Statistics Advanced - 1| Assignment

Question 1: What is a random variable in probability theory?
Soln. : A random variable is a mathematical function that assigns a unique numerical value to each possible outcome in the sample space of a random experiment. Instead of looking at qualitative results (like "heads" or "tails"), we use a random variable to map these results to numbers (like 1 and 0).

Types of Random variable:

- *Discrete Random Variables*: These take on a countable number of distinct values. For example, the number of successful free throws in a game can only be whole numbers (0,1,2…….).
- *Continuous Random Variables:* These can take any value within a specific range or interval. For example, the exact time it takes for a chemical reaction to occur could be any value like 12.3321 seconds or 12 seconds.

Question 2: What are the types of random variables?
Soln. : Types of Random variable:

- *Discrete Random Variables*: These take on a countable number of distinct values. For example, the number of successful free throws in a game can only be whole numbers (0,1,2…….).
- *Continuous Random Variables:* These can take any value within a specific range or interval. For example, the exact time it takes for a chemical reaction to occur could be any value like 12.3321 seconds or 12 seconds.

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

- *Discrete Distributions :* A discrete distribution describes variables that can only take on specific, separate values. These are usually "counted" data points.
  Key Logic: You can define the probability of a specific outcome, such as the chance of rolling exactly a 4 on a die (P(X=4)=1/6).
  Examples: Binomial, Poisson, and Bernoulli distributions.

- *Continuous Distributions* : A continuous distribution describes variables that can take any value within a range. These are usually "measured" data points like time, weight, or distance.

Because there are infinite possible values (like 1.7000... vs 1.70001...), the probability of hitting a *single exact point* is mathematically zero. Instead, we calculate the probability that a value falls within a certain interval.
Examples: Normal, Exponential, and Uniform distributions

Question 4: What is a binomial distribution, and how is it used in probability?
Soln. : A binomial distribution is a discrete probability distribution that models the number of "successes" in a fixed number of independent trials. It is one of the most widely used distributions for analyzing binary outcomes, where each trial has only two possibilities: success or failure.

For a situation to follow a binomial distribution, it must meet four specific criteria:

1. Fixed Number of Trials (n): The number of experiments is set in advance (e.g., flipping a coin exactly 10 times).
2. Two Possible Outcomes: Each trial results in either a "success" or a "failure" (e.g., yes/no, pass/fail, defective/non-defective).
3. Independent Trials: The outcome of one trial does not affect any other trial.
4. Constant Probability (p): The probability of success remains the same for every single trial.

Question 5: What is the standard normal distribution, and why is it important?
Soln. : The Standard Normal Distribution is a specific type of continuous probability distribution. It is a special case of the General Normal Distribution where the parameters are fixed:
Mean = 0
Standard Deviation = 1 (and consequently, Variance)
It is represented by the random variable. Geometrically, it is a perfectly symmetrical, bell-shaped curve centered at the origin on the horizontal axis.

- The distribution is perfectly symmetrical around Z=0. This implies that the Mean, Median, and Mode are all equal to 0.
- The "tails" of the curve approach the horizontal axis but never actually touch it, extending to infinity in both directions.

Significance of the Z-score:

- A positive Z-score indicates the value is above the mean.
- A negative Z-score indicates the value is below the mean.
- The numerical value tells you exactly how many standard deviations the data point is away from the mean.

Importance:

- Instead of calculating complex integrals for every different dataset, we convert data to Z and use a Standard Normal Table (Z-table) to find probabilities.
- It allows for the comparison of data from different scales. For example, comparing a GRE score (out of 340) with an SAT score (out of 1600) by seeing which has a higher Z-score.
- It is the basis for calculating Confidence Intervals and performing Hypothesis Testing (Z-tests).
- The CLT establishes that as sample sizes grow, the sampling distribution of the mean approaches a normal distribution. Standardizing this result allows us to make predictions about population parameters regardless of the original population's shape.

Question 6: What is the Central Limit Theorem (CLT), and why is it critical in statistics?
Soln. : The Central Limit Theorem states that if you take sufficiently large random samples from any population (regardless of its underlying distribution be it skewed, uniform, or random), the distribution of the sample means will follow a Normal Distribution as the sample size (n) increases.

