Probability and Statistics: Detailed Mid-Semester Syllabus Notes

IIT Mandi

March 2, 2025

Contents

1 Descriptive Statistics

Descriptive statistics summarize and organize data so that it can be easily understood. This section covers measures of central tendency, measures of dispersion, and graphical methods.

1.1 Measures of Central Tendency

These measures indicate the center or typical value of a dataset.

- **Mean:** The arithmetic average of a dataset.

$$\text{Mean } (\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- **Median:** The middle value when the data is ordered. If n is even, it is the average of the two middle numbers.
- **Mode:** The most frequently occurring value in the dataset.

Example: For the dataset $\{4, 8, 6, 5, 3, 7, 8\}$:

- Mean: $\frac{4+8+6+5+3+7+8}{7} = \frac{41}{7} \approx 5.86$.
- Median: When sorted $\{3, 4, 5, 6, 7, 8, 8\}$, the median is 6.
- Mode: The value 8 occurs most frequently.

1.2 Measures of Dispersion

These measures describe the spread of data.

- **Range:** Difference between the maximum and minimum values.

$$\text{Range} = \text{Max} - \text{Min}.$$

- **Variance:** The average of the squared differences from the mean.

$$\text{Variance } (s^2) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad \text{(sample variance)}.$$

- **Standard Deviation:** The square root of the variance.

$$s = \sqrt{s^2}.$$

- **Interquartile Range (IQR):** The difference between the first quartile (Q_1) and the third quartile (Q_3).

$$\text{IQR} = Q_3 - Q_1.$$

Example: For the sorted dataset $\{3, 4, 5, 6, 7, 8, 8\}$:

- Range: $8 - 3 = 5$.
- To find the IQR, split the data at the median (6). The lower half is $\{3, 4, 5\}$ and the upper half is $\{7, 8, 8\}$. Thus, $Q_1 = 4$ and $Q_3 = 8$. So, $\text{IQR} = 8 - 4 = 4$.

1.3 Graphical Representations

Visual methods help understand the data's distribution.

- **Histogram:** A bar graph representing the frequency distribution.
- **Box Plot:** Displays the median, quartiles, and potential outliers.
- **Scatter Plot:** Shows the relationship between two quantitative variables.

1.4 Practice Questions

Q1: Given the dataset $\{12, 15, 12, 18, 20, 22, 22, 19, 17, 14\}$, calculate the mean, median, mode, range, and standard deviation.

Q1: Draw a box plot for the dataset $\{4, 8, 15, 16, 23, 42\}$ and identify the quartiles.

2 Basics of Probability Theory

Probability theory is the foundation of statistics. It helps quantify uncertainty using mathematical principles.

2.1 Probability Axioms (Kolmogorov's Axioms)

Any probability measure P on a sample space Ω must satisfy:

1. **Non-negativity:** $P(A) \geq 0$ for any event $A \subseteq \Omega$.
2. **Normalization:** $P(\Omega) = 1$.
3. **Additivity:** If $A_1, A_2, \ldots$ are mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

2.2 Derived Properties

- **Complement Rule:** $P(A^c) = 1 - P(A)$.
- **Inclusion-Exclusion Principle:** For any two events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

2.3 Example Problem

Problem: If $P(A) = 0.3$ and $P(B) = 0.5$ with $P(A \cap B) = 0.15$, find $P(A \cup B)$.
Solution:

$$P(A \cup B) = 0.3 + 0.5 - 0.15 = 0.65.$$

2.4 Practice Questions

Q2: Given $P(A) = 0.4$ and $P(B) = 0.6$ with $P(A \cap B) = 0.25$, calculate $P(A^c)$ and $P(A \cup B)$.

Q2: If $P(A \cup B) = 0.8$ and $P(A) = 0.5$ and $P(B) = 0.5$, determine $P(A \cap B)$.

3 Conditional Probability and Bayes' Theorem

Conditional probability helps update the likelihood of events given new information.

3.1 Conditional Probability

The conditional probability of A given B (with $P(B) > 0$) is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

3.2 Bayes' Theorem

Bayes' theorem allows us to update probabilities as new evidence is available:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where the total probability of B is given by:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

3.3 Worked Example

Scenario: A disease affects 1% of a population. A test detects the disease correctly 90% of the time (true positive) and falsely indicates disease 5% of the time (false positive).
Task: Find the probability that a person has the disease given a positive test result.
Solution:

$$\begin{aligned} P(D) = 0.01, \quad P(T|D) = 0.9, \quad P(T|D^c) = 0.05. \\ P(T) = 0.9(0.01) + 0.05(0.99) = 0.009 + 0.0495 = 0.0585, \\ P(D|T) = \frac{0.9 \times 0.01}{0.0585} \approx 0.1538 \quad (15.38\%). \end{aligned}$$

3.4 Practice Questions

Q3: In a factory, 2% of products are defective. A test identifies defects with 95% accuracy but has a 3% false alarm rate. Calculate the probability that a product is actually defective given a positive test.

Q3: Given $P(A) = 0.4$, $P(B) = 0.5$ and $P(B|A) = 0.6$, find $P(A|B)$.

4 Independence

Two events A and B are independent if the occurrence of one does not affect the occurrence of the other.

$$\text{Definition: } P(A \cap B) = P(A)P(B).$$

4.1 Example

Problem: Suppose $P(A) = 0.3$ and $P(B) = 0.4$. If A and B are independent, then:

$$P(A \cap B) = 0.3 \times 0.4 = 0.12.$$

4.2 Practice Question

Q4: If two dice are rolled, are the events "the first die shows a 3" and "the sum of the dice is 7" independent? Justify your answer.

