Probability and Statistics - BSc Maths Hons, DU

Probability and Statistics

BSc Maths Hons, DU - Semester I: Exam-Focused Notes

This document provides a comprehensive and detailed overview of essential concepts in probability and statistics, tailored for students of BSc Maths Honours.

1. Descriptive Statistics

Descriptive statistics are used to summarize and describe the main features of a dataset.

Measures of Central Tendency

These indicate the central or typical value of a dataset.

- Mean ($\overline{x}$): The arithmetic average of all values in a dataset.
  - Formula: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - It is sensitive to outliers.
- Median: The middle value of a dataset when it is ordered from least to greatest. If there's an even number of observations, it's the average of the two middle values. It is robust to outliers.
- Mode: The value that appears most frequently in a dataset. A dataset can have one mode (unimodal), more than one mode (multimodal), or no mode.

Measures of Dispersion

These describe the spread or variability of a dataset.

- Variance (σ^2): The average of the squared differences from the mean. It gives an idea of how much individual data points deviate from the mean.
  - Formula: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- Standard Deviation (σ): The square root of the variance. It is in the same units as the data, making it more interpretable than variance.
  - Formula: $\sigma = \sqrt{\sigma^2}$

- Coefficient of Variation (CV): A standardized measure of dispersion that expresses the standard deviation as a percentage of the mean. It is useful for comparing the variability of datasets with different units or vastly different means.
  - Formula: $CV = \frac{\sigma}{\bar{x}} \times 100\%$

PYQ: Calculate mean, variance, and CV for given data.

2. Probability Theory

Probability theory provides a mathematical framework for quantifying uncertainty.

Basics

- Sample Space (S): The set of all possible outcomes of a random experiment.
- Events (A, B): Subsets of the sample space, representing specific outcomes or collections of outcomes.
- Classical Probability: Applicable when all outcomes in the sample space are equally likely.
  - Formula: $P(E) = \frac{|E|}{|S|}$ (where $|E|$ is the number of outcomes in event E, and $|S|$ is the total number of outcomes in the sample space).

Axioms of Probability (Kolmogorov)

These fundamental rules govern any probability assignment:

1. Non-negativity: The probability of any event A is a non-negative real number between 0 and 1, inclusive.
   - $0 \leq P(A) \leq 1$
2. Normalization: The probability of the sample space (the certain event) is 1.
   - $P(S) = 1$
3. Additivity (for mutually exclusive events): For a sequence of mutually exclusive events (events that cannot occur simultaneously), the probability of their union is the sum of their individual probabilities.
   - $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

PYQ: Solve basic probability problems using classical and axiomatic definitions.

3. Conditional Probability & Independence

These concepts explore relationships between events.

Conditional Probability

- The probability of an event A occurring, given that another event B has already occurred.
- Formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ where $P(B) > 0$

Independence

- Two events A and B are independent if the occurrence of one does not affect the probability of the other.
- Mathematically: $P(A \cap B) = P(A)P(B)$

PYQ: Solve for conditional probability and test for independence.

4. Bayes' Theorem

Bayes' Theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It is crucial for updating beliefs.

- Formula: $P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{k=1}^{n} P(A_k)P(B|A_k)}$
  - $P(A_i|B)$: Posterior probability of event A_i given evidence B.
  - $P(B|A_i)$: Likelihood of evidence B given event A_i.
  - $P(A_i)$: Prior probability of event A_i.
  - $\sum_{k=1}^{n} P(A_k)P(B|A_k)$: Total probability of evidence B.

PYQ: Apply Bayes' Theorem for given data, find revised probabilities.

5. Random Variables

A random variable is a variable whose value is subject to variations due to chance (randomness).

Types

- Discrete Random Variable: Takes on a finite or countably infinite number of distinct values (e.g., number of heads in coin flips, number of defects in a sample).
- Continuous Random Variable: Takes on any value within a given interval (e.g., height, weight, time).

Probability Distribution

Describes the possible values a random variable can take and their corresponding probabilities.

