Statistics Advanced - 1 |Assignment

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

Answer:

A random variable is a function that assigns a numerical value to each outcome in a sample space of a probabilistic experiment. It maps outcomes to real numbers and is used to quantify randomness.

Question 2:What are the types of random variables?

Answer:

Two main types: (a) Discrete random variables — take countable values (e.g., number of heads). (b) Continuous random variables — take values in a continuum (e.g., heights, weights).

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

Answer:

Discrete distributions describe probabilities for countable outcomes and use probability mass functions (PMFs). Continuous distributions describe densities over intervals and use probability density functions (PDFs); probabilities are given by integrals of the PDF.

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

Answer:

Binomial distribution models the number of successes in a fixed number of independent Bernoulli trials with the same success probability p. It's used for yes/no outcomes, e.g., number of defective items in a batch.

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

Answer:

The standard normal distribution is the normal distribution with mean 0 and standard deviation 1. It is important because many other normals can be standardized to it, and it is central to inferential methods (z-scores, tables).

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

Answer:

The Central Limit Theorem states that the distribution of the sample mean approaches a normal distribution as sample size increases (regardless of the population distribution), with mean equal to the population mean and variance equal to population variance divided by n. CLT justifies using normal-based inference for sample means when n is reasonably large.

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

Answer:

Confidence intervals provide a range of plausible values for an unknown population parameter (e.g., mean) constructed so that a specified proportion (e.g., 95%) of such intervals from repeated samples will contain the true parameter.

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

Answer:

The expected value (or expectation) of a random variable is the long-run average value it takes, computed as the weighted average of all possible values with their probabilities (sum for discrete, integral for continuous).

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.

Answer:

Python code:

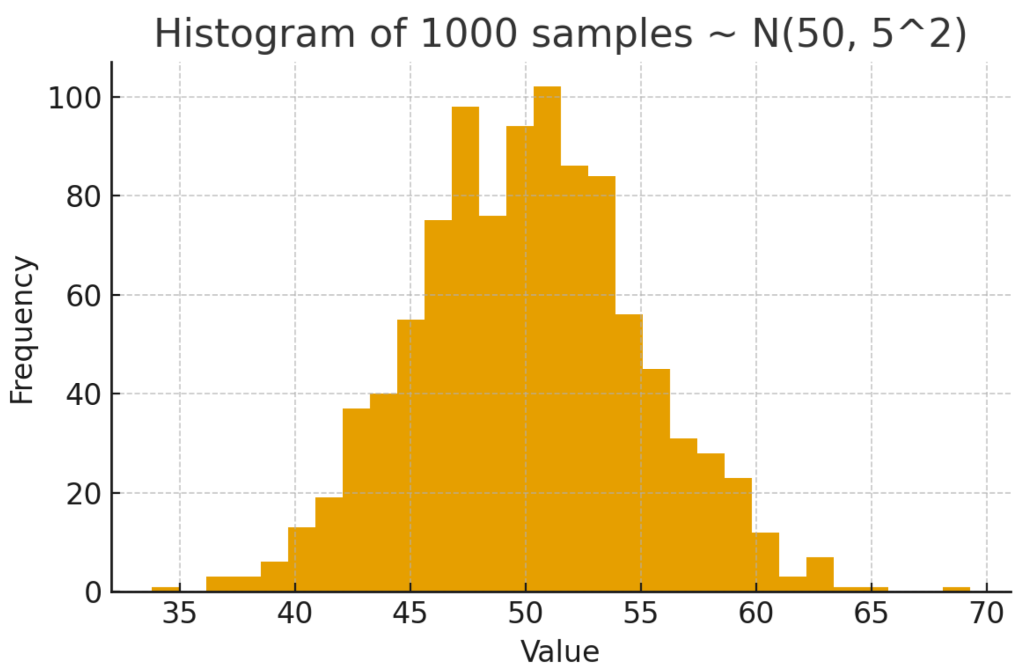
```
import numpy as np
np.random.seed(42)
samples = np.random.normal(loc=50, scale=5, size=1000)
print("Mean:", samples.mean())
print("Std (sample):", samples.std(ddof=1))
# plot histogram using matplotlib
```

Output:

Mean of generated sample (approx): 50.096660

Sample standard deviation (ddof=1) (approx): 4.896080

Histogram:



<u>Question</u> 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 data (provided):
220, 245, 210, 265, 230, 250, 260, 275, 240, 255, 235, 260, 245, 250, 225, 270, 265, 255, 250, 260

Tasks:
- 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.

Answer (Explanation):

Using the Central Limit Theorem: the sampling distribution of the sample mean is approximately normal for reasonably large sample sizes. For small samples (like n=20) we use the t-distribution with df = n-1 to account for extra uncertainty when the population standard deviation is unknown. The 95% confidence interval for the mean is: sample_mean ± t_{0.975, df} * (sample_sd / sqrt(n)).

Python code:

```
from statistics import mean, stdev
from math import sqrt
daily_sales =
[220,245,210,265,230,250,260,275,240,255,235,260,245,250,225,270,265,255,250,260]
n = len(daily_sales)
m = mean(daily_sales)
s = stdev(daily_sales)
se = s / sqrt(n)
# For 95% CI with df=n-1, find t_crit and compute m ± t_crit*se
```

Output:

Sample size (n): 20

Sample mean: 248.250000

Sample standard deviation (ddof=1): 17.265344

Degrees of freedom: 19

T-critical (approx for 95% CI): 2.093024

Standard error (SE): 3.860648

95% Confidence Interval for the mean: (240.169570, 256.330430)