1. Course Introduction and Overview

Purpose of the Course:

The **Introduction to Statistical Methods** course introduces students to fundamental concepts in probability and statistics, essential for data analysis, decision-making, and predictive modeling. The course covers essential statistical techniques for analyzing and interpreting data and includes practical applications.

Key Concepts:

- **Probability**: Understanding how likely events are to occur.
- **Statistics**: Organizing, analyzing, and interpreting data.
- **Inference**: Making predictions or conclusions about a population based on a sample.

2. Course Modules

The course is divided into six major modules:

1. **Basic Probability & Statistics**
   Introduces basic concepts of probability, data collection, and basic statistics techniques.
2. **Conditional Probability & Bayes' Theorem**
   Focuses on conditional probability and how to calculate probabilities in complex situations using Bayes' Theorem.
3. **Probability Distributions**
   Understanding the different types of probability distributions (e.g., normal, binomial, etc.) used to model various phenomena.
4. **Hypothesis Testing**
   Methods for testing statistical hypotheses and drawing conclusions from data.
5. **Prediction & Forecasting**
   Focuses on predictive modeling techniques used in machine learning and statistics.
6. **Prediction & Forecasting using Gaussian Mixture Models and Expectation Maximization**
   Advanced techniques for statistical prediction, particularly in handling complex datasets.

3. Textbooks and References

- **T1**: *Statistics for Data Scientists* by Maurits Kaptein et al. (2022)
- **T2**: *Probability and Statistics for Engineering and Sciences* by Jay L. Devore (8th Edition)
- **T3**: *Introduction to Time Series and Forecasting* by Peter J. Brockwell & Richard A. Davis (2nd Edition)

These textbooks provide in-depth coverage of the course material and are excellent resources for further study.

---

4. Data Variables

Understanding data variables is crucial for proper data analysis:

- **Categorical Variables**: These describe categories or groups without a numerical value (e.g., gender, political party).
  - *Nominal*: No order, just categories (e.g., red, blue, green).
  - *Ordinal*: Categories with an inherent order (e.g., satisfaction levels like "very satisfied", "neutral").
- **Numerical Variables**: These represent data that can be measured and quantified (e.g., weight, height).
  - *Discrete*: Countable values (e.g., number of students).
  - *Continuous*: Measurable quantities that can take any value within a range (e.g., temperature, time).

---

5. Levels of Measurement

Understanding the level of measurement determines the statistical tests that can be applied:

1. **Nominal**: Used for labeling and categorizing without a meaningful order (e.g., gender, nationality).
2. **Ordinal**: Data can be ordered, but the intervals between data points are not consistent (e.g., ranking).
3. **Interval**: Data can be ordered with meaningful intervals but no true zero (e.g., temperature in Celsius).

4. **Ratio**: Similar to interval data, but with an absolute zero point (e.g., weight, age).

---

6. Central Tendency

Central tendency refers to the "center" of a data set and is represented by the **mean**, **median**, and **mode**.

- **Mean**: The arithmetic average, calculated as the sum of all values divided by the total number of values.

$$\text{Mean} = \frac{\sum \text{data points}}{N}$$

- **Median**: The middle value in an ordered data set. If the number of data points is odd, it is the middle value; if even, it is the average of the two middle values.
- **Mode**: The most frequent value in the data set.

Example:

Data: 2, 4, 6, 8, 10

- Mean = $\frac{2+4+6+8+10}{5} = 6$
- Median = 6 (middle value)
- Mode = No mode (all values occur only once)

---

7. Measures of Variability

These measures describe the spread of data:

- **Range**: The difference between the highest and lowest values.

Formula:

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

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

  Formula:

$$\text{Variance} = \frac{\sum(X - \text{Mean})^2}{N}$$

- **Standard Deviation**: The square root of the variance. It shows the average distance from the mean.

  Formula:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

---

8. Probability Distributions

Probability distributions describe how probabilities are distributed over the values of the data. Some common types:

- **Normal Distribution**: Bell-shaped curve where most values are clustered around the mean.
- **Binomial Distribution**: Describes the number of successes in a fixed number of trials.

Example (Normal Distribution):

The heights of adult women are normally distributed with a mean of 64 inches and a standard deviation of 3 inches. If you were to randomly select a woman, the probability of her being between 61 and 67 inches tall can be calculated using the normal distribution formula.

---

9. Hypothesis Testing

Hypothesis testing allows us to make inferences about populations based on sample data.

- **Null Hypothesis (H_0)**: A statement suggesting no effect or no difference.
- **Alternative Hypothesis (H_1)**: A statement suggesting some effect or difference.

The steps in hypothesis testing:

1. State the null and alternative hypotheses.
2. Choose the significance level (e.g., $\alpha = 0.05$).
3. Compute the test statistic.
4. Compare the test statistic to the critical value.
5. Reject or fail to reject the null hypothesis.

---

10. Confidence Intervals

A confidence interval provides a range of values that likely contains the population parameter.

Example:

For a sample mean of 50 and a standard deviation of 10, a 95% confidence interval can be calculated as:

$$\text{CI} = \text{Mean} \pm Z \times \frac{\sigma}{\sqrt{N}}$$

Where Z is the Z-score corresponding to the desired confidence level.

---

11. Prediction and Forecasting

Prediction refers to estimating future outcomes based on current data, while forecasting typically refers to long-term predictions.

Example:

In time-series forecasting, historical sales data can be used to predict future sales using methods like **ARIMA (Auto-Regressive Integrated Moving Average)** or **exponential smoothing**.

---

12. Gaussian Mixture Models and Expectation Maximization

Gaussian Mixture Models (GMMs) are used for clustering, where the data is assumed to be generated from multiple Gaussian distributions.

Expectation Maximization (EM) is an algorithm used to find maximum likelihood estimates of parameters in statistical models, especially when the data has missing or incomplete values.

---

13. Conclusion

The **Introduction to Statistical Methods** course equips students with essential tools for analyzing data. By understanding concepts like **central tendency**, **measures of variability**, **probability distributions**, and **hypothesis testing**, students can apply these techniques to real-world problems. The knowledge gained from this course serves as a strong foundation for more advanced statistical and machine learning methods.

End of the guide

This guide covers all the key statistical concepts introduced during the course, providing clarity and mathematical examples for easy understanding. This step-by-step approach will serve as a reliable reference for both beginners and those preparing for exams or practical applications.