- Probability Mass Function (PMF) (for Discrete): For a discrete random variable X, the PMF $P(X = x_i) = p_i$ gives the probability that X takes on a specific value x_i.
- Probability Density Function (PDF) (for Continuous): For a continuous random variable X, the PDF $f(x)$ describes the relative likelihood for the random variable to take on a given value. The probability of X falling within an interval $[a, b]$ is given by the integral of the PDF over that interval: $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Expectation and Variance

- Expectation (E[X] or Mean): The weighted average of all possible values a random variable can take, with the probabilities as weights. It represents the "long-run" average value of the random variable.
  - For Discrete: $E(X) = \sum x_i p_i$
  - For Continuous: $E(X) = \int_{-\infty}^{\infty} x f(x) dx$
- Variance (Var(X)): Measures the spread or dispersion of the random variable's values around its expected value.
  - Formula: $Var(X) = E[(X-\mu)^2] = E(X^2) - [E(X)]^2$

PYQ: Find expectation and variance of given distribution.

6. Discrete Distributions

Binomial Distribution

- Used for a fixed number of independent Bernoulli trials (experiments with only two possible outcomes, success or failure).
- Formula: $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$, where $k = 0, 1, ..., n$
  - n: number of trials
  - k: number of successes
  - p: probability of success on a single trial
- Parameters: Expected Value $E(X) = np$, Variance $Var(X) = np(1-p)$

Poisson Distribution

- Used to model the number of events occurring in a fixed interval of time or space, given a constant average rate of occurrence and independence of events.
- Formula: $P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$, where $k = 0, 1, 2, ...$
  - λ: average number of events in the interval
- Parameters: Expected Value $E(X) = \lambda$, Variance $Var(X) = \lambda$

PYQ: Identify type of distribution, calculate probability, mean, variance.

7. Continuous Distributions

Uniform Distribution

- Describes a scenario where all outcomes within a given interval $[a, b]$ are equally likely.
- Probability Density Function: $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$

Normal Distribution (Gaussian Distribution)

- A symmetric, bell-shaped distribution that is fundamental in statistics due to the Central Limit Theorem. Many natural phenomena approximate this distribution.
- Probability Density Function: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  - μ: mean (location parameter)
  - σ: standard deviation (scale parameter)

Properties of Normal Distribution

- Symmetry: It is symmetric about its mean μ.
- Empirical Rule (or 68-95-99.7 Rule):
  - Approximately 68.26% of the data falls within one standard deviation of the mean. $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6826$
  - Approximately 95.44% of the data falls within two standard deviations of the mean. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9544$
  - Approximately 99.7% of the data falls within three standard deviations of the mean.

PYQ: Apply standard normal distribution table to compute probabilities.

8. Correlation and Regression

These are used to analyze the relationship between two or more variables.

Correlation

- Measures the strength and direction of a linear relationship between two quantitative variables.
- Pearson Correlation Coefficient (r): Ranges from -1 to +1.
  - $r = +1$: Perfect positive linear correlation.
  - $r = -1$: Perfect negative linear correlation.
  - $r = 0$: No linear correlation.
- Formula: $r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$

Regression

- A statistical method used to predict the value of a dependent variable based on the value of one or more independent variables.
- Linear Regression Equations (Line of Best Fit):
  - Line of y on x (predicting y from x): $y - \overline{y} = r\frac{\sigma_y}{\sigma_x}(x - \overline{x})$
  - Line of x on y (predicting x from y): $x - \overline{x} = r\frac{\sigma_x}{\sigma_y}(y - \overline{y})$

- Here, $\overline{x}$ and $\overline{y}$ are the means of x and y respectively, and σ_x and σ_y are their standard deviations.

PYQ: Find regression equations and interpret correlation coefficient.

9. PYQ Practice Topics (Frequent)

The following topics are frequently asked in examinations and require thorough preparation:

Topic	Type of Question
Mean/Variance	Compute for raw/grouped data
Probability	Classical/axiomatic/sampling problems
Conditional Prob.	Compute $P(A\|B)$, test independence
Bayes' Theorem	Apply for revised probabilities
Discrete Distributions	Binomial/Poisson applications
Continuous Dist.	Normal distribution problems
Correlation/Regression	Find coefficients and equations