As the sample size n becomes larger (typically n>30):

1. The mean of the sample means will equal the population mean.
2. As n increases, the spread narrows, meaning our estimate becomes more precise.
3. The shape of the distribution will become a symmetric bell curve (N(0,1) when standardized), even if the original data looks like a "ramp" or "waves."

The CLT is arguably the most important theorem in statistics for these reasons:

- Most real-world data (like wealth distribution or website clicks) is not "Normal" it’s usually messy or skewed. The CLT allows us to use Normal distribution tools (like Z-scores) on that messy data anyway, provided we look at *averages* of samples rather than individual points.
- **Enables Inferential Statistics:** Without the CLT, we could not perform **Hypothesis Testing** or calculate **Confidence Intervals** for general populations.

It gives us the mathematical justification to say, "Based on this sample, we are 95% sure the population mean is X."

- **Predictability:** It tells us that averages are more stable and predictable than individual observations. This is why insurance companies and casinos can predict their long-term profits with near-certainty, even though individual events (a car crash or a slot machine win) are random.

Question 7: What is the significance of confidence intervals in statistical analysis?
Soln. : Confidence Interval (CI) is a range of values, derived from sample data, that is likely to contain the value of an unknown population parameter.
If a point estimate (like a sample mean) is a "single guess," a confidence interval is a "calculated range" that accounts for uncertainty.

- It tells us how much "wiggle room" there is in our data. A narrow interval suggests a very precise estimate, while a wide interval suggests high variability or a small sample size.
- Reporting only a mean (e.g., "The average user stays for 5 minutes") can be misleading. Reporting a CI (e.g., "5 minutes ± 10 seconds") provides the necessary context of reliability.
- In clinical trials or manufacturing, a CI helps determine if a result is practically significant. For example, if a new drug lowers blood pressure by a range of [0.1,10.5], the interval includes values close to zero, suggesting the drug might not be effective enough.

Question 8: What is the concept of expected value in a probability distribution?
Soln. : Expected Value (denoted as E[x]) represents the long-term average or "center of gravity" of a random variable.
If you were to repeat an experiment an infinite number of times, the average of all those outcomes would converge to the expected value.
For Discrete Random Variables: It is the weighted average of all possible outcomes, where each outcome is multiplied by its probability.
For Continuous Random Variables: It is the integral of the value multiplied by the probability density function.

