WEEK 3 Lecture Transcripts explanation:

Statistical Modelling

## **Introduction to Random Variables and Probability**

* Random variables characterize random phenomena using probability measures and probability density functions
* Deterministic phenomena have predictable outcomes, while stochastic phenomena have multiple possible outcomes with limited confidence
* Modeling and measurement errors lead to the use of probability density functions to predict outcomes with confidence intervals

## **Types of Random Phenomena**

* Discrete random variables have finite outcomes, like coin tosses or dice rolls
* Continuous random variables have infinite outcomes, like body temperature measurements

## **Probability Concepts**

* Probability assigns real values between 0 and 1 to outcomes, with the sum of all outcomes equaling 1
* Probabilities can be interpreted as frequencies in the limit as the number of trials goes to infinity
* Independent events have no influence on each other's occurrence
* Mutually exclusive events preclude each other's occurrence

## **Probability Rules**

* The probability of a complement is 1 minus the probability of the event
* If event B is a subset of event A, then P(B) ≤ P(A)
* P(A or B) = P(A) + P(B) - P(A and B), for non-mutually exclusive events
* Conditional probability P(B|A) = P(A and B) / P(A), assuming P(A) > 0
* Bayes' rule: P(A|B)P(B) = P(B|A)P(A)

## **Example: Manufacturing Process**

* A manufacturing process produces 1000 parts, with 50 being defective
* Randomly selecting a part, the probability of picking a defective part is 50/1000
* If the first part picked is defective, the probability of the second part being defective changes to 49/999
* If the first part picked is non-defective, the probability of the second part being defective remains at 50/999

## **Introduction to Random Variables**

* Random variables map outcomes from a sample space to real numbers, allowing numerical computations.
* For example, in a coin toss, heads (H) can be mapped to 0 and tails (T) to 1.

## **Types of Random Variables**

* **Discrete Random Variables**: Map outcomes of discrete phenomena (e.g., coin tosses) to real numbers.
* **Continuous Random Variables**: Map outcomes of continuous phenomena (e.g., measurements) to intervals on the real line.

## **Probability Mass and Density Functions**

* **Probability Mass Function (PMF)**: Used for discrete random variables, assigning probabilities to specific outcomes.
* **Probability Density Function (PDF)**: Used for continuous random variables, defining probabilities over intervals. The probability that a continuous random variable lies within an interval [�,�][*a*,*b*] is calculated as the integral of the PDF from �*a* to �*b*.

## **Cumulative Density Function (CDF)**

* The CDF, denoted as �(�)*F*(*b*), represents the probability that a random variable �*X* is less than or equal to �*b*. It is defined as the integral of the PDF from −∞−∞ to �*b*.

## **Example: Coin Tosses**

* In �*n* coin tosses, the probability of getting exactly �*k* heads can be computed using the binomial distribution. The formula involves combinations and the probabilities of heads and tails.

## **Normal Distribution**

* The normal distribution is characterized by its bell-shaped curve and is defined by two parameters: mean (�*μ*) and variance (�2*σ*2).
* The standard normal distribution is a special case where �=0*μ*=0 and �=1*σ*=1.

## **Moments of a Distribution**

* Moments are statistical measures that describe the shape of the distribution. The first moment is the mean, and the second moment is the variance.
* The variance is calculated as the expectation of the squared deviation from the mean.

## **Joint Distribution**

* The joint distribution of two random variables is described by a joint probability density function, which allows for the computation of probabilities involving both variables.

## **Functions for Probability Computation**

* Functions like pnorm and qnorm in statistical software compute probabilities and quantiles for normal distributions, respectively. These functions can also be adapted for other distributions like chi-squared and exponential.

This structured overview provides a concise summary of the key concepts discussed in the lecture, suitable for pasting and reference.