5 Counting Techniques: Permutations and Combinations

Counting techniques are essential for determining the number of possible outcomes.

5.1 Permutations

Permutations count the number of ways to arrange n distinct items. For n items:

$$n! = n \times (n-1) \times \cdots \times 1.$$

For arranging r items from n:

$$P(n,r) = \frac{n!}{(n-r)!}.$$

5.2 Combinations

Combinations count the number of ways to choose r items from n when order does not matter:

$$C(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

5.3 Examples

- **Permutation Example:** How many ways can you arrange 3 out of 5 books?

$$P(5,3) = \frac{5!}{(5-3)!} = \frac{120}{2} = 60.$$

- **Combination Example:** How many ways can you select 3 students from a class of 10?

$$C(10,3) = \frac{10!}{3!7!} = 120.$$

5.4 Practice Questions

Q5: A committee of 4 is to be formed from 12 candidates. How many different committees are possible?

Q5: In how many ways can the letters of the word *STATISTICS* be arranged (note: some letters repeat)?

6 Bernoulli and Binomial Distributions

6.1 Bernoulli Distribution

A Bernoulli distribution describes a single trial that has only two outcomes: success (with probability p) or failure (with probability $1-p$). Its probability mass function (pmf) is:

$$P(X=1) = p, \quad P(X=0) = 1-p.$$

6.2 Binomial Distribution

The binomial distribution models the number of successes in n independent Bernoulli trials.

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

Key properties:

- **Mean:** $\mu = np$.
- **Variance:** $\sigma^2 = np(1-p)$.

6.3 Worked Example

Problem: Suppose a basketball player has a 70% chance of making a free throw. If she takes 10 free throws, what is the probability that she makes exactly 7?
Solution:

$$P(X=7) = \binom{10}{7}(0.7)^7(0.3)^3.$$

Compute the combination:

$$\binom{10}{7} = \frac{10!}{7!3!} = 120.$$

Thus,

$$P(X=7) = 120 \times (0.7)^7 \times (0.3)^3.$$

6.4 Practice Questions

Q6: In a manufacturing process with a 5% defect rate, what is the probability of finding exactly 2 defective items in a random sample of 20?

Q6: If a fair coin is tossed 8 times, what is the probability of getting exactly 5 heads?

7 Random Variables, Cumulative Distribution Functions (CDF), and Probability Density Functions (PDF)

7.1 Random Variables

A random variable assigns a numerical value to each outcome in the sample space. It can be discrete (taking countable values) or continuous (taking any value in an interval).

7.2 Cumulative Distribution Function (CDF)

For a random variable X, the CDF is defined as:

$$F_X(x) = P(X \leq x).$$

For discrete variables, it is a step function; for continuous variables, it is a continuous, non-decreasing function.

7.3 Probability Density Function (PDF)

For a continuous random variable X, the PDF $f_X(x)$ satisfies:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt,$$

with the property:

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1.$$

Example (Uniform Distribution): If X is uniformly distributed on $[a, b]$, then:

$$f_X(x) = \frac{1}{b-a} \quad \text{for } a \leq x \leq b,$$

and the CDF is:

$$F_X(x) = \frac{x-a}{b-a} \quad \text{for } a \leq x \leq b.$$

7.4 Example Problem

Problem: Consider a continuous random variable X with PDF:

$$f_X(x) = \begin{cases} 2x, & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

1. Verify that $f_X(x)$ is a valid PDF.
2. Find the CDF $F_X(x)$.

Solution:

1. Verify normalization:

$$\int_0^1 2x\,dx = x^2\big|_0^1 = 1.$$

2. For $0 \leq x \leq 1$:

$$F_X(x) = \int_0^x 2t\,dt = x^2.$$

For $x < 0$, $F_X(x) = 0$; for $x > 1$, $F_X(x) = 1$.

7.5 Practice Questions

Q7: A random variable Y has the PDF $f_Y(y) = 3y^2$ for $0 \leq y \leq 1$. Find the CDF $F_Y(y)$.

Q7: If X is normally distributed with mean 50 and standard deviation 5, what is the probability $P(45 \leq X \leq 55)$ using the properties of the normal distribution?

8 Summary and Final Practice

8.1 Key Concepts to Remember

- **Descriptive Statistics:** Measures of central tendency (mean, median, mode) and dispersion (range, variance, standard deviation, IQR).
- **Probability Theory:** Fundamental axioms, the complement rule, and inclusion-exclusion principle.
- **Conditional Probability:** How new information changes the probability and the use of Bayes' theorem.
- **Independence:** Understanding when events do not affect each other.
- **Counting Techniques:** Permutations for ordered arrangements and combinations for selections.
- **Distributions:** Bernoulli and binomial distributions for discrete outcomes; understanding PDFs and CDFs for continuous variables.

8.2 General Practice Questions

Q8: For a dataset, explain why the median might be a better measure of central tendency than the mean.

Q8: How does the inclusion-exclusion principle help in calculating the probability of the union of two events?

Q8: In your own words, describe the difference between discrete and continuous random variables.

Q8: Solve a real-world problem: A survey shows 60% of people prefer tea over coffee. If 10 people are selected at random, what is the probability that exactly 7 prefer tea?