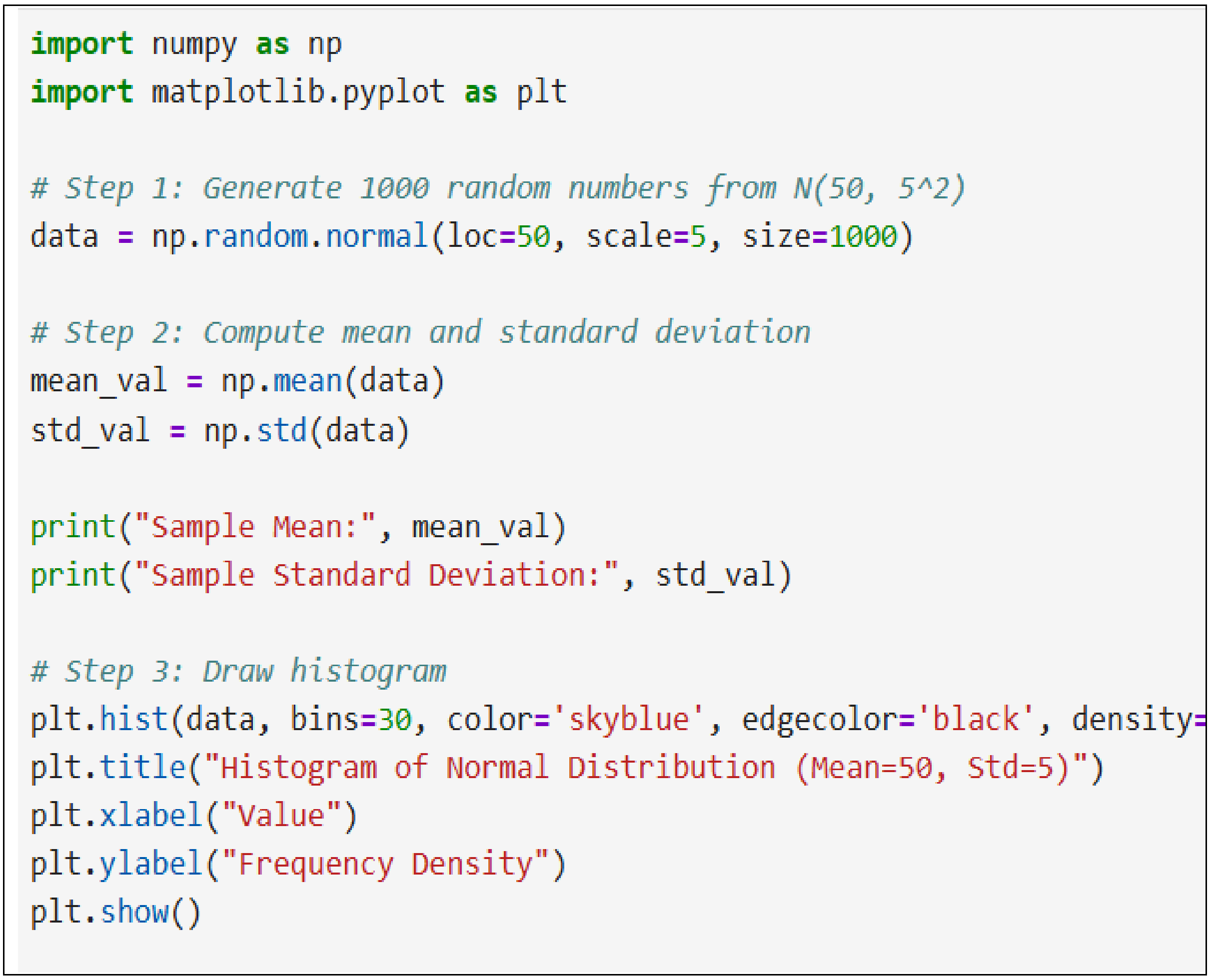
1. Measure of Central Tendency – It is the theoretical mean of the distribution.
2. Decision Making – Used in economics, business, and game theory to evaluate risks.
3. Foundation for Other Measures – Variance, standard deviation, and moment calculations are based on expected value.
4. Real-Life Applications – Insurance premium calculation, stock market risk analysis, gambling outcomes, quality control.

6. Conclusion

The expected value is a fundamental concept in probability that provides a single number summarizing the long-run average of a random variable. It is crucial for prediction, decision-making, and statistical modeling, making it one of the most widely used tools in probability theory.

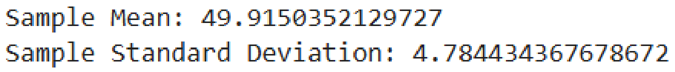
Question 9: Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution. (Include your Python code and output in the code box below.)