Key Points from the Lecture

Introduction to Statistical Analysis

* Probability provides a theoretical framework for performing statistical analysis of data
* Statistical analysis deals with the analysis of experimental observations
* When analyzing data, we typically do not know the entire sample space or the parameters of the distribution

Sampling and Inferences

* The sample set should be sufficiently representative of the entire population
* Inferences drawn from the sample set are uncertain and should be accompanied by confidence intervals
* Statistical analysis can be divided into graphical analysis (using plots and graphs) and numerical computations (using summary statistics)

Measures of Central Tendency

1. **Mean (x̄)**: The sum of all data points divided by the number of data points
   * An unbiased estimate of the population mean μ
   * Susceptible to outliers
2. **Median**: The value below which 50% of the data points lie and above which 50% lie
   * A robust measure against outliers
   * The best estimate in terms of minimizing absolute deviations
3. **Mode**: The value that occurs most often or the most probable value

Measures of Spread

1. **Sample Variance (s²)**: The sum of squared deviations from the mean, divided by N-1
   * An unbiased estimate of the population variance
   * Susceptible to outliers
2. **Mean Absolute Deviation**: The sum of absolute deviations from the mean, divided by N or N-1
   * A more robust measure against outliers
3. **Range**: The difference between the maximum and minimum values

Properties of Sample Mean and Variance

* If observations are drawn from a normal distribution with parameters μ and σ², and x̄ is the sample mean:
  + E[x̄] = μ (unbiased estimate)
  + Var[x̄] = σ²/N (variance decreases with increasing sample size)
* If s² is the sample variance:
  + (N-1)s²/σ² follows a χ² distribution with N-1 degrees of freedom

Pastable Output

text

**x̄ = ∑x\_i / N # Sample mean**

**median = middle value when data is ordered # Value below/above which 50% data lies**

**mode = most frequent value # Most probable value**

**s² = ∑(x\_i - x̄)² / (N-1) # Sample variance**

**MAD = ∑|x\_i - x̄| / N # Mean absolute deviation**

**range = max(x) - min(x) # Difference between max and min values**

**If x\_i ~ N(μ, σ²):**

**E[x̄] = μ # Mean of sample mean is population mean**

**Var[x̄] = σ²/N # Variance of sample mean decreases with N**

**(N-1)s²/σ² ~ χ²(N-1) # Sample variance follows scaled χ² distribution**

## Notes on Hypothesis Testing

### Introduction

- Hypothesis testing is a statistical method for decision-making based on data.

- It involves assessing a hypothesis about a population parameter (mean, variance, etc.) using sample data.

### Key Steps in Hypothesis Testing

1. \*\*Identify Parameter of Interest\*\*: Determine what you want to test (e.g., mean, variance).

2. \*\*Formulate Hypotheses\*\*:

- \*\*Null Hypothesis (H0)\*\*: The default assumption (e.g., no effect or no difference).

- \*\*Alternative Hypothesis (H1)\*\*: What you want to prove (e.g., there is an effect or difference).

3. \*\*Collect Data\*\*: Gather sample data relevant to the hypothesis.

4. \*\*Compute Test Statistic\*\*: A function of the sample data (e.g., sample mean, sample variance).

5. \*\*Determine Distribution\*\*: Find the distribution of the test statistic under the null hypothesis.

6. \*\*Set Threshold/Criterion\*\*: Choose a significance level (α) to determine when to reject H0.

7. \*\*Make a Decision\*\*: Compare the test statistic to the threshold to decide whether to reject H0.

### Types of Hypothesis Tests

- \*\*Two-Sided Test\*\*: Tests if the parameter is different from a specified value (e.g., H0: μ = 0 vs. H1: μ ≠ 0).

- \*\*One-Sided Test\*\*: Tests if the parameter is greater than or less than a specified value (e.g., H0: μ = 0 vs. H1: μ > 0).

### Errors in Hypothesis Testing

- \*\*Type I Error (α)\*\*: Rejecting H0 when it is true (false positive).

- \*\*Type II Error (β)\*\*: Not rejecting H0 when H1 is true (false negative).

- \*\*Statistical Power\*\*: The probability of correctly rejecting H0 when H1 is true (1 - β).

### Trade-offs

- Reducing Type I error increases Type II error and vice versa.

- The choice of significance level affects the sensitivity of the test.

### Example: Solid Propellant Burning Rate

- A manufacturer tests if a solid propellant burns at 50 cm/s.

- Collects 25 samples from the mixing bowl and computes the sample mean.

- Decision made based on whether the sample mean significantly differs from 50 cm/s.

### Summary

- Hypothesis testing is essential for making data-driven decisions.

- Understanding the types of tests, potential errors, and trade-offs is crucial for effective analysis.

Citations:

[1] https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31453066/a1c65552-f1a0-49a6-92ff-cf364e5a2205/paste.txt

[2] https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31453066/45a7d74e-ba9a-4ff5-9a83-d31909f27115/paste.txt