Advanced Data Science

Statistics

Session 3

Kiran Waghmare
Program Manager
C-DAC Mumbai

Agenda

- **Statistics**

Unit I:
Advance Methods of Data Science and Algorithms
Statistics
Central Limit Theorem;
AB Testing;
Linear Regression

Fundamentals of Statistics

- Data science refers to dealing with data.
- Statistical analysis helps in enhancing predictability, pattern analysis, and concluding and interpreting the data.
- The two fundamental statistics concepts that play a key role in data science are descriptive and inferential statistics.

Statistics

1. Overview of Statistics

- **Definition:** Branch of mathematics for collecting, analyzing, interpreting, and presenting data.
- **Importance:** Essential for extracting insights, making predictions, and uncovering patterns in data science.

2. Types of Statistics

- **Descriptive Statistics:** Summarizes data to provide insights into the data set.
- **Inferential Statistics:** Makes predictions or inferences about a population based on a sample.

3. Descriptive Statistics

Measures of Central Tendency

- Mean (μ):

$$\text{Mean} = \frac{\text{Sum of Values}}{\text{Number of Values}}$$

- **Median**: Middle value when data is sorted.
  - For odd number of data points:

$$\text{Median} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ value}$$

  - For even number of data points:

$$\text{Median} = \text{Average of } \left(\frac{n}{2}\right)^{\text{th}} \text{ value and next value}$$

- **Mode**: Most frequently occurring value.

```
import numpy as np

data = [10, 20, 30, 40, 50]
mean = np.mean(data)
print("Mean:", mean)
```

```
variance = np.var(data)
print("Variance:", variance)
```

```
median = np.median(data)
print("Median:", median)
```

```
from scipy import stats

mode = stats.mode(data)
print("Mode:", mode)
```

Measures of Dispersion

- **Range**: Difference between maximum and minimum values.
- **Mean Absolute Deviation (MAD)**:

$$\text{MAD} = \frac{\sum |X_i - \bar{X}|}{n}$$

- **Standard Deviation (σ)**:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{n}}$$

- **Variance (σ^2)**:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{n}$$

- **Interquartile Range (IQR)**:

$$\text{IQR} = Q3 - Q1$$

- **Coefficient of Variation (CV)**:

$$\text{CV} = \left(\frac{\sigma}{\mu}\right) \times 100$$

- **Z-score**:

$$Z = \frac{X - \mu}{\sigma}$$

Measures of Shape

- **Kurtosis**: Measures the "tailedness" of the distribution.
- **Skewness**: Measures the asymmetry of the distribution.
    - **Positive Skew**: Right tail is longer (Mean > Median).
    - **Negative Skew**: Left tail is longer (Mean < Median).
    - **Zero Skew**: Symmetric distribution (Mean = Median).

4. Covariance and Correlation

- Covariance (Cov(x, y)):

$$\text{Cov}(x, y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

- Correlation (ρ(X, Y)):

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

5. Regression Analysis

- Regression Coefficient (β):

$$y = \alpha + \beta x$$

$$\beta = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

6. Probability and Distributions

Probability Functions

- **Probability Mass Function (PMF)**: For discrete variables.
- **Probability Density Function (PDF)**: For continuous variables.
- **Cumulative Distribution Function (CDF)**: Probability that a variable takes a value ≤ x.

Bayes' Theorem

- **Bayes' Theorem**:

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

Probability Distributions

- Uniform Distribution:

$$f(X) = \frac{1}{b-a}$$

- Binomial Distribution:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Poisson Distribution:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

- Normal Distribution:

$$f(X|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5\left(\frac{X-\mu}{\sigma}\right)^2}$$

7. Central Limit Theorem (CLT)

- CLT: Sample mean distribution approaches normality as sample size increases.

8. Hypothesis Testing

- **Null Hypothesis (H_0)**: No effect or difference.
- **Alternative Hypothesis (H_1)**: Indicates effect or difference.
- **Type I Error (α)**: False positive.
- **Type II Error (β)**: False negative.
- **p-value**: Probability of obtaining the observed result under H_0.
- **Confidence Interval (CI)**: Range in which the population parameter likely lies.

9. Parametric Tests

- **Z-test**:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

- **T-test**:

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

- **F-test**: Compares variances of two samples.

$$F = \frac{s_1^2}{s_2^2}$$

- **ANOVA**: Analyzes differences among group means.

10. Non-Parametric Tests

- Chi-Squared Test (χ^2):

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- **Mann-Whitney U Test**: Compares two independent groups.
- **Kruskal-Wallis Test**: Compares three or more groups.

11. A/B Testing (Split Testing)

- Compares two versions to determine which one performs better.

These bullet points cover the essentials of statistical concepts, measures, and tests commonly used in data science.