Answer:

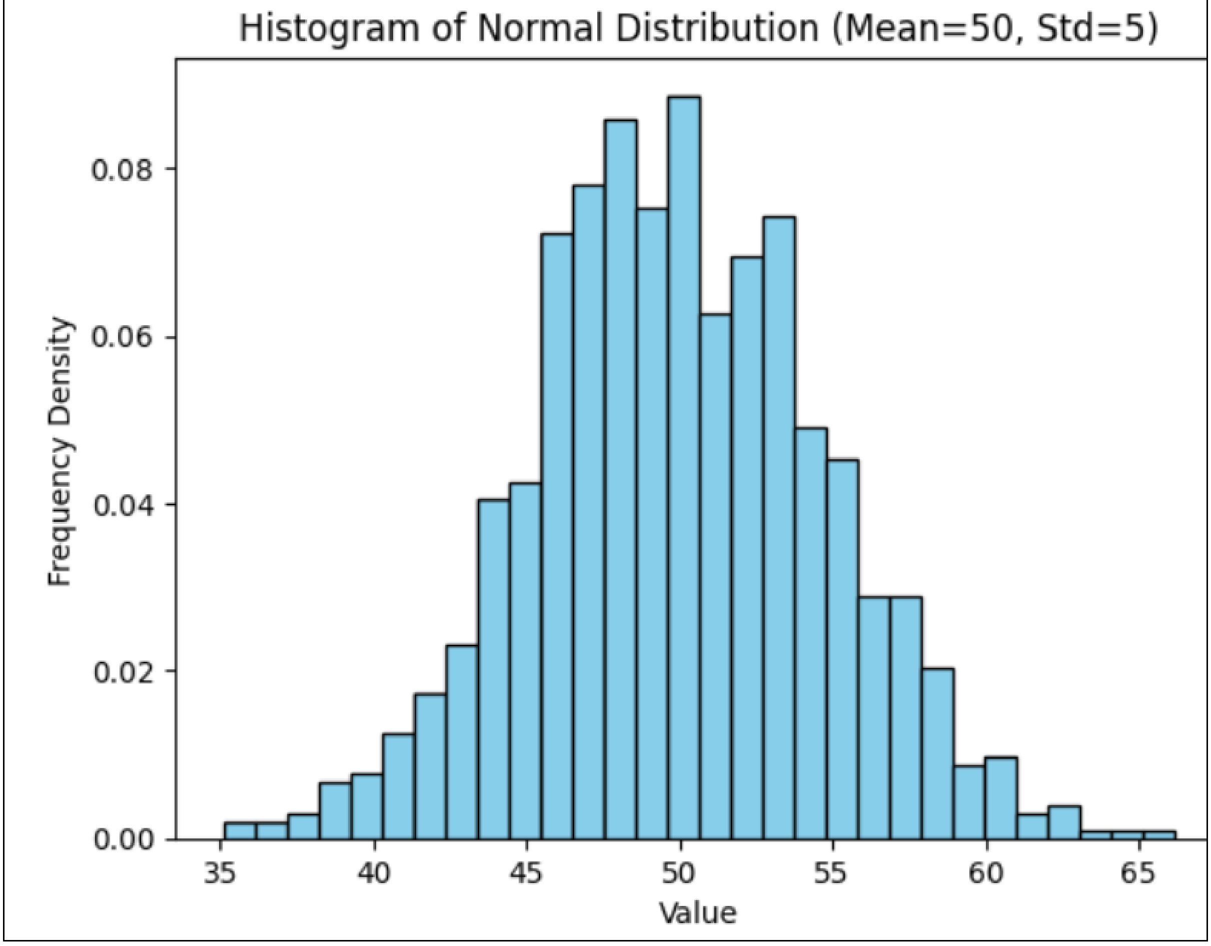


```
import numpy as np
import matplotlib.pyplot as plt

# Step 1: Generate 1000 random numbers from N(50, 5^2)
data = np.random.normal(loc=50, scale=5, size=1000)

# Step 2: Compute mean and standard deviation
mean_val = np.mean(data)
std_val = np.std(data)

print("Sample Mean:", mean_val)
print("Sample Standard Deviation:", std_val)

# Step 3: Draw histogram
plt.hist(data, bins=30, color='skyblue', edgecolor='black', density=
plt.title("Histogram of Normal Distribution (Mean=50, Std=5)")
plt.xlabel("Value")
plt.ylabel("Frequency Density")
plt.show()
```

Output:



```
Sample Mean: 49.9150352129727
Sample Standard Deviation: 4.784434367678672
```



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:

1. Applying the Central Limit Theorem (CLT)

- The company has collected daily sales data. If we take repeated random samples of sales data and compute their means, the sampling distribution of the sample mean will approach a normal distribution (by CLT), regardless of the shape of the original data.

