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 function that maps outcomes of a random experiment (elements of a sample space) to real numbers. It provides a numerical description of the outcome. Random variables are conventionally denoted by uppercase letters (e.g., XXX, YYY). They allow probabilities to be assigned to numerical events such as X≤xX \leq xX≤x. Formally, a random variable must be measurable so that events of the form {X≤x}\{X \le x\}{X≤x} are in the probability sigma-algebra.

Question 2: What are the types of random variables?

Answer:

- Discrete random variables — take countable values (finite or countably infinite), e.g., number of heads in 10 coin flips. Characterized by a probability mass function (PMF) P(X=x)P(X=x)P(X=x).

- Continuous random variables — take values on a continuum (an interval or union of intervals). Characterized by a probability density function (PDF) f(x)f(x)f(x) where P(a≤X≤b)=∫abf(x) dxP(a \le X \le b)=\int_a^b f(x)\,dxP(a≤X≤b)=∫abf(x)dx.

- Mixed random variables — have both discrete and continuous components (a point mass plus a density).

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

Answer:

- For **discrete distributions**, probabilities are attached to individual points. The PMF p(x)=P(X=x)p(x)=P(X=x)p(x)=P(X=x) gives probabilities for each possible value. Probabilities of single points can be nonzero. The CDF is a step function.

- For **continuous distributions**, probabilities of single points are zero; probabilities are given over intervals via the PDF f(x)f(x)f(x). The CDF is continuous and differentiable (where PDF exists), and P(X=x)=0P(X=x)=0P(X=x)=0 for any single xxx.

In practice, inference and calculation use sums for discrete and integrals for continuous distributions.

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

Answer:

The binomial distribution models the number of successes in nnn independent Bernoulli trials each with success probability ppp. The PMF is:
P(X=k)=(nk)pk(1−p)n−k,k=0,1,…,n.P(X=k) = \binom{n}{k} p^k (1-p)^{n-k},\qquad k=0,1,\dots,n.P(X=k)=(kn)pk(1−p)n−k,k=0,1,…,n.
It is used to model counts of successes (e.g., number of defective items in a batch, number of heads in coin flips). Mean =np=np=np, variance =np(1−p)=np(1-p)=np(1−p). It's often used for hypothesis testing and confidence intervals for proportions.

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

Answer:

The standard normal distribution is a normal (Gaussian) distribution with mean 000 and variance 111. Its PDF is:
ϕ(z)=12πe−z2/2.\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.ϕ(z)=2π1e−z2/2.
It is important because many statistics (via standardization) reduce to the standard normal; the Central Limit Theorem implies that standardized sums/means approach a normal distribution, allowing use of standard normal tables for inference. Converting an arbitrary normal N(μ,σ2)N(\mu,\sigma^2)N(μ,σ2) to the standard normal via Z=X−μσZ=\frac{X-\mu}{\sigma}Z=σX−μ simplifies probability calculations.

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

Answer:

The Central Limit Theorem states that the sampling distribution of the sample mean (or sum) of i.i.d. random variables with finite mean and variance approaches a normal distribution as sample size nnn grows, regardless of the parent distribution. Concretely, if X1,…,XnX_1,\dots,X_nX1,…,Xn are i.i.d. with mean μ\muμ and variance σ2\sigma^2σ2, then
X‾−μσ/n→dN(0,1)\frac{\bar X - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)σ/nX‾−μdN(0,1)
as n→∞n\to\inftyn→∞. It is critical because it justifies using normal-based inference (confidence intervals, hypothesis tests) for many statistics even when the original data are not normal, provided nnn is sufficiently large.

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

Answer:
A confidence interval (CI) gives a range of plausible values for an unknown population parameter (e.g., mean) based on sample data, together with a confidence level (e.g., 95%). A 95% CI constructed by a procedure means that, under repeated sampling and interval construction, 95% of such intervals will contain the true parameter. CIs quantify sampling uncertainty and are more informative than point estimates because they express precision and reliability.

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

Answer:

The expected value (mean) of a random variable XXX, denoted E[X]E[X]E[X] or μ\muμ, is the long-run average value of XXX over repeated sampling. For a discrete RV: E[X]=∑xx P(X=x)E[X]=\sum_x x\,P(X=x)E[X]=∑xxP(X=x). For a continuous RV: E[X]=∫−∞∞x f(x) dxE[X]=\int_{-\infty}^{\infty} x\,f(x)\,dxE[X]=∫−∞∞xf(x)dx. The expectation summarizes the central tendency and is linear: E[aX+b]=aE[X]+bE[aX+b]=aE[X]+bE[aX+b]=aE[X]+b.

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:

- Sample size = 1000
- Sample mean ≈ **49.7737**
- Sample standard deviation (sample, ddof=1) ≈ **4.9376**

A histogram was also plotted showing a bell-shaped distribution around 50.

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:

- Sample mean sales $\bar{x}$=\bar x =$\bar{x}$= **248.25**
- Sample standard deviation s=s =s= **17.2653**
- Standard error =s/n≈3.8606= s/\sqrt{n} \approx 3.8606=s/n≈3.8606
- t-critical (df = 19, two-sided 95%) ≈\approx≈ **2.0930**
- 95% confidence interval for mean sales: **(240.1696, 256.3304)**

So, with 95% confidence, the true average daily sales lies between about ₹240.17 and ₹256.33 (currency/unit same as sales data).
The Python code used to compute this and print the CI was executed above; the printed numbers are the values shown.