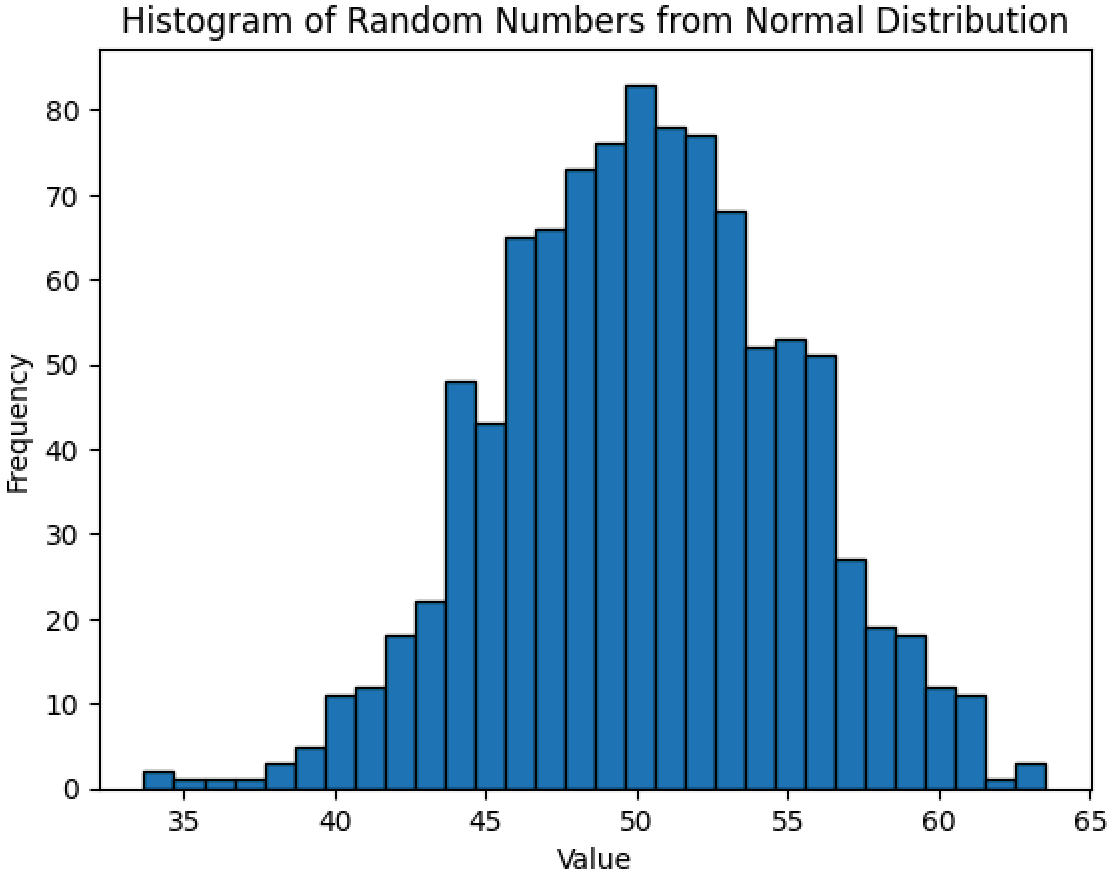
Significance:

1. In finance and insurance, expected value helps determine the "fair price" of a contract or the potential return on an investment.
2. This law states that as the number of trials increases, the actual observed average will get closer and closer to the expected value.
3. You cannot calculate Variance (the spread of data) without first knowing the Expected Value, as variance measures how much outcomes deviate from this mean.

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.
Soln. :

```
import numpy as np
import matplotlib.pyplot as plt
mean = 50
std_dev = 5
random_numbers = np.random.normal(mean, std_dev, 1000)
computed_mean = np.mean(random_numbers)
computed_std_dev = np.std(random_numbers, ddof=1)
plt.hist(random_numbers, bins=30, edgecolor='black')
plt.title('Histogram of Random Numbers from Normal Distribution')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.show()
print(f'Computed Mean: {computed_mean}')
print(f'Computed Standard Deviation: {computed_std_dev}')
```

Output:



```
Computed Mean: 50.23404832881679
Computed Standard Deviation: 4.804729276418653
```

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.
Soln. :
First, we must calculate the basic metrics from the provided `daily_sales` data:
Sample Size (n): 20 days.
Sample Mean (x bar): The sum of sales divided by 20.
Sample Standard Deviation (s): The measure of spread in our 2-year data.

Now we apply CLT

For a 95% confidence level, we want the middle 95% of the standard normal distribution.
This leaves 5% in the tails (2.5% on each side).

Looking at a standard Z-table, the critical value for 95% confidence is 1.96.

Margin of error is calculated: The Margin of Error is the distance from the mean to the edge of our confidence interval.
Then we calculate the Confidence Interval.

```
import numpy as np
import scipy.stats as stats
daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,
                235, 260, 245, 250, 225, 270, 265, 255, 250, 260]
mean_sales = np.mean(daily_sales)
std_dev_sales = np.std(daily_sales, ddof=1)
n = len(daily_sales)
confidence_level = 0.95
z_score = stats.norm.ppf(1 - (1 - confidence_level) / 2)
margin_of_error = z_score * (std_dev_sales / np.sqrt(n))
confidence_interval = (mean_sales - margin_of_error, mean_sales +
margin_of_error)
print(f'Mean Sales: {mean_sales}')
print(f'95% Confidence Interval: {confidence_interval}')
```

```
Mean Sales: 248.25
95% Confidence Interval: (np.float64(240.68326838343515),
np.float64(255.81673161656485))
Margin of Error: 7.566731616564841
```