- The sample mean (X¯) is an unbiased estimate of the population mean.
- The 95% confidence interval (CI) is given by:

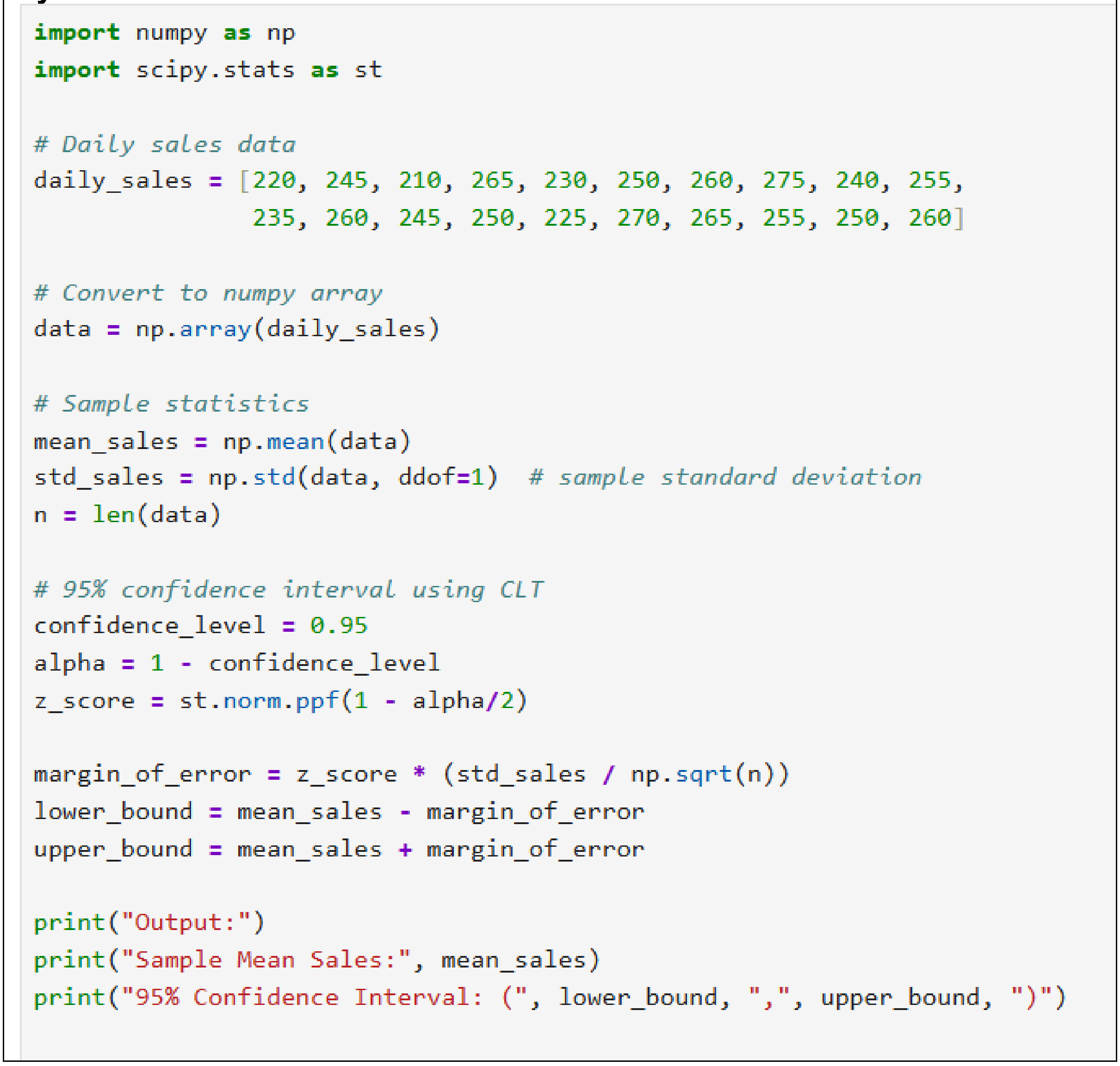
$$CI = \bar{X} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where:

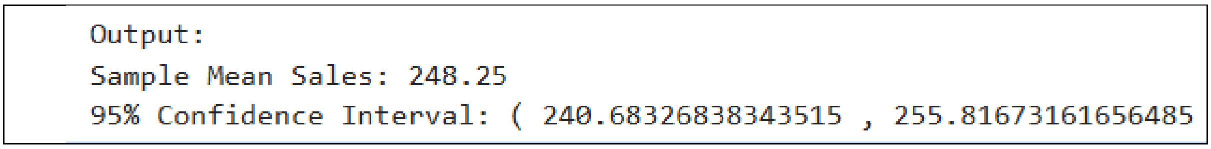
- X¯ = sample mean
- s = sample standard deviation
- n = sample size
- Zα/2=1.96 for 95% CI

This interval gives a range within which the true average daily sales is expected to lie with 95% confidence.

Python Code:



```
import numpy as np
import scipy.stats as st

# 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]

# Convert to numpy array
data = np.array(daily_sales)

# Sample statistics
mean_sales = np.mean(data)
std_sales = np.std(data, ddof=1)  # sample standard deviation
n = len(data)

# 95% confidence interval using CLT
confidence_level = 0.95
alpha = 1 - confidence_level
z_score = st.norm.ppf(1 - alpha/2)

margin_of_error = z_score * (std_sales / np.sqrt(n))
lower_bound = mean_sales - margin_of_error
upper_bound = mean_sales + margin_of_error

print("Output:")
print("Sample Mean Sales:", mean_sales)
print("95% Confidence Interval: (", lower_bound, ",", upper_bound, ")")
```



```
Output:
Sample Mean Sales: 248.25
95% Confidence Interval: ( 240.68326838343515 , 255.81673161656485
```