Statistics Exam Prep Study Notes

Table of Contents

Basic Statistical Concepts

Key Framework

Statistics → Learning from Data

- **Collection → Description → Analysis → Conclusion**

Population vs Sample

- **Population**: Large, unknown group we want to study
- **Sample**: Smaller subset used for analysis
- **Inference**: Drawing conclusions about population from sample data

Important Note

To draw valid inferences about a large population, we need **random** (unbiased) samples.

Sampling Methods

1. Random Sampling

- Each entity in population has **equal chance** to enter sample
- Use depends on experiment goals
- **Example**: Selecting students for height study at IIT

2. Stratified Sampling

- Taking samples from different **categories/strata**
- Each stratum gets random sampling
- **Example**: Crop study with 50% rice, 30% wheat, 20% others
  - If n=500: Rice=250, Wheat=150, Others=100

3. Sequential Sampling

- Used when **time and cost** are critical factors
- **Example**: Bulb factory defect testing - test one by one until decision can be made

---

Data Types

Main Categories

Numerical Data

- **Discrete**: Countable values (tickets sold, students, balls bowled)
- **Continuous**: Can take decimal values (height, weight, age)

Categorical Data

- **Nominal**: Categories with no order (T/F, M/F, hair color, religion)
- **Ordinal**: Categories with order (customer rating, award categories)

Measurement Levels

Level	Ordering	Equal Intervals	True Zero	Arithmetic
Nominal	No	No	No	None
Ordinal	Yes	No	No	Limited
Interval	Yes	Yes	No	+/-
Ratio	Yes	Yes	Yes	×/÷

Key Examples:

- **Interval**: Temperature (0°C ≠ absence of temperature), IQ scores, dates
- **Ratio**: Height, weight, age (true zero exists)

Descriptive Statistics

Frequency Representations

1. **Frequency Table**: xi | fi
2. **Relative Frequency**: fi/n
3. **Cumulative Frequency**: Ci

Graphical Representations

- **Line Graph**: Points connected with lines
- **Bar Graph**: Discrete bars
- **Frequency Polygon**: Connected frequency points
- **Histogram**: For grouped/continuous data
- **Pie Charts**: For categorical data

For Large Datasets

- **Class Intervals**: Group data into ranges
- **Histogram**: Bar graph for grouped data
- **Ogive**: Cumulative frequency plot
- **Stem-and-Leaf Plot**: For small/medium datasets

Measures of Central Tendency & Spread

Central Tendency

Sample Mean: $\bar{x} = \Sigma xi/n$

Sample Median:

- If n is odd: $x(n+1)/2$
- If n is even: $(xn/2 + x(n+1)/2)/2$
- *Data must be sorted first*

Sample Mode: Data value with maximum frequency

- Can be multimodal

Spread (Variability)

Sample Variance: $s^2 = \Sigma(xi - \bar{x})^2/(n-1) = [\Sigma xi^2 - n\bar{x}^2]/(n-1)$

Sample Standard Deviation: $s = \sqrt{s^2}$

Properties of Linear Transformations

If $y_i = ax_i + b$, then:

- $\bar{y} = a\bar{x} + b$
- $s_y^2 = a^2 s_x^2$

Percentiles and Box Plots

Percentile Calculation

For 100p percentile:

1. Ensure data is sorted
2. Find $n \cdot p$
3. If $n \cdot p \notin \mathbb{N}$: Take $T[np]$ (round up)
4. If $n \cdot p \in \mathbb{N}$: Take $(T[np] + T[np+1])/2$

Box Plot Components

- **Q1**: First quartile (25th percentile)
- **Q2**: Median (50th percentile)
- **Q3**: Third quartile (75th percentile)
- **Range**: $T_n - T_1$
- **IQR**: $Q_3 - Q_1$ (Interquartile Range)

Inequalities

Chebyshev's Inequality: For any dataset, at least $(1 - 1/k^2) \times 100\%$ of data lies within $(\bar{x} \pm ks)$

Empirical Rule (for approximately normal distributions):

- $\bar{x} \pm s$: ~68% of data
- $\bar{x} \pm 2s$: ~95% of data
- $\bar{x} \pm 3s$: ~99.7% of data

Probability Theory

Basic Concepts

- **Sample Space (S)**: All possible outcomes
- **Events (A, B, C)**: Subsets of sample space
- **$P(A) = n(A)/n(S)$** (counting method)

Key Relationships

- **Mutually Exclusive**: $P(A \cap B) = 0$
- **Independent**: $P(A \cap B) = P(A) \cdot P(B)$
- **Addition Rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Complement**: $P(Ac) = 1 - P(A)$

Conditional Probability

$P(A|B) = P(A \cap B)/P(B)$

Bayes' Formula: $P(A \cap B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B)$

Random Variables

Types

- **Discrete**: Takes sequence of values (finite or countably infinite)
- **Continuous**: Takes continuum of values in intervals

Distribution Functions

CDF: $F(x) = P(X \leq x)$

- **Useful**: $P(a < X \leq b) = F(b) - F(a)$

For Discrete RV:

- **PMF**: $p(x) = P(X = x)$
- **$F(a) = \Sigma p(x)$** for $x \leq a$

For Continuous RV:

- **PDF**: $f(x)$ where $\int f(x)dx = 1$
- **$F(a) = \int_{-\infty}^{a} f(x)dx$**
- **$P(X = a) = 0$** (probability of exact point is zero)

Expectation and Variance

Expectation:

- Discrete: $E(X) = \Sigma xi \cdot p(xi)$
- Continuous: $E(X) = \int x \cdot f(x)dx$

Properties:

- $E(aX + b) = aE(X) + b$
- $E(X_1 + X_2 + \ldots + Xn) = \Sigma E(Xi)$

Variance: $V(X) = E(X^2) - [E(X)]^2$

Properties:

- $V(aX + b) = a^2V(X)$
- $V(X + Y) = V(X) + V(Y)$ if X, Y independent

Sample Problems with Solutions

Problem 1: Sampling

Question: A university has 10,000 students: 6,000 undergraduates and 4,000 graduates. Design a stratified sample of size 500.

Solution:

- Undergraduate proportion: 6,000/10,000 = 0.6
- Graduate proportion: 4,000/10,000 = 0.4
- Undergraduate sample: 500 × 0.6 = 300
- Graduate sample: 500 × 0.4 = 200

Problem 2: Descriptive Statistics

Question: Data set: {2, 4, 4, 6, 8, 10, 12}. Find mean, median, mode, and standard deviation.

Solution:

- **Mean**: $\bar{x} = (2+4+4+6+8+10+12)/7 = 46/7 \approx 6.57$
- **Median**: Middle value = 6 (4th position)
- **Mode**: 4 (appears twice)
- **Variance**: $s^2 = [(2-6.57)^2 + (4-6.57)^2 + (4-6.57)^2 + (6-6.57)^2 + (8-6.57)^2 + (10-6.57)^2 + (12-6.57)^2]/(7-1)$
  - $s^2 = [20.88 + 6.60 + 6.60 + 0.32 + 2.04 + 11.76 + 29.49]/6 = 77.69/6 \approx 12.95$
- **Standard Deviation**: $s = \sqrt{12.95} \approx 3.60$

Problem 3: Probability

Question: In a deck of 52 cards, what's the probability of drawing a red card or a face card?

Solution:

- $P(Red) = 26/52 = 1/2$
- $P(Face) = 12/52 = 3/13$
- $P(Red \cap Face) = 6/52 = 3/26$ (red face cards)
- $P(Red \cup Face) = P(Red) + P(Face) - P(Red \cap Face)$
- $P(Red \cup Face) = 26/52 + 12/52 - 6/52 = 32/52 = 8/13$

Problem 4: Random Variables

Question: Let X be the number of heads in 3 coin flips. Find E(X) and V(X).

Solution:

- X can take values: 0, 1, 2, 3
- $P(X=0) = 1/8$, $P(X=1) = 3/8$, $P(X=2) = 3/8$, $P(X=3) = 1/8$
- **$E(X)$** $= 0\times(1/8) + 1\times(3/8) + 2\times(3/8) + 3\times(1/8) = 12/8 = 1.5$
- **$E(X^2)$** $= 0^2\times(1/8) + 1^2\times(3/8) + 2^2\times(3/8) + 3^2\times(1/8) = 24/8 = 3$
- **$V(X)$** $= E(X^2) - [E(X)]^2 = 3 - (1.5)^2 = 3 - 2.25 = 0.75$

Problem 5: Correlation

Question: Given data points (1,2), (2,4), (3,5), (4,7), find the correlation coefficient.

Solution:

- $\bar{x} = (1+2+3+4)/4 = 2.5$, $\bar{y} = (2+4+5+7)/4 = 4.5$
- $sx^2 = [(1-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (4-2.5)^2]/3 = 5/3$
- $sy^2 = [(2-4.5)^2 + (4-4.5)^2 + (5-4.5)^2 + (7-4.5)^2]/3 = 23/3$
- $\Sigma(xi-\bar{x})(yi-\bar{y}) = (-1.5)(-2.5) + (-0.5)(-0.5) + (0.5)(0.5) + (1.5)(2.5) = 8$
- **$r = \Sigma(xi-\bar{x})(yi-\bar{y})/[(n-1)sxsy] = 8/[3\times\sqrt{(5/3)}\times\sqrt{(23/3)}] \approx 0.982$**

Exam Tips

1. **Always check if data needs to be sorted** (for median, percentiles)
2. **Read probability problems carefully** - distinguish between "and" (intersection) vs "or" (union)
3. **For random variables, verify if discrete or continuous** before choosing formulas
4. **Remember the n-1 correction** for sample variance
5. **Use Chebyshev when distribution unknown**, Empirical Rule only for normal distributions
6. **In correlation problems, correlation ≠ causation**
7. **For conditional probability, clearly define events** before